Reconocimiento de Formas

Clasificación y Aprendizaje

Francisco Mario Hernández Tejera
José Javier Lorenzo Navarro

Universidad de
Las Palmas de Gran Canaria

Reconocimiento de Formas: Clasificación y Aprendizaje

Francisco Mario Hernández Tejera
José Javier Lorenzo Navarro

ISBN: 84-699-8881-6

Impreso en Las Palmas de Gran Canaria, Mayo 2002

Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
España

Índice

Tema1: Conceptos Básicos en Reconocimiento de Formas

Tema 2: Reglas de Decisión

Tema 3: Aprendizaje Supervisado de Clasificadores

Tema1

Conceptos Básicos en Reconocimiento de Formas

1.1.- INTRODUCCION

Un aspecto importante de la actividad humana lo constituye el continuo interés por el diseño y desarrollo de herramientas y máquinas (entendidas en su sentido más amplio) con la finalidad de disminuir el esfuerzo físico y/o realizar procesos más rápido y/o mejor. Una orientación de ello, la primera históricamente, se refleja en el desarrollo de ingenios capaces de posibilitar, reducir o eliminar el esfuerzo en tareas de naturaleza física. Ejemplos pueden ser desde los tópicos del martillo, la rueda y la polea hasta la máquina de transporte más sofisticada que pueda diseñarse. Históricamente, este aspecto se ha englobado dentro de las tareas de interés tecnológico de lo que hoy en día se consideran las ingenierías clásicas. La otra orientación, cualitativamente diferente, es la que se refiere a las máquinas capaces de procesar información, categoría que puede agrupar por ejemplo, desde el ábaco hasta el computador tecnológicamente más avanzado.

El avance de las civilizaciones y de las ciencias genera una gran cantidad de información. En efecto, si quisiéramos medir el nivel de desarrollo de una cierta sociedad, un parámetro que probablemente debiéramos tener en cuenta es el referido a la cantidad de información que genera, hasta tal punto, que se podría sentenciar que sin información, la civilización no existe.

Los niveles de desarrollo de las sociedades industriales, fundamentalmente en la segunda mitad del presente siglo, han conllevado una explosión en el crecimiento de la cantidad de información generada. Como datos que avalan esta afirmación se podrían analizar los datos de crecimiento comparativo de ejemplos tambien tópicos, como son los correspondientes a la cantidad de información que se utiliza en la cada vez más avanzada Medicina para realizar diágnósticos, definir tratamientos o efectuar intervenciones quirúrgicas, o por otro lado, al número de revistas y artículos de pensamiento o científico-técnicos publicados en todo el mundo en diferentes lenguas, el número de operaciones bancarias, de efectos postales, de operaciones de teletransmisión, etc ... efectuados todos ellos a lo largo de un año. En la década de los años sesenta, con el desarrollo de la tercera generación de computadores, los diversos sectores económicos comenzaron a mostrar un interés cada vez mayor hacia la manipulación automatizada de la información. Este interés se transforma cada vez más en necesidad, con lo que se está produciendo un incremento creciente del nivel de penetración de los computadores en prácticamente todos los aspectos de nuestra vida cotidiana.

La disciplina raiz que engloba este aspecto es la **Informática**, entendida como *la disciplina del tratamiento y la representación mecanizadas de la información.* Históricamente, la informática no aparece espontaneamente, sino más bien como un desgajamiento de la **Cibernética**, definida por su creador N. Wiener como *el área*

dedicada al estudio de los procesos de control e información en los seres vivos y las máquinas.

Los computadores han permitido por primera vez abordar, gracias a sus prestaciones, un conjunto de problemas con los que ha especulado el hombre desde muy antiguo; *los relacionados con el diseño y realización de máquinas capaces de incorporar procesos análogos a los biológicos de información.* El área de la informática que se ocupa, entre otros, de estos problemas es la **Inteligencia Artificial**.

Entre los procesos comentados se encuentran los englobados bajo el epígrafe de **Percepción Artificial**, entendiendo como sistemas perceptuales a aquellos que *realizan la interpretación de impresiones sensoriales*, adquiriendo información acerca del entorno y, en cooperación con otros sistemas efectores, actuar sobre aquel y por tanto influenciarlo. El área dedicada al estudio de los procesos de percepción mecanizada es la del **Reconocimiento de Formas** (Pattern Recognition).

El reconocimiento es un atributo básico de los humanos y, en general, de los seres vivos. Realizamos actos de reconocimiento en cualquier instante de nuestras vidas; reconocemos los objetos del entorno que nos rodea y nos movemos o actuamos en relación a ellos, podemos distinguir a una persona conocida entre una multitud, la voz de un amigo, los gestos de una cara, un texto escrito, el olor del bizcocho de la abuela, el sabor de una naranja o el tacto de un trozo de hielo. En este sentido, los humanos somos un sistema de información muy sofisticado, con prestaciones de reconocimiento muy elevadas. Atendiendo al sentido de la palabra reconocimiento y según la naturaleza de las formas a reconocer **[TOU-74]**, podemos dividir los actos de reconocimiento en dos tipos: los referentes a items concretos y los referentes a items abstractos. Como ejemplos del primer tipo se encuentran el reconocimiento de textos escritos a mano, de los objetos que nos rodean o de una pieza musical. Es decir, el primer tipo recoge los actos de reconocimiento sensorial, a los que nos hemos referido en el párrafo anterior, los cuales hacen referencia a los procesos de identificación y clasificación de formas espaciales y/o temporales, y son el objeto de los contenidos de la asignatura que nos ocupa.

Como ejemplos del segundo tipo podemos citar los argumentos lógicos de una antigua reflexión, el reconocimiento, ante una cierta integral funcional, de la metodología de solución, aprendida tiempo atrás en un curso de Cálculo Integral. Los actos de este tipo se incluyen en el denominado Reconocimiento Conceptual, en contraste con el tipo anteriormente mencionado.

El reconocimiento de formas concretas por los seres humanos puede considerarse un problema psicofisiológico, en el que se establece una relación entre una persona y un estímulo físico. Cuando un sujeto percibe una forma, realiza un proceso de inferencia inductiva y asocia su percepción con una serie de pistas y

conceptos generales derivados de experiencias perceptuales pasadas. Podríamos interpretar, por tanto, que los actos humanos de reconocimiento son, en realidad, procesos de estimación de los parecidos entre los datos de entrada y los conceptos generales realizados en base a las pistas, constituyendo ambos la información a priori para el reconocimiento. En definitiva, se puede decir que el problema de reconocimiento de formas puede asimilarse a un proceso de discriminación de los datos de entrada entre poblaciones de conceptos, mediante la extracción de características o atributos individuales significativos de dichos datos de entrada.

El estudio de los problemas de reconocimiento puede dividirse en dos grandes áreas:

1.- El estudio de las habilidades o capacidades de reconocimiento de los seres humanos o seres vivos en general, que es un objetivo incluido en disciplinas como la psicología, la fisiología o la biología.

2.- El desarrollo de teorías, métodos y técnicas para el diseño de ingenios capaces de realizar ciertas tareas de reconocimiento en aplicaciones específicas, objetivo que cae dentro de las áreas de interés de la Informática en general y de la Inteligencia Artificial y Reconocimiento de Formas en particular. Este, por tanto es el objetivo que nos mueve en esta asignatura

Dicho objetivo, no obstante no consiste en una mera emulación de los procesos biológicos de reconocimiento. Al contrario, el objetivo de la disciplina, que es el diseño de máquinas con capacidades de reconocimiento se nutre en el desarrollo de las teorías, métodos y técnicas, anteriormente mencionados, del conocimiento disponible acerca del funcionamiento de los sistemas biológicos, pero este conocimiento no se orienta a la replicación de dichos sistemas naturales, sino más bien de sus capacidades. Una analogía que puede servir como ejemplo para concretar el comentario anterior es la que se refiere al vuelo. Los aviones y las aves vuelan en base al mismo principio, el de sustentación aerodinámica, que es el fundamento que utilizan los ingenieros aeronáuticos para el diseño de sus máquinas. Sin embargo, los diseñadores de aviones no han copiado el mecanismo de impulsión de las aves basado en el batir de alas, del que las ha dotado la evolución. Ello se debe a que la solución propuesta por aquellos resulta mucho más adecuada para el diseño de los aviones, tanto desde la óptica de eficacia en nuestro contexto económico como de los problemas mecánico-estructurales, en nuestro contexto tecnológico. En definitiva, tanto las aves como los aviones cumplen el mismo objetivo; volar, pero el camino seguido por los diseñadores de las máquinas para conseguirlo es claramente diferente al suministrado por la naturaleza.

Desde el punto de vista que nos ocupa, se puede entender como sistema de **Reconocimiento de Formas** a *aquel conjunto de procesos orientados a la*

transformación de señales o datos en experiencias o entidades con significado. Los *datos* a que se hace referencia presentan una relación causa-efecto con respecto a los hechos del mundo sobre los que el sistema de RF va a desarrollar su actividad, y se corresponden con la salida suministrada por el sistema sensor o de sensores que adquieren la información del *mundo* o *entorno* del sistema, como puede observarse en el esquema de la figura 1.1. Un ejemplo es la matriz de pixels suministrada por un sistema de adquisición de imágenes o el vector que corresponde a una señal muestreada por un sistema de adquisición de señales monodimensionales.

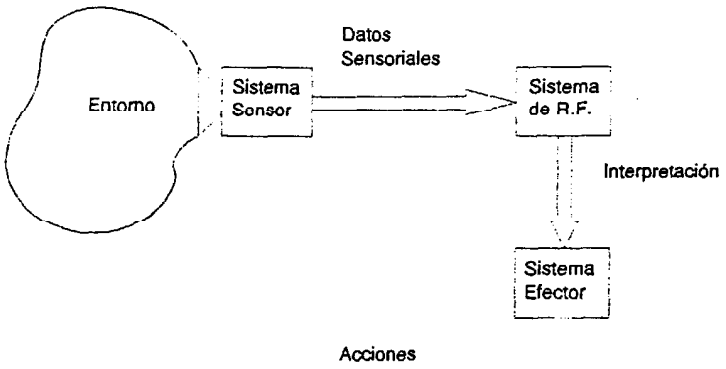


Figura 1.1.- Esquema de bloques de la ubicación de un sistema de Reconicimiento de Formas en un cierto entorno

Por otro lado, las denominadas *entidades con significado* se refieren a las salidas efectoras del sistema de Reconocimiento de Formas hacia aquel otro sistema que recibe la información suministrada por el primero para fines específicos, ya sea de simple monitorización o para actuar convenientemente sobre el *entorno*. Por ejemplo, un usuario que monitoriza con algún objetivo los resultados suministrados por un sistema de Reconocimiento de Formas, o un Sistema Robótico que actúa sobre el entorno de trabajo ,asistido por un sistema de Visión Artificial en una cadena de montaje y/o inspección visual.

El salto entre los datos sensoriales y la interpretación de los mismos en entidades con significado para el sistema efector se efectúa por un proceso de **Contrastación o Clasificación**. Dicho salto se suele efectuar a través del uso de una **representación** intermedia por los siguientes motivos:

1) Superar el abismo semántico (semantic gap) entre la estructura de los datos sensoriales y la estructura de las interpretaciones.

2) Permitir el diseño de una estructura de representación que simplifique y robustezca los procesos de contrastación o clasificación.

3) Compactar la información relevante a los efectos de interpretación de entre la disponible en los datos sensoriales, eliminando además aquella información presente en los datos de entrada que no es relevante a los efectos de interpretación.

El esquema de bloques general de un sistema de RF puede ser el indicado en la figura 1.2, donde el reconocimiento se efectúa según la secuencia siguiente: de los datos captados por el sensor del entorno o sistema físico se pasa a la representación de los mismos y con ella se realiza el proceso de contrastación o clasificación, con los modelos del entorno como referencia, para generar la interpretación. Ahora bien, una característica interesante de los sistemas de RF es su posibilidad de adaptación a diferentes entornos. Ello implica que los modelos se puedan modificar o adaptar a diferentes situaciones. Esto se puede conseguir si al sistema se le dota en el diseño con un segundo modo de funcionamiento, denominado **modo de análisis o de aprendizaje**, con el cual se efectúa la adaptación del modelo a las características propias del problema de reconocimiento, definiendo las categorías de formas propias del sistema y las reglas de asignamiento de formas incógnita para efectuar el proceso de contrastación.

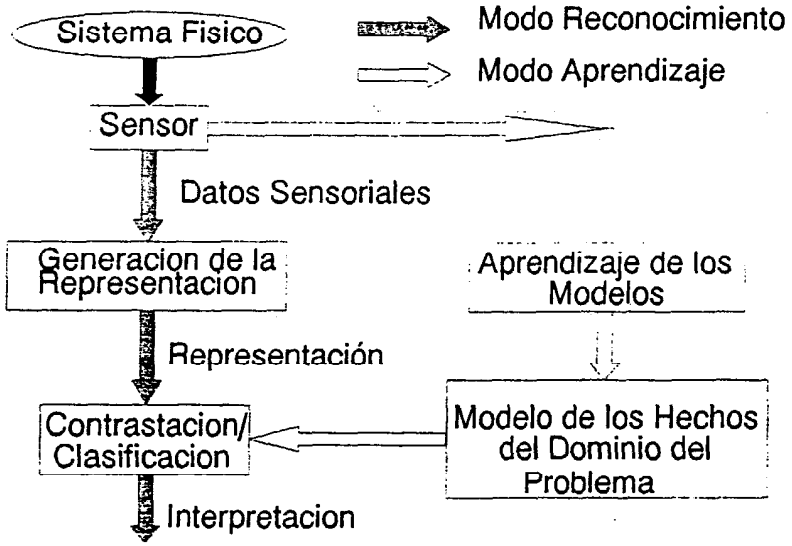


Figura 1.2.- Esquema General de un Sistema de Reconocimiento de Formas

Para concluir este apartado haremos una breve recopilación de otras definiciones de Reconocimiento de Formas citadas en la bibliografía especializada. De entre ellas podemos destacar:

I) *El Reconocimiento de Formas está relacionado con la descripción y análisis de procesos físicos y mentales* **[FU-84a]**.

II) *Actividades relacionadas con los aspectos matemáticos y técnicos de la percepción* **[NIEM-81]**.

III) *Categorización de datos de entrada en clases identificables mediante la extracción de características o atributos de los datos, respecto al fondo o detalles irrelevantes* **[TOU-74]**.

1.2.- CONCEPTOS BÁSICOS

Desde nuestro punto de vista, entendemos por forma a *cada una de las descripciones, cuantitativas o estructurales, de entidades o hechos del mundo o entorno del sistema*. Del mismo modo, se puede definir una clase de formas como *un conjunto de formas que poseen alguna propiedad común, y que pueden tener asociada alguna categoría semántica*. El objetivo del RF es, como hemos dicho, la asignación de formas a sus clases respectivas de manera mecanizada, es decir, automática y por tanto, con la menor intervención humana posible.

Por ejemplo, sea el problema de diseño de una máquina de RF que sea capaz de reconocer visualmente caracteres alfanuméricos. En este caso cada carácter, en todas las configuraciones posibles (p.e. mayúsculas y minúsculas), será una forma, y existirán tantas clases de formas como caracteres distintos se pueden presentar, es decir, 38: 27 correspondientes a los caracteres del alfabeto castellano y 10 correspondientes a los dígitos numéricos. El objetivo del sistema será la identificación de cada carácter de entrada (forma adquirida) como perteneciente a alguna de las clases de formas definidas.

La definición anterior tambien nos da pie a una reflexión acerca del fundamento último de los sistemas de RF. El objetivo que se plantea en dicha definición es la asignación de los datos adquiridos a una de entre un conjunto de clases de ellos previamente determinadas. Esto presupone que los datos admiten esa categorización, es decir, que existe algún tipo de **regularidad** entre ellos en base a la cual se pueden definir las categorias.

Las aproximaciones utilizables en la descripción de las formas en una "máquina" de RF concreta pueden ser muy variadas, y en muchos casos, la definición de una en concreto estará muy influenciada por el ámbito de aplicación de la misma, como se verá posteriormente. Así, por ejemplo, para la máquina anterior dedicada al reconocimiento de caracteres, los datos de entrada pueden ser los obtenidos con un dispositivo fotosensible conectado a un sistema de digitalización, que suministra al sistema datos en estructura matricial y en los que cada celda se corresponde con la respuesta de un elemento de la matriz fotosensible a una zona del plano donde se encuentra escrito o impreso el caracter en cuestión. Estos datos se pueden utilizar directamente para efectuar el proceso de reconocimiento, o se puede realizar previamente algún tipo de transformación sobre ellos, obteniendo un conjunto de medidas significativas que constituyan una representación intermedia. Como ejemplo de ellas podemos mencionar algún conjunto de medidas o relaciones geométricas extraidas de cada carácter.

Por otro lado, en el contexto que nos ocupa, se entiende por **forma simple** *a aquella que puede considerarse como un todo y, por tanto, como una sola clase o unidad a efectos del sistema de Reconocimiento de Formas.* Por el contrario, una **forma compleja** *es aquella que a efectos del sistema no puede considerarse como una unidad elemental, sino que está constituida por una agragación de formas simples con un conjunto de relaciones entre ellas.*

La consideración de un elemento objeto de reconocimiento como forma simple o compleja es a veces dependiente del problema de RF concreto que se plantea. En una cierta aplicación, puede considerarse simple, mientras que en otra, puede resultar más adecuado considerarla como compleja. Ejemplos de ello pueden ser los siguientes. Supongamos un sistema de RF orientado a reconocer los caracteres tipográficos producidos por una cierta máquina de escribir. Es posible en este caso considerar que los caracteres son formas simples, perfectamente diferenciados unos de otros, en los que es posible definir una forma como un cierto conjunto de características. Un caso diferente sería el de una máquina con el objetivo de reconocer caracteres escritos a mano. En este caso, existe una variedad estructural entre los caracteres constitutivos de una clase concreta, por ejemplo la de las letras "a", ya que no todas las personas escriben el caracter de la misma manera. Sería necesario en este caso determinar cuales son las reglas generales de constitución de las diferentes escrituras de la letra mencionada, en base a los trazos que la constituyen y las relaciones entre ellos. Por tanto, para realizar el reconocimiento resulta pués recomendable considerar a las formas como complejas.

1.3.- FORMALIZACIÓN DEL PROBLEMA

Para la formalización consideremos una situación como la mostrada en el diagrama de la figura 1.3, que recoge las relaciones entre los conjuntos de datos involucrados en el esquema de la figura 1.2.

Sean:

- Σ el *Dominio o Espacio Sensorial*, es decir, el conjunto de cantidades o datos que pueden ser medidos por parte de los dispositivos del sistema sensorial.

- χ el *Dominio o Espacio de las Representaciones o Características.*

- Λ el *Dominio de las Interpretaciones o de las Categorías de Formas*, que se corresponde con el de las etiquetas linguísticas asociadas a nivel de salida del

sistema a cada una de las clases de formas involucradas en el marco del sistema de Reconocimiento de Formas $\Lambda = \{ L_j; j=1,...,c \}$

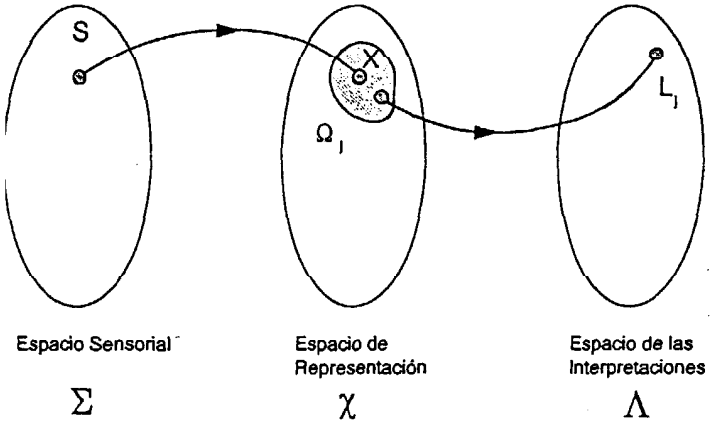


Figura 1.3.- Esquema de las relaciones entre los conjuntos de datos involucrados en un sistema de Reconocimiento de Formas

El conjunto de las interpretaciones debe cumplir las siguientes condiciones:

$$1) \quad L_j \neq \varnothing, \quad \forall\, j=1,...,c$$

$$2) \quad L_j \cap L_i = \varnothing, \quad \forall\, j \neq i$$

$$3) \quad \bigcup_{j=1}^{c} L_j = \Lambda$$

Es decir, no existe la interpretación nula, todas las interpretaciones son disjuntas y el conjunto de interpretaciones es completo.

En este contexto, un **Clasificador** o **Regla de Clasificación** d( ) es un proceso que tiene por objeto asignar una interpretación $L_j \in \Lambda$ a cada representación $\mathbf{X} \in \chi$, es decir:

$$d: \chi \rightarrow \Lambda$$

Tal que, $\forall$ **X** $\in \chi$, si **X** corresponde a una forma cuya interpretación es L_j, entonces:

$$d(\mathbf{X}) = L_j$$

Un clasificador d( ) es pues un mecanismo sistemático de asignación de interpretaciones a datos sensoriales.

En general, entre el dominio de las Representaciones y el dominio de las Interpretaciones se produce una compactación del volumen de la información asociada a cada forma en cuestión. Además en este proceso, a cada interpretación I_j se la suele hacer corresponder con un subconjunto del espacio o dominio de representación $\Omega_j \subseteq \chi$ más que con un solo elemento **X** (figura 1.3). Los motivos de ello los podemos resumir en los dos siguientes:

1) "Ruido" que se introduce, por diferentes motivos, en los procesos del sistema sensorial y que provoca perturbaciones en los datos sensoriales. Ello a su vez se refleja en variaciones en las representaciones, por lo que una misma identidad linguística o interpretación se podrá corresponder con diferentes representaciones que se corresponderán a su vez con versiones "degradadas" de una hipotética representación ideal.

2) Pequeñas variaciones o distorsiones en la forma original que provocan las correspondientes en los datos sensoriales, que, de la misma manera se transmiten a las representaciones. Dichas variaciones, sin embargo, no modifican las identidades linguísticas I_j de las formas.

En ninguno de los dos casos, el sistema deberá modificar por estos hechos la identidad asociada con la forma original.

Así, existirán tantos subconjuntos $\{\Omega_j; j=1,...,c\}$ en el espacio de representación como interpretaciones posibles. Cada uno de esos subconjuntos se corresponderá con una clase de formas, que agrupará a todas aquellas variaciones o versiones de la forma original, generadas por las causas mencionadas. Teniendo en cuenta lo anterior, podríamos definir una **clase de formas** como *cada una de las c particiones del espacio de representaciones.*

El conjunto de las clases de formas cumplen, en general, las siguientes propiedades:

Es decir, ninguno es vacío y todos cubren el dominio completo de las representaciones.

$$1)- \quad \Omega_j \neq \varnothing$$

$$2)- \quad \bigcup_{j=1}^{c} \Omega_j = \chi$$

En este contexto, podemos reescribir la definición de **clasificación** como *el proceso por el cual una forma incógnita definida por los datos de un acto sensorial, resulta asignada a una de entre las clases de formas, teneindo cada una de ellas asignada una interpretación.*

La **partición** del espacio de representación según el conjunto de clases de formas $\{\Omega_j;\ j=1,...,c\}$ se efectúa, en general, mediante procesos de aprendizaje, cuyo objetivo es determinar dicha partición a partir de las regularidades detectadas en las representaciones, las cuales están relacionadas con los dos fenómenos citados anteriormente.

Si el funcional de clasificación es tal que le asociamos una **Función de Confianza** $f \in [0,1] \subseteq \mathbb{R}$, cuyo objeto es asignar una *medida de confianza* a la interpretación I_j correspondiente a una representación **X**:

$$d_f : \chi \rightarrow <\Lambda, F>$$

De forma que:

$$\forall\ X \in \chi,\ d(X) = < L_j , f_j >$$

Se podría generalizar la definición de los procedimientos de clasificación en Reconocimiento de Formas. Así, si **f** se interpreta como una probabilidad, se definiría el marco de clasificación estadístico, mientras que si se interpreta como una posibilidad, tendríamos un marco de clasificación borroso.

En definitiva, el objetivo nuclear del Reconocimiento de Formas está relacionado con la construcción de algoritmos que implementen clasificadores d( ) y de aquellos otros que permitan generarlos, es decir, algoritmos de aprendizaje.

1.4.- APROXIMACIONES EN RECONOCIMIENTO DE FORMAS

Existen dos aproximaciones básicas a la solución de los problemas de RF, definidas en función del tipo de representación utilizada para las formas:

I) Aproximación de Teoría de la Decisión. Se considera cuando el sistema va a manipular formas simples. En este caso, las mismas formas son los elementos primitivos de la representación, no planteándose descomposición estructural de las mismas. Se utilizan funciones de decisión para clasificar las formas, representadas como vectores de características.

II) Aproximación Estructural. Su utilidad se encuentra en problemas en los que las formas en estudio se consideran complejas, es decir, constituidas por formas simples y relaciones estructurales entre las formas simples. En este caso, las formas resultan representadas como cadenas, árboles, grafos, etc., estableciéndose una analogía entre la estructura de la forma y la sintaxis de un lenguaje, o entre la estructura de la forma y una red estructural con contenido semántico. El proceso de reconocimiento resulta, en el caso sintáctico, uno de análisis (parsing), o en el caso estructural, un problema de comparación entre grafos o de inferencia.

En ambas aproximaciones, todo sistema de RF puede estas dotado, como se ha comentado anteriormente, de dos modos de funcionamiento: el denominado de reconocimiento propiamente dicho y el de análisis o aprendizaje. Las aproximaciones metodológicas al diseño de las etapas de aprendizaje pueden ser:

I) Aprendizaje Supervisado (Learning with a Teacher). En este caso se dispone de una muestra controlada de cada una de las clases de formas en el diseño del clasificador de la cual se conocen a priori las etiquetas de pertenencia a clase de todos y cada uno de los elementos de dicha muestra. El proceso de aprendizaje consiste pues en la determinación, en base a la muestra, de las reglas de clasificación a partir de las regularidades de la misma.

II) Aprendizaje no Supervisado (Learning without a Teacher). En este caso no se dispone de las etiquetas de pertenencia a clase de los elementos de la muestra controlada, por lo que antes de obtener las reglas de clasificación es preciso analizar el conjunto de datos para determinar el número de clases de formas que la constituyen así como sus regularidades.

III) Aprendizaje mediante Refuerzo (Reinforcement Learning). Es un concepto extraido de la literatura de aprendizaje animal, y se refiere a una clase de tareas y algoritmos de aprendizaje, para los cuales este último se efectúa maximizando una evaluación escalar o **refuerzo** de la *calidad* de lo aprendido acerca del problema en cuestión. Si la medida de calidad se obtiene tras cada paso del proceso de aprendizaje por una muestra, el procedimiento se denomina de refuerzo inmediato, mientras que si se obtiene el refuerzo, no para cada muestra sino para un colectivo de ellas, el procedimiento se denomina de refuerzo retardado (delayed reinforcement).

1.5.- POSTULADOS DE NIEMANN

Cualquier aproximación al diseño de sistemas de RF se debe plantear teniendo en cuenta en lo posible los postulados siguientes **[NIEM-81]**:

1) Representatividad. El diseño de un proceso de RF en un campo dado requiere de una muestra representativa de formas.

2) Discriminabilidad. Una forma simple tiene siempre características que permiten determinar su grado de pertenencia a una clase, esto es, siempre es posible encontrar un conjunto de características que discriminen una forma simple en algún tipo de espacio.

3) Compacidad y Separación. Las características de las formas de una clase ocupan un dominio compacto de la representación, y los dominios ocupados por clases diferentes están separados. El que se verifique este postulado es condición necesaria para que pueda tener lugar un proceso de RF. En la práctica resulta de extrema importancia la elección de características que lo cumplan.

Así, según Niemann, el problema central del RF es encontrar o generar una representación que cumpla con el postulado 3, por cuanto si la encontramos, el problema de la clasificación está resuelto, al menos formalmente.

4) Una forma compleja se compone de constituyentes más simples, entre los cuales existen ciertas relaciones de estructura.

5) A un conjunto de componentes más unas relaciones entre ellos le corresponde una y solo una forma compleja.

Para una aplicación general con formas complejas son necesarios los cinco postulados, mientras que para una con formas simples son sólo necesarios los tres primeros.

Como ilustración a aspectos relacionados con el postulado de discriminabilidad transcribimos un cuento recogido por B. K. P. Horn en su texto **[HORN-87]**, y que contiene un principio básico útil para el diseño de clasificadores para sistemas de RF y que se hará patente durante el curso:

Había una vez dos granjeros vecinos, Jed y Ned, que eran propietarios cada uno de un caballo. Los dos caballos tenían por costumbre saltar a un lado y otro de la valla que separaba las granjas, por tanto, los granjeros necesitaban de algún medio para discriminar cual era el caballo de cada uno.

Así que Jed y Ned se reunieron y decidieron un esquema de discriminación, que consistió en lo siguiente. Jed hizo un corte en la oreja de su caballo, no muy grande, pero lo suficiente para que se pudiese observar a simple vista. Pero, al día siguiente de que Jed realizase dicha operación, el caballo de Ned se hirió la misma oreja en el alambre de espino de la valla, resultando así ambos animales con una cicatriz idéntica. De nuevo algo había que hacer para distinguirles, por lo que decidieron que Ned atara una cinta azul al cuello de su caballo como elemento distintivo. Pero al día siguiente, el caballo de Jed saltó la valla, corrió hacia el lugar donde estaba pastando el de Ned, le arrancó la cinta con los dientes y se la comió.

Desesperados, los granjeros volvieron a reunirse y Jed sugirió que debían buscar una característica que fuese menos sensible a cambios. La altura les pareció buena, y decidieron comprobar si los caballos eran diferentes en altura. Cada granjero procedió a medir su caballo y, efectivamente, eran distintos en altura, ya que ¡el caballo marrón era exactamente dos pulgadas más alto que el caballo blanco!.

Y, como de muchos cuentos, de este se puede extraer una moraleja, que traspasada a nuestro objeto de interés se puede expresar como:

Cuando se tengan dificultades en la clasificación, no intentemos resolver de entrada el problema a traves de intrincados y esotéricos trucos matemáticos, sino, por el contrario, busquemos mejores características.

1.6.- APROXIMACION DE TEORIA DE LA DECISION

La aproximación de Teoría de la Decisión al RF se basa en la utilización de funciones de decisión para la clasificación de formas previamente representadas por vectores de características n-dimensionales:

$$\boldsymbol{X} = [X_1 \;\; X_2 \;\; . \;\; . \;\; X_i \;\; . \;\; . \;\; X_n]^t$$

Donde x_i representa la i-ésima característica del vector. Un diagrama de bloques de la estructura de un sistema de RF según esta aproximación se muestra en la figura 1.4.

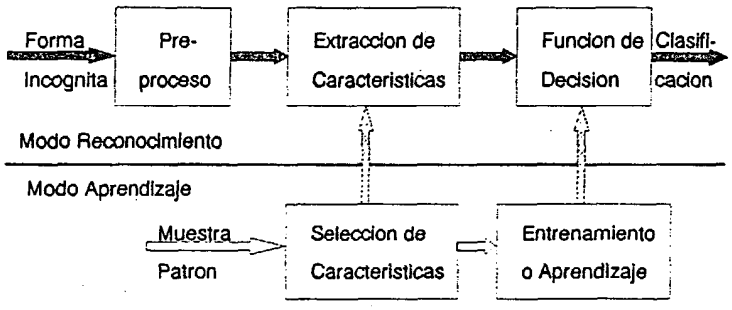


Figura 1.4.- Diagrama de Bloques de un Sistema de RF en Aproximación de Teoría de la Decisión

El problema se puede abordar según uno de los siguientes esquemas:

1) Esquema Geométrico. En este caso, la clasificación se puede plantear como la asignación del vector de características incógnita, entendido este como un punto del espacio de características, a una de las particiones (regiones) mútuamente exclusivas previamente obtenidas en dicho espacio. Cada una de esas particiones se corresponde con una de las c clases de formas Ω_j (j=1,2,...c). Para efectuar el proceso de asignación, durante el proceso de aprendizaje se define un conjunto de c funciones discriminantes $d_j(\mathbf{X})$, asociada cada una a una de las c clases de formas. La regla de clasificación se construye, a continuación en base a esas funciones discriminantes, con la cual se efectuará posteriormente el proceso de clasificación de las formas incógnitas.

2) Esquema Estadístico. En este caso se considera que los elementos del vector de características son variables aleatorias, siendo x_i una medida *ruidosa* de la i-ésima característica. Para cada clase de formas Ω_j (j=1,...,c) se hace uso del conocimiento de su estructura probabilistica de distribución y la probabilidad a priori de de ocurrencia de las muestras. En base a estos atos se construye una regla de decisión de naturaleza probabilística basada en el objetivo de minimizar las probabilidades de reconocimiento erróneo.

3) Esquema de Redes Neuronales. En este caso, se considera la utilización de elementos provenientes del paradigma de las redes neuronales artificiales para resolver problemas de RF.

1.7.- APROXIMACION ESTRUCTURAL

Cuando las formas a reconocer presentan cierto grado de complejidad, la caracterización de las formas en base a vectores de un cierto espacio de representación dan lugar a esquemas resolutivos del problema de RF impracticables. Puede deberse, por ejemplo, a que el número de características requeridas para una representación adecuada de los objetos sea muy grande y/o porque resulte excesivo el número de clases en que se debería dividir el espacio de representación para reflejar toda la variabilidad significativa de los objetos.

En estos casos, el proceso de interpretación no se corresponde con una simple clasificación o identificación, como ocurrre en los planteamientos de Teoría de la Decisión, sino con uno de **descripción estructural**, en el cual, los objetos o formas complejas se describen como composición de subobjetos, siendo los subobjetos más simples las denominadas *primitivas*, o formas básicas de la representación.

Un esquema en diagrama de bloques de un sistema de RF según una aproximación estructural se muestra en la figura 1.5.

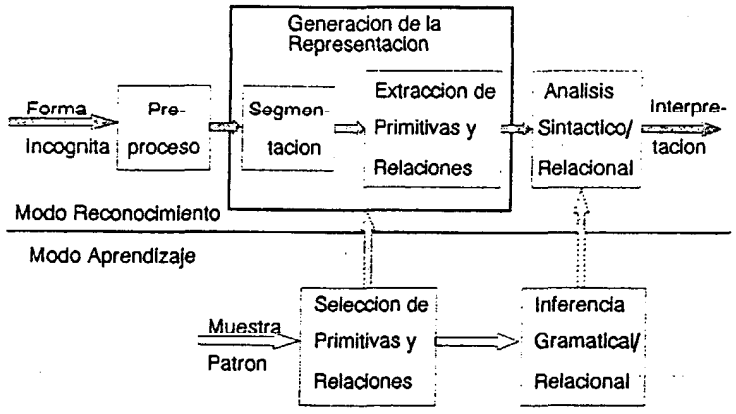


Figura 1.5.- Diagrama de Bloques de un Sistema de Reconocimiento de Formas en Aproximación Estructural.

En dicho diagrama se observa coma la etapa de generación de la representación está constituida por un proceso de segmentación en el que se

detectan las partes o primitivas de la representación y en un proceso posterior en el que se codifica la estructura de primitivas y relaciones entre ellas. El proceso de contrastación posterior se puede efectuar según alguno de los dos siguientes esquemas:

1) **Esquema Sintáctico.** En este caso, la representación estructural de un objeto o escena, en base a las primitivas y las operaciones de composición, se corresponde con una frase de un cierto *Lenguaje de Descripción de las Formas*, y el conjunto de reglas de composición se corresponde con las *reglas de producción* de una cierta gramática. La descripción estructural entonces se puede plantear como un problema de Análisis Sintáctico.

2) **Esquema Relacional.** Para los problemas discutidos se puede escoger un segundo esquema, que se basa en considerar que las estructuras de representación son grafos relacionales, redes semánticas u otro tipo (como los *frames*), de las utilizadas ampliamente en el marco de la Inteligencia Artificial. En este caso, el proceso de contrastación se efectúa como si de un problema de isomorfismo entre grafos, o en su caso, como uno de inferencia.

Por último, hay que hacer notar que, si bien la aproximación de Teoría de la Decisión y la Estructural están claramente diferenciadas, la segunda aproximación precisa, en general, hacer uso de elementos de la primera en la etapa de generación de la representación simbólica, tanto para establecer las clases de primitivas y relaciones en el aprendizaje como para etiquetar los elementos durante el proceso de obtención de la descripción.

1.8.- ALGUNAS APLICACIONES DEL RECONOCIMIENTO DE FORMAS

A continuación incluimos, con intención ilustrativa, un conjunto de problemas en los que la aplicación de las técnicas de Reconocimiento de Formas resulta de interés económico. La lista incluida no pretende ser exahustiva, pero si lo suficientemente amplia para dar una visión del interés creciente que estas técnicas está generando en los campos de aplicación más variados. Hay que hacer notar que, algunas de las aplicaciones mencionadas pueden hacer uso, no de todos los procesos de un sistema de RF, sino más bien de algunos de ellos y, por tanto, de lo que hacen uso es de ciertas técnicas que forman parte de la disciplina.

1) **Reconocimiento de Caracteres:** Desde los sistemas clásicos OCR (Optical Character Recognition) en las máquinas de lectura automática de los caracteres codificados de los cheques bancarios (por ejemplo, los basados en

el conjunto estandar de caracteres ABA E-13B), a los sistemas OCR comerciales actuales de lectura automática de textos, a aquellos otros más sofisticados como los lectores automáticos de periódicos, de documentos complejos, por ejemplo, con fórmulas matemáticas, o en caracteres orientales como chino o japonés, o los más complejos de lectura de textos manuscritos.

2) Reconocimiento de Huellas Digitales y/o de Caras: cuyo interes es claro y evidente, cuando se trata de diseñar sistemas automáticos para la detección de la identidad a partir de ellas, teniendo en cuenta lo amplios que suelen ser los ficheros de identificación, o para su utilización en control de accesos a zonas o medios restringidos.

3) Clasificación e Identificación en Aplicaciones Relacionadas con Imágenes Aéreas o de Satélite: que van desde la identificación y seguimiento de objetivos hasta las de investigación de recursos naturales o estados climáticos a partir de imágenes multiespectrales.

4) Aplicaciones Industriales: que pueden ser de control de calidad de productos, asociadas a la automatización de procesos de producción, asistencia automática a procesos de ensamblaje, etc.

5) Reconocimiento Auditivo de Patrones: es decir, sistemas relacionados con el reconocimiento automático del habla humana, en todas sus variantes posibles: de palabras aisladas o discurso continuo, de un solo sujeto o de múltiples, etc...

6) Identificación de Objetivos: a partir de las señales suministradas por equipos de sonar, radar u otras bandas em.

7) Análisis de Escenas en Movimiento: como asistencia visual en vehículos autoguiados o sistemas autónomos móviles en general.

8) Aplicaciones en Biomedicina: como las de análisis de ECG, EEG, análisis y recuentos cromosómicos, tests clínicos en general, etc.

9) Aplicaciones de Control o Toma de Decisión Basadas en la Información Suministrada por Diferentes Tipos de Sensores: como pueden ser los sensores térmicos, los táctiles, detectores-medidores de emanaciones gaseosas, radiaciones o ciertas partículas o compuestos químicos.

1.9.- REFERENCIAS

[ANZA-89] Anzai Y., **Pattern Recognition and Machine Learning**, Academic Press Inc., San Diego, CA, 1989.

[BOW-92] Bow S., **Pattern Recognition and Image Processing**, Marcel Dekker Inc., New York, 1992.

[BREI-84] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., **Classification and Regression Trees**, Wadsworth & Brooks/Cole Advanced Books & Sotfware, Pacific Grove, CA, 1984.

[CHEN-93] Chen C. H., Pau L. F., Wang P. S. P., **Handbook of Pattern Recognition and Computer Vision**, World Scientific Pub. Co., Singapore, 1993.

[DUDA-73] Duda R. O., Hart P. E., **Pattern Classification and Scene Analysis**, John Wiley & Sons, New York, 1973.

[FU-82] Fu K. S., **Syntactic Pattern Recognition and Applications**, Prentice-Hall, Englewood Cliffs, N. J., 1982.

[FU-84] Fu K. S., Rosenfeld A., *Pattern Recognition and Computer Vision*, IEEE Computer, vol. 17, nº 10, pp. 274-282. Oct. 1984.

[GONZ-77] Gonzalez R. C., Wintz P., **Digital Image Processing**, Addison-Wesley Pub. Co., London, 1977.

[HORN-87] Horn B. K. P., **Robot Vision**, The MIT Press, Cambridge, Mass., 1987.

[MEND-70] Mendel J. M., Fu K. S., **Adaptive, Learning and Pattern Recognition Systems**, Academic Press, New York, 1970.

[NIEM-81] Niemann H., **Pattern Analysis**, Springer-Verlag, 1981.

[PAL-86] Pal S. K., Majumder D. K. D., **Fuzzy Mathematical Approach to Pattern Recognition**, Wiley Eastern Ltd, New Delhi, India, 1986.

[PAVL-77] Pavlidis T., **Structural Pattern Recognition**, Springer-Verlag, Berlin, 1977.

[PERL-94] Perlovsky L. I., *Computational Concepts in Classification: Neural Networks, Statistical Pattern Recognition, and Model-Based Vision*, Journal of Mathematical Imaging and Vision, Vol. 4, pp. 81-110, 1994.

[SIMO-86] Simon J.-C., **Patterns and Operators. The Foundations of Data Representation**, North Oxford Academic Pub. Ltd., London, 1986.

[SCHA-92] Schalkoff R., **Pattern Recognition. Statistical, Structural and Neural Approaches**, John Wiley & Sons, Inc., New York, 1992.

[TOU-74] Tou J. T., Gonzalez R. C., **Pattern Recognition Principles**, Addison-Wesley, 1974.

[WEIS-91] Weiss S. M., Kulikowski C. A., **Computer Systems that Learn**, Morgan Kaufmann Pub. Inc., San Francisco, CA, 1991.

[YOUN-86] Young T. Y., Fu K. S. (eds.), **Handbook of Pattern Recognition and Image Processing**, Academic Press, London, 1986.

Tema 2

Reglas de Decisión

2.1.- INTRODUCCION

El núcleo de todo sistema de Reconocimiento de Formas lo constituye el módulo generador de las decisiones, que asigna las formas incógnitas a las clases de formas previamente definidas, según reglas preestablecidas. El estudio de las reglas de decisión es el objetivo de este tema y dado que en esta parte de la asignatura nos centramos en el estudio de los métodos de RF según la aproximación de Teoría de la Decisión, analizaremos:

a) Reglas de Decisión en Problemas con planteamiento geométrico: Funciones Discriminantes y Criterios de Distancia.

b) Reglas de Decisión en problemas con Planteamiento Estadístico.

2.2.- FUNCIONES DISCRIMINANTES Y SUPERFICIES DE DECISION

Dado un espacio n-dimensional E (p. e. $\mathbb{R}^n$) donde se ha definido un conjunto de c clases $\{\Omega_1,\Omega_2,...,\Omega_c\}$ y asociada a cada clase i se encuentra un funcional $d_i(\boldsymbol{X})$, donde $\boldsymbol{X}$ representa a un vector de medibles o características del espacio, se puede establecer una ***Regla de Clasificación*** basada en estos funcionales de la siguiente manera: *el clasificador asigna el vector de características* **X** *de la forma incógnita a la clase* Ω_i *con la que se cumple*:

$$d_i(\boldsymbol{X})>d_j(\boldsymbol{X}) \quad \forall j\neq i \qquad [2.1]$$

Al conjunto de funcionales mencionados se les denomina ***Funciones de Decisión o Funciones Discriminantes.***

Si planteamos cada inecuación de la siguiente manera:

$$d_i(\boldsymbol{X})-d_j(\boldsymbol{X})>0 \qquad [2.2]$$

Su límite inferior vendrá definido por la ecuación:

$$d_{ij}(\boldsymbol{X})=d_i(\boldsymbol{X})-d_j(\boldsymbol{X})=0 \qquad [2.3]$$

Que en el espacio n-dimensional de la representación se corresponde con una hipersuperficie $d_{ij}(\boldsymbol{X})$ que separa las clases i y j. A esta hipersuperficie se la denomina ***Frontera de Decisión o Superficie de Decisión.***

La naturaleza de las funciones discriminantes se define en base a la aproximación que se haga al problema:

a) Si se considera al espacio de representación como uno de naturaleza estadística, donde las distribuciones de las clases son conocidas o determinables por aplicación de ciertas técnicas, el problema de clasificación es de naturaleza estadística paramétrica, y las funciones discriminantes serán funcionales estadísticos.

b) Si, por el contrario, no se considera dicha naturaleza estadística, el problema se plantea como uno de decisión geométrica, en el que las funciones discriminantes son funcionales deterministas paramétricos.

El éxito de los esquemas de clasificación de formas mediante funciones de decisión depende de dos factores:

I) La forma de la función de decisión, directamente relacionada con las propiedades de las clases en consideración. Si no se posee información previa acerca de las clases en cuestión así como de su distribución en el espacio, la única manera de establecer la efectividad de una función de decisión es mediante prueba directa.

II) La determinación de los parámetros de la función, que se resuelve mediante esquemas de aprendizaje, normalmente a partir de muestras de formas, cuestión que abordaremos en el próximo tema.

2.3.- DISCRIMINANTE LINEAL BICLASICO

Sea un problema de clasificación entre dos clases Ω_1 y Ω_2 en un espacio bidimensional, es decir, donde el vector de características es de la forma:

$$X=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad [2.4]$$

Y asociados a las clases se encuentran las funciones discriminantes $d_1(\mathbf{X})$ y $d_2(\mathbf{X})$, que intervienen en la clasificación a través de la siguiente regla:
Si las funciones discriminantes son de la forma:
Es decir, combinaciones lineales de las coordenadas del vector de características, a la función se la denomina ***Discriminante Lineal***, y la correspondiente ***Superficie de***

$$\forall X: \begin{cases} si\ d_1(X)>d_2(X) \rightarrow X\in\Omega_1 \\ si\ d_1(X)<d_2(X) \rightarrow X\in\Omega_2 \end{cases} \quad [2.5]$$

$$d_i(X)=\omega_{i1}x_1+\omega_{i2}x_2+\omega_{i0} \quad \forall i=1,2 \quad [2.6]$$

Decisión será:

$$\begin{gathered} d(X)=d_{12}(X)=d_1(X)-d_2(X)= \\ (\omega_{11}-\omega_{21})x_1+(\omega_{12}-\omega_{22})x_2+(\omega_{10}-\omega_{20})= \\ \omega_1x_1+\omega_2x_2+\omega_0=0 \end{gathered} \quad [2.7]$$

Que es la ecuación de una recta como la mostrada en la figura 2.1.

La ecuación de la superficie de decisión puede tambien usarse como base de una regla de decisión, tal y como se muestra a continuación:

$$\forall X: \begin{cases} si\ \ d(X)>0 \rightarrow X\in\Omega_1 \\ si\ \ d(X)<0 \rightarrow X\in\Omega_2 \end{cases} \quad [2.8]$$

Un esquema del clasificador lineal biclásico se muestra en la figura 2.2.

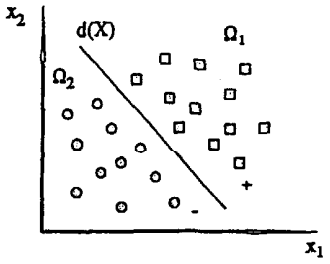


Figura 2.1: Recta de Decisión en el Plano de Características.

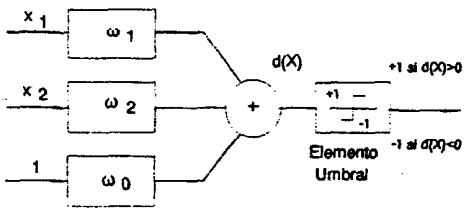


Figura 2.2: Esquema del Clasificador Lineal entre dos clases.

La función discriminante biclásica puede generalizarse a un espacio n-dimensional, en cuyo caso y por extrapolación del caso anterior, su expresión puede ser:

$$d_i(X)=\omega_{i1}x_1+\omega_{i2}x_2+\ldots+\omega_{in}x_n+\omega_{i0} \quad \forall i=1,2 \quad [2.9]$$

Que puede expresarse en forma vectorial como:

$$d_i(\mathbf{X}) = \boldsymbol{\omega}_i^t \mathbf{X} + \omega_{i0} \quad \forall i=1,2 \qquad [2.10]$$

Donde $\omega_i^t=[\omega_{i1}\ \omega_{i2}\ \ldots\ \omega_{in}]$ es la traspuesta del denominado *Vector de Pesos o de Parámetros* y ω_{i0}, el *Peso o Parámetro Umbral* de la clase Ω_i.

La regla de decisión basada en las funciones discriminantes tendrá por expresión la 2.5. Así mismo, la superficie de decisión será:

$$\begin{gathered} d(\mathbf{X}) = d_{12}(\mathbf{X}) = \\ (\omega_{11}-\omega_{21})X_1+(\omega_{12}-\omega_{22})X_2+ \ \ldots\ +(\omega_{1n}-\omega_{2n})X_n+(\omega_{10}-\omega_{20}) = \\ \omega_1 X_1+\omega_2 X_2+ \ \ldots\ +\omega_n X_n-\omega_0 = \boldsymbol{\omega}^t \mathbf{X}+\omega_0 = 0 \end{gathered} \qquad [2.11]$$

Que se corresponde con un *hiperplano* del espacio n-dimensional, siendo el vector de pesos normal al hiperplano.

Si expresamos la ecuación 2.11 en la forma:

$$\omega^t \mathbf{X} = -\omega_0 \qquad [2.12]$$

Y expresamos al vector de pesos en función del vector unitario **u** en su dirección, podemos poner:

$$\mathbf{u}^t \mathbf{X} = -\frac{\omega_0}{\|\omega\|} = D_u \qquad [2.13]$$

Ecuación que nos dice que, el cociente, cambiado de signo, entre en peso umbral y el módulo del vector de pesos se corresponde con la distancia del hiperplano al origen de referencia, como puede observarse, para un caso bidimensional, en la figura 2.3.

Los signos opuestos que presentan los valores de la función $d(\mathbf{X})$ en las dos particiones del espacio separadas por la superficie de decisión se pueden analizar en el siguiente ejemplo bidimensional ilustrado en la figura 2.3.

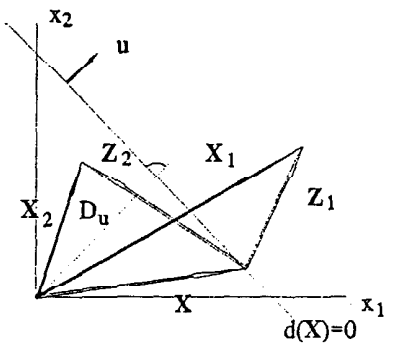


Figura 2.3: Configuración de vectores para el análisis de la Función de Decisión.

Sean los vectores X_1 y X_2 que se corresponden a puntos situados a uno y otro lado, respectivamente de la recta de decisión. Ambos vectores se pueden representar como la suma de otros dos vectores:

$$\begin{aligned} X_1 &= X + Z_1 \\ X_2 &= X + Z_2 \end{aligned} \qquad [2.14]$$

Para el primer vector, X_1, podemos poner:

$$\frac{d(X_1)}{\|\omega\|} = \frac{\omega^t X_1}{\|\omega\|} + \frac{\omega_0}{\|\omega\|} = uX_1 - D_u = \qquad [2.15]$$

$$(uX - D_u) + uZ_1$$

El término entre paréntesis es nulo, como se desprende de la ecuación 2.13 para todo vector **X** situado sobre la recta de decisión, con lo cual resulta:

$$\frac{d(X_1)}{\|\omega\|} = uZ_1 \qquad [2.16]$$

Como el ángulo formado por el vector unitario **u** y el vector Z_1 es agudo, el producto escalar será positivo, con lo cual $d(X_1)>0$. Análogamente se puede realizar el análisis para X_2, obteniéndose que $d(X_2)<0$, para todo vector del semiplano correspondiente.

El discriminante lineal de las expresiones 2.9 y 2.10 puede representarse en la denominada *forma homogénea*, que consiste en expresar la función discriminante como un producto matricial. Si partimos de la expresión matricial del discriminante según 2.9:

$$d(X)=(\omega_1 \quad \omega_2 \quad . \quad . \quad . \quad \omega_n)\begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} + \omega_0 = \sum_{i=1}^{n} \omega_i x_i + \omega_0 \qquad [2.17]$$

Y si establecemos las siguientes igualdades:

$$\begin{aligned} a_i &= \omega_i \quad \forall i=0,1,...,n \\ y_0 &= 1 \\ y_i &= x_i \quad \forall i=1,2,...,n \end{aligned} \qquad [2.18]$$

Podemos construir dos nuevos vectores: $\boldsymbol{a}$ e $\boldsymbol{Y}$, correspondientes a pesos y características, de dimensión (n+1), con lo cual podemos escribir el discriminante como:

$$d(X) = \sum_{i=0}^{n} a_i y_i = \boldsymbol{a}^t \boldsymbol{Y} =$$

$$(\omega_0 \quad \omega_1 \quad . \quad . \quad . \quad \omega_n) \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} = (\omega_0 \quad \boldsymbol{\omega}^t) \begin{pmatrix} 1 \\ \boldsymbol{X} \end{pmatrix} \qquad [2.19]$$

Lo que equivale a realizar un mapeo del espacio original de características n-dimensional a otro espacio n'-dimensional (n'=n+1). Este mapeo preserva todas las relaciones de distancia entre las muestras y la superficie de decisión, que será:

$$\boldsymbol{a}^t \boldsymbol{Y} = 0 \qquad [2.20]$$

Y que pasará por el origen del nuevo sistema referencial.

2.2.2.- DISCRIMINACION LINEAL MULTICLASICA

Cuando el problema de decisión se establece entre c clases (c>2), se presentan varias opciones para establecer reglas de decisión, las cuales analizaremos a continuación.

a) A cada clase Ω_i se asocia un discriminante lineal $d_i(\mathbf{X})$ que presenta el siguiente comportamiento para cualquier vector $\boldsymbol{X}$:

$$d_i(\boldsymbol{X})\begin{cases} >0 \ si \ \boldsymbol{X}\in\Omega_i \\ <0 \ si \ \boldsymbol{X}\notin\Omega_i \end{cases} \quad \forall i=1,2,...,c \qquad [2.21]$$

O dicho de otra manera, existe un hiperplano de decisión $d_i(\mathbf{X})=0$ que separa a cada clase Ω_i de las c-1 clases restantes. La regla de decisión del clasificador se establece como:

$$\boldsymbol{X}\in\Omega_i \ si \begin{cases} d_i(\boldsymbol{X})>0 \\ d_j(\boldsymbol{X})<0 \quad \forall j\neq i \end{cases} \qquad [2.22]$$

Podemos analizar la regla partiendo de un ejemplo bidimensional triclásico como el mostrado en la figura 2.4. En ella se muestran las regiones correspondientes a las clases: Ω_1, Ω_2 y Ω_3 y las correspondientes fronteras de decisión. Dada la naturaleza de la regla de clasificación y la estructuración de las fronteras de decisión, que separan cada clase de todas las demas, en el espacio de representación aparecerán regiones que no cumplen la regla, es decir, no asignables a ninguna de las clases. Dichas regiones se etiquetan en la figura como R.I. (Regiones Indeterminadas). En la figura tambien se indican las desigualdades correspondientes a las funciones de decisión que permiten corroborar la afirmación anterior.

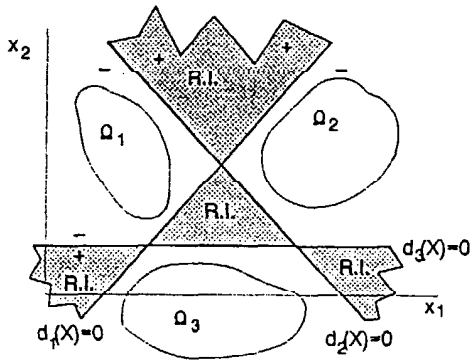


Figura 2.4: Ejemplo de Fronteras de Decisión para un Clasificador según el modelo a.

b) Cada clase se separa de cada una de las c-1 restantes mediante una superficie de decisión. El número de superficies de decisión será:

$$C_2^c = \frac{c(c-1)}{2} \qquad \text{[2.23]}$$

Las fronteras de decisión $d_{ij}(\mathbf{X})$ presenta las siguientes características:

$$\begin{aligned} &\text{I)} \quad \text{Si } \mathbf{X} \in \Omega_i \Leftrightarrow d_{ij}(\mathbf{X}) > 0 \ \ \forall j \neq i \\ &\text{II)} \quad d_{ij}(\mathbf{X}) = -d_{ji}(\mathbf{X}) \end{aligned} \qquad \textbf{[2.24]}$$

Y la regla de decisión es:

$$\text{Si} \quad d_{ij}(\mathbf{X}) > 0 \quad \forall j \neq i \Leftrightarrow \mathbf{X} \in \Omega_i \qquad \textbf{[2.25]}$$

En este caso y en condiciones generales, se presentan tambien regiones indeterminadas, como puede observarse en la figura 2.5. El único caso en el que no existirían dichas regiones sería el muy particular correspondiente a que todas las rectas de decisión intersectasen en un mismo punto.

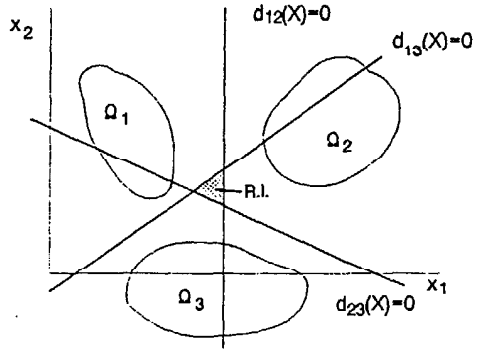


Figura 2.5: Ejemplo de Fronteras de Decisión para un Clasificador según el modelo b.

c) Existen c funciones de decisión $d_i(\mathbf{X})$ con la propiedad de que:

$$Si \quad X \in \Omega_i \Rightarrow d_i(X) > d_j(X) \quad \forall\, i,j=1,\ldots,c;\ j \neq i \qquad [2.26]$$

Existe una relación entre este caso y el anterior (**b**), ya que, a partir de las funciones de la expresión 2.26 podemos definir las superficies de decisión:

$$d_{ij}(X) = d_i(X) - d_j(X) \quad \forall j \neq i \qquad [2.27]$$

Que verifican la condición $d_{ij}(\mathbf{X})>0\ \forall\ i\neq j$, ya que $d_i(X) > d_j(X)\ \forall\ i\neq j$. Luego, si las clases resultan separables bajo estas condiciones, lo son bajo las del caso **b**. Lo contrario no es necesariamente cierto. En este caso, la regla de decisión puede escribirse de la siguiente manera:

$$Si \quad d_i(X) = \max_{\forall j=1,2,\ldots,c} \{d_j(X)\} \Rightarrow X \in \Omega_i \qquad [2.28]$$

El clasificador resultante se denomina *MAQUINA LINEAL*, mostrándose un diagrama operacional del mismo en la figura 2.6.

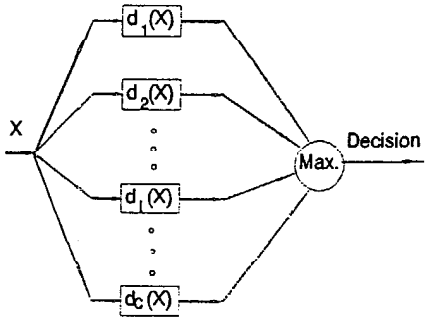


Figura 2.6: Diagrama Operacional de la Máquina Lineal

En la figura 2.7 se muestra un ejemplo bidimensional triclásico que ilustra el modelo de clasificador. La estructura del mismo, analizada desde el punto de vista de los signos de las desigualdades permite determinar las regiones de cada clase. Se observa como la única posible región indeterminada es alguna área triángular definida por las tres rectas, una de las cuales se ilustra en la figura 2.8.

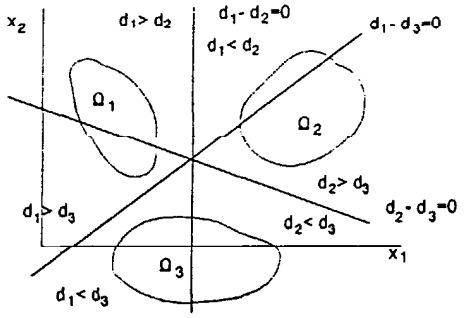


Figura 2.7: Ejemplo de Fronteras de Decisión para un Clasificador según el modelo c.

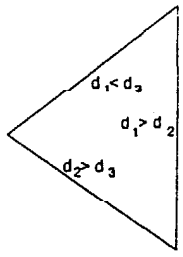


Figura 2.8: Análisis de las desigualdades en una de las posibles regiones de intersección definida por las tres rectas de decisión de la Figura 2.7.

El análisis de las desigualdades en su interior, según cada frontera da lugar a:

$$\begin{aligned} d_1(X) &< d_3(X) \\ d_2(X) &> d_3(X) \\ d_1(X) &> d_2(X) \end{aligned} \qquad [2.29]$$

De la segunda y la tercera desigualdad se obtiene: $d_1(\mathbf{X}) > d_3(\mathbf{X})$, lo que resulta contradictorio con la primera, con lo que podemos concluir que la única posibilidad es que todas las rectas de decisión se intersecten en el mismo punto. Por tanto, con este esquema de clasificador no aparecerán regiones indeterminadas.

2.2.3.- FUNCIONES DISCRIMINANTES GENERALIZADAS

No todos los problemas de clasificación se pueden resolver utilizando fronteras de decisión como las analizadas anteriormente. En efecto, existen configuraciones de clases no linealmente separables, como se muestra en la figura 2.9.

La complejidad de las fronteras puede ir desde el mencionado caso lineal al de hipersuperficies altamente no lineales.

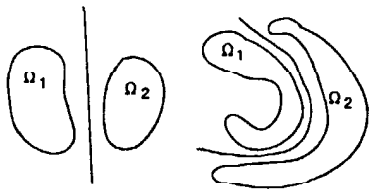


Figura 2.9: A la izquierda, dos regiones linealmente separables. A la derecha, no linealmente separables.

Se puede generalizar convenientemente el funcional de decisión, partiendo de la expresión 2.9, sustituyendo la combinación lineal de las características por la de funcionales genéricos $f_i(\mathbf{X})$ del vector de características n-dimensional, es decir:

$$d(\mathbf{X})=\omega_1 f_1(\mathbf{X})+\omega_2 f_2(\mathbf{X})+\ldots+\omega_k f_k(\mathbf{X})+\omega_0=\sum_{i=0}^{i=k}\omega_i f_i(\mathbf{X}) \qquad [2.30]$$

Donde $f_0(\mathbf{X})=1$ y $\{f_i(\mathbf{X});i=1,\ldots k\}$ son funciones reales unievaluadas. La expresión 2.30 engloba una variedad infinita de funciones de decisión, cuya naturaleza depende de la elección de los funcionales $f_i(\mathbf{X})$ y del número k de términos de la expresión. La función de decisión obtenida presenta una naturaleza no lineal respecto a **X**, sin embargo, es posible expresarla en forma *lineal generalizada* mediante una transformación del espacio de representación. En efecto, definiendo las igualdades:

$$\begin{aligned} &a_i=\omega_i \ ; \ \forall i=0,1,...,k \\ &y_0=1 \\ &y_i=f_i(X) \ ; \ \forall i=1,2,...,k \end{aligned} \qquad [2.31]$$

La función discriminante resulta entonces expresada en base al vector Y como:

$$d(X) = \sum_{i=0}^{k} a_i y_i = a^t Y = (a_0 \ a_1 \ . \ . \ . \ a_k) \begin{pmatrix} y_0 \\ y_1 \\ . \\ . \\ y_k \end{pmatrix} \qquad [2.32]$$

Es decir, en forma lineal. Como se observa, el proceso de linealización se basa en un mapeo desde el espacio original de representación n-dimensional, en el cual la función de decisión es no lineal, a un espacio transformado n^*-dimensional ($n^*=k+1$), en el que el discriminante es lineal, concretamente un hiperplano que pasa por el origen.

Una de las soluciones más estudiadas corresponde al caso en que el discriminante es de naturaleza polinómica, que se genera añadiendo términos de grados superiores al discriminante lineal. El caso más sencillo es el de grado dos, de expresión:

$$d(X) = \omega_0+\sum_{i=1}^{n} \omega_i x_i+\sum_{i=1}^{n}\sum_{i=1}^{n} \omega_{ij} x_i x_j \qquad [2.33]$$

Como $x_i=x_j$, podemos suponer, sin pérdida de generalidad que $\omega_{ij}=\omega_{ji}$. Con ello podemos expresar la función como:

$$d(X)=\omega_0+\sum_{i=1}^{n} \omega_i x_i+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \omega_{ij} x_i x_j+\sum_{i=1}^{n} \omega_{ii} x_i^2 \qquad [2.34]$$

La frontera de decisión $d(X)=0$, correspondiente a esta ecuación, es una superficie hipercuadrática, cuya forma viene definida por sus términos. Así, si definimos las matrices:

$$\begin{aligned} &W=\{\omega_{ij}\} \ ; \ i,j=1,2,...,n \\ &\omega=\{\omega_i\} \ ; \ i=1,2,...,n \end{aligned} \qquad [2.35]$$

Podemos expresar la frontera de decisión como:

$$d(X) = X^t WX + \omega^t X + \omega_0 = 0 \qquad [2.36]$$

Las propiedades de **W** determinan la forma[1] de la frontera de decisión. Así, si **W** es matriz identidad, la frontera es una hiperesfera. Si es definida positiva, describe un hiperelipsoide con los ejes en la dirección de los autovectores de **W**. Si es semidefinida positiva, es un cilindro hiperelipsoidal, cuyas secciones son hiperelipsoides de dimensión inferior cuyos ejes están en la dirección de los autovectores de **W** correspondientes a autovalores no nulos. Por último, si es definida negativa, la frontera es un hiperhiperboloide.

Análogamente a las funciones discriminantes polinómicas de grado dos, se pueden definir otras de cualquier grado superior. Dichas funciones son expresables recursivamente:

$$d^r(X) = \left(\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_r=i_{r-1}}^{n} \omega_{i_1 i_2 \ldots i_r} x_{i_1} x_{i_2} \ldots x_{i_r} \right) + d^{r-1}(X) \qquad [2.37]$$

Donde r indica el grado de no linealidad y:

$$d^0(X) = \omega_0 \qquad [2.38]$$

Ejemplo: Usando la expresión recursiva, podemos generar una función cuadrática (r=2), en un espacio bidimensional (n=2). De 2.37 con r=2 obtenemos:

$$d^2(X) = \sum_{i_1=1}^{2} \sum_{i_2 i_1}^{2} \omega_{i_1 = i_2} x_{i_1} x_{i_2} + d^1(X) \qquad [2.39]$$

Y con r=1:

---

[1] Se dice que W es definida positiva si:

$$X^t WX > 0 \quad \forall X \neq 0$$

Se dice que es semidefinida positiva si:

$$X^t WX \geq 0 \quad \forall X \neq 0$$

Análogamente se puede decir para definida negativa y semidefinida negativa, para los casos respectivos de las desigualdades < y ≤.

$$d^1(X)=\sum_{i_1=1}^{2}\omega_{i_1}x_{i_1}+d^0(X) \qquad [2.40]$$

Con lo que, sustituyendo y desarrollando, obtenemos:

$$d^2(X)=\omega_{11}x_1^2+\omega_{12}x_1x_2+\omega_{22}x_2^2+\omega_1x_1+\omega_2x_2+\omega_0 \qquad [2.41]$$

El número de coeficientes de peso de las funciones de decisión polinómicas crece rápidamente con la dimensionalidad (n) del espacio de representación y con el grado r. Baste para ello hacer recuento en el caso del discriminante lineal, donde el número de términos es de n+1, mientras que en el caso del cuadrático (ecuación 2.34) es:

$$1+n+\frac{n(n-1)}{2}+n=\frac{(n+1)(n+2)}{2} \qquad [2.42]$$

Como caso ilustrativo es válido el de n=2 anterior, donde el discriminante lineal contiene 3 parámetros a determinar, mientras que el cuadrático posee 6. En general, el número de terminos de un discriminante de grado r en un espacio n-dimensional contiene el número de coeficientes definidos por la expresión:

$$N=C_r^{n+r}=\frac{(n+r)!}{r!n!} \qquad [2.43]$$

Se deja al lector la comprobación, por simple sustitución de valores en la ecuación anterior, de cómo se incrementa N con el incremento de los valores de n y r. Un número alto de coeficientes conlleva un incremento de la complejidad computacional del proceso de aprendizaje de dichos términos.

2.3.- CLASIFICACION POR FUNCIONES DE DISTANCIA

La manera más simple e intuitiva de realizar un proceso de clasificación de formas es probablemente la basada en los conceptos de distancia entre la forma a clasificar y conjuntos de prototipos de las clases en consideración. El motivo de ello es que la proximidad o similaridad, en algún sentido, entre vectores de características es una manera obvia de determinar pertenencia a categorías.

En este caso, los vectores de características se consideran como puntos de un espacio con estructura tal que admite la definición de una métrica o pseudométrica.

Además, para obtener resultados satisfactorios, las clases deben ser lo más compactas y separadas entre si, es decir, deben de ser tales que cumplan adecuadamente el tercer postulado de Niemann (alto valor de la relación entre dispersiones interclásicas a dispersiones intraclásicas). Para ilustrar, intuitivamente, la aseveración anterior, basta con observar los ejemplos mostrados en la figura 2.10. En el de la izquierda, que cumple la afirmación, resulta bastante adecuada la conclusión de asignar la muestra incógnita X a la clase Ω_i, en base a su mayor proximidad a las muestras patrón de esta clase. En el de la derecha, sin embargo, una conclusión de este tipo es más arriesgada.

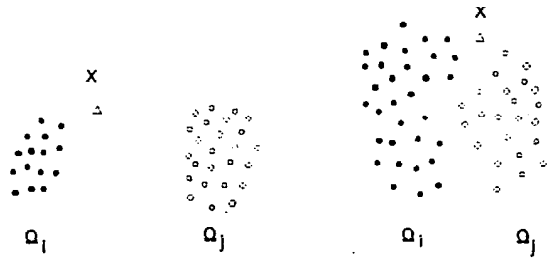


Figura 2.10: A la izquierda, muestra clasificable fácilmente por concepto de proximidad. A la derecha, muestra no fácilmente clasificable.

2.3.1.- SIMILARIDAD Y DISTANCIA

Antes de centrarnos en las reglas de clasificación por distancia, resulta adecuado definir conceptos. Para ello, sea U ($U=\mathbb{R}^n$) un conjunto, finito o infinito, de elementos, que se corresponden con los vectores de características asociados a las muestras de un problema dado. Sea $\mathbb{R}$ el conjunto de los números reales y $\mathbb{R}^+$ el de los reales positivos. Se denomina ***función de distancia*** D a un mapeo $D:U^*U\rightarrow\mathbb{R}$, que para un par arbitrario X, $Y\in U$ cumple:

$$\begin{aligned} &1)\ D(X,Y)\geq D_0 \\ &2)\ D(X,X)=D_0 \\ &3)\ D(X,Y)=D(Y,X) \end{aligned} \qquad [2.44]$$

Donde D_0 es un número real finito arbitrario, que puede ser negativo. La primera propiedad indica que la función de distancia posee una cota inferior, la segunda que la distancia es mínima para el caso de elementos idénticos, y la tercera es la de simetría.

La función de distancia se dice que es ***métrica*** si además:

$$\begin{aligned} &4)\ \ Si\ \ D(X,Y)=D_0 \rightarrow X=Y \\ &5)\ \ \forall X,Y,Z\in U:\ D(X,Z)\le D(X,Y)+D(Y,Z) \end{aligned} \qquad [2.45]$$

La cuarta indica que siempre que el funcional de distancia posea el valor mínimo, los elementos considerados son idénticos y la quinta corresponde a la desigualdad triangular.

Los funcionales de distancia, por su definición, asignan valor numérico a la noción de disimilaridad o lejanía entre dos formas representadas por vectores de características, de manera que una gran similaridad o semejanza entre ambas formas se refleja en un valor pequeño del funcional de distancia.

Si además D_0=0, se tiene el concepto de métrica del análisis funcional. Si D_0 es negativo, se puede construir una distancia métrica como:

$$D'(X,Y)=D(X,Y)-D_0 \qquad [2.46]$$

La más conocida de las métricas es la Euclídea, que corresponde a la generalización a **n** dimensiones de la distancia entre dos puntos es un plano, y derivada de la norma L_2 de un vector, es decir de:

$$\|X\|_2=\sqrt{X^tX}=\sqrt{\sum_{i=1}^{n}x_i^2} \qquad [2.47]$$

La correspondiente distancia entre dos vectores X e Y es:

$$D_2(X,Y)=\|X-Y\|_2=\sqrt{(X-Y)^t(X-Y)}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \qquad [2.48]$$

Esta métrica Euclídea, así como la norma, presenta la propiedad de que sus valores son invariantes respecto a mapeos ortogonales (rotaciones) de los vectores, los cuales se describen por matrices Q de dimensión n*n, tales que $Q^{t*}Q=I$ (matriz identidad). Así, tenemos que:

$$\begin{aligned} \|QX\|_2&=\|X\|_2 \\ D_2(QX,QY)&=D_2(X,Y) \end{aligned} \qquad [2.49]$$

Además, norma y métrica tambien resultan invariantes frente a traslaciones.

La métrica Euclídea puede generalizarse de dos maneras:

1) Basándonos en la norma L_p:

$$\|X\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \quad (p \geq 1) \qquad \textbf{[2.50]}$$

Según la cual, por analogía nos permite definir:

$$D_p(X,Y) = \|X-Y\|_p \qquad \textbf{[2.51.a]}$$

Que constituyen la familia de coeficientes de disimilaridad de Minkowski. De ellos, dos que resultan particularmente interesantes son los siguientes:

a) Distancia Manhatan (tambien conocida como "city blocks", "taxicab" o ciudad"), que corresponde al caso p=1 y cuya expresión es:

$$D_1(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \qquad \textbf{[2.51.b]}$$

b) Distancia de Tablero de Ajedrez o Chebycheff, que corresponde al caso $p=\infty$ y con expresión:

$$D_\infty(X,Y) = \max_{i=1}^{n} |x_i - y_i| \qquad \textbf{[2.51.c]}$$

A efectos comparativos, se muestran en la figura 2.11.a las curvas correspondientes a los puntos del plano que se encuentran a distancias unitarias euclídea, Manhatan y Tablero de Ajedrez respectivamente.

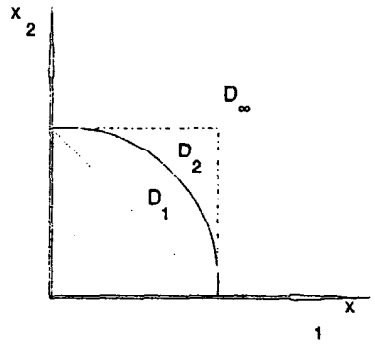


Figura 2.11.a: Curvas correspondientes a los puntos de un plano situados a distancia unitaria del origen de coordenadas según tres de las distancias de Minkowsky

No se conocen otras transformaciones distintas de las traslaciones para los cuales las norma y distancia anteriores resulten invariantes.

2) Definiendo:

$$\|X\|_B=\sqrt{X^tBX} \qquad [2.52]$$

Donde B es una matriz definida positiva.

La métrica correspondiente a 2.52 es:

$$D_B(X,Y)=\sqrt{(X-Y)^tB(X-Y)} \qquad [2.53]$$

En los casos más sencillos B es una matriz diagonal, en la que los elementos de la diagonal se corresponden a diferentes pesos positivos para las componentes del vector.

Una métrica de este tipo que es muy utilizada es la distancia estadística general de Mahalanobis **[CUAD-81]**. Para describirla, supongamos un conjunto de m vectores o muestras de formas n-dimensionales { $\mathbf{X}_k$; k=1,2,...,m }. Denominemos a la componente i-ésima del vector k-ésimo como x_{ik}. Denominemos M a una matriz de dimensiones m*n, en la que cada columna se corresponde con uno de los vectores de características $\mathbf{X}_k$. Sea μ el vector de medias de las componentes o características, es decir:

$$\mu=E[X] \qquad [2.54]$$

Con lo que, cada componente de μ es:

$$\mu_i=\frac{1}{m}\sum_{k=1}^{m} x_{ik} \qquad [2.55]$$

Sea a su vez, Σ la matriz n*n de covarianzas de variables, es decir:

$$\Sigma=E[(X-\mu)(X-\mu)^t]=(\sigma_{ij}) \qquad [2.56]$$

Siendo cada elemento de dicha matriz (σ_{ij}), la covarianza entre las componentes i y j, indicadas por los subíndices, es decir:

$$\sigma_{ij}=\frac{1}{m}\sum_{k=1}^{m}(x_{ik}-\mu_i)(x_{jk}-\mu_j) \qquad [2.57]$$

Si definimos la matriz M^* como:

$$M^*=(x_{ik}-\mu_i) \qquad [2.58]$$

La matriz de covarianzas se puede expresar como:

$$\Sigma=\frac{1}{m}\, M^*(M^*)^t \qquad [2.59]$$

Cuando los vectores X^* son linealmente independientes (algo que se puede asumir cuando m>>n **[SPAT-80]**), la matriz de covarianzas de variables es definida positiva y, por tanto, no singular, con lo cual posee inversa definida positiva Σ^{-1}. Además, Σ^{-1} es una matriz simétrica, ya que $\sigma_{ij}=\sigma_{ji}$ $\forall i,j=1,2,...,n$.

En base a lo anterior, se define la ***Distancia de Mahalanobis*** entre dos muestras X e Y como:

$$D_\Sigma(X,Y)=\sqrt{(X-Y)^t\Sigma^{-1}(X-Y)} \qquad [2.60]$$

Esta distancia cumple las propiedades 2.44 y 2.45 con $D_0=0$. Además, resulta invariante ante cualquier transformación lineal no singular de las variables o características, sobre todas las muestras. Para demostrarlo, sea una matriz de transformación C genérica como las mencionadas, de dimensión n*n, tal que, la relación entre vectores originales (X e Y) y transformados, que denominaremos λ y ξ:

$$\lambda = CX \Rightarrow \lambda^t = X^t C^t$$
$$\xi = CY \Rightarrow \xi^t = Y^t C^t \qquad [2.61]$$

Entonces, la matriz de covarianzas para las muestras transformadas será:

$$\Sigma_C = \frac{1}{m} M_C^* (M_C^*)^t \qquad [2.62]$$

Y dado que:

$$M_C^* = CM^*$$
$$(M_C^*)^t = (M^*)^t C^t \qquad [2.63]$$

Sustituyendo en 2.62 se obtiene:

$$\Sigma_C = \frac{1}{m} CM^* (M^*)^t C^t = C\Sigma C^t \qquad [2.64]$$

Y su inversa:

$$\Sigma_C^{-1} = [C\Sigma C^t]^{-1} = (C^t)^{-1}\Sigma^{-1}C^{-1} \qquad [2.65]$$

Con lo que, la distancia de Mahalanobis entre λ y ξ resulta:

$$D_{\Sigma_C}(\lambda,\xi) = \sqrt{(\lambda-\xi)^t \Sigma_C^{-1} (\lambda-\xi)} =$$
$$\sqrt{(X-Y)^t C^t [(C^t)^{-1}\Sigma^{-1}C^{-1}] C(X-Y)} = \sqrt{(X-Y)^t \Sigma^{-1}(X-Y)} = D_\Sigma(X,Y) \qquad [2.66]$$

c.q.d.

Si, en particular, **C** es una matriz diagonal con elementos no nulos en la diagonal, la transformación de **X** por **C** significa que el valor de cada componente del vector se multiplica por una constante, es decir, la matriz de transformación efectúa un cambio de *escala*. Como se observa en la expresión anterior, aún ante esta transformación, la distancia de Mahalanobis resulta invariante. Hay que hacer notar cómo otras métricas, incluidas las euclideas, no poseen esta importante propiedad.

Como comentario añadido ante este funcional de distancia, se puede decir que está expresada en unidades de desviación típica, y tiene en cuenta las correlaciones (es decir, interdependencia o redundancia) entre las variables,

de forma que la distancia disminuye a medida que aumenta la correlación de las variables. Además, la distancia es un funcional monótono creciente con la dimensionalidad del espacio. Como se puede observar en la figura 2.11.b.

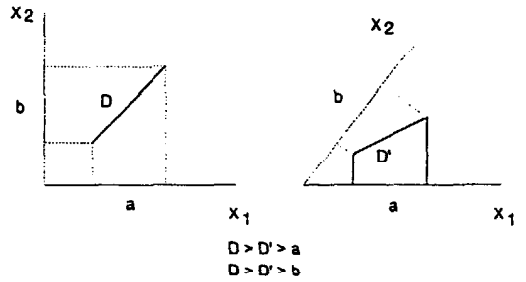


Fig 2.11.b: Distancia de Mahalanobis para características no correlacionadas (a la izquierda) y correlacionadas (derecha).

Es posible tambien definir, en vez de funcionales que asignen valor numérico a la disimilaridad, como son los de distancia anteriormente comentados, funcionales que cuantifique la similaridad, es decir, que presenten mayor valor a mayor similaridad y menor valor a menor similaridad, como son los denominados funcionales de semejanza. Formalmente y por analogía con la función de distancia, se define una *función de semejanza* S para un conjunto U de elementos como un mapeo $S:U^*U \rightarrow \mathbf{R}$ que ,para una par arbitrario X, $Y \in U$ posee las siguientes propiedades:

$$\begin{aligned} &1)\ S(X,Y) \leq S_0 \\ &2)\ S(X,X) = S_0 \\ &3)\ S(X,Y) = S(Y,X) \end{aligned} \qquad [2.67]$$

Siendo S_0 un número real finito arbitrario. La función de semejanza se dice ***métrica*** si, además:

$$\begin{aligned} &4)\ Si\ \ S(X,Y) = S_0 \Rightarrow X = Y \\ &5)\ \forall X,Y,Z \in U : [S(X,Y) + S(Y,Z)].S(X,Z) \geq S(X,Y).S(Y,Z) \end{aligned} \qquad [2.68]$$

La cuarta corresponde a la proposición de que la máxima semejanza sólo pueden poseerla elementos idénticos. La quinta se define estableciendo analogía con la correspondiente de la definición de distancia métrica.

La relación entre semejanza y distancia es, pués, evidente. Así, si D es una función de distancia (métrica) definida en el rango de valores de $\mathbf{R}^+$ o de $\mathbf{R}$, entonces $1/D$ es

una función de semejanza (métrica). Si D es una métrica que está definida en $\mathbb{R}^+$, entonces:

$$e^{-d} \qquad [2.69]$$

Es una función de semejanza también métrica. Por otro lado, si D está definida en un rango finito de valores reales, entonces, son métricas de semejanza:

$$\begin{matrix} \max\{D\}-D \\ \sqrt{\max\{D\}-D} \\ \max\{D\}-D^2 \end{matrix} \qquad [2.70]$$

Ahora bien, las medidas de similaridad no tienen por que limitarse a estar expresadas en función de distancias predefinidas. Por ejemplo, sea la semejanza:

$$S(X,Y)=\frac{X^tY}{\|X\|\,\|Y\|} \qquad [2.71]$$

Que se corresponde con el coseno del ángulo que forman los vectores X e Y, y que es máxima cuando ambos vectores están orientados en la misma dirección respecto al origen del sistema de referencia. En este sentido, resultará útil cuando las clases constituyen regiones alargadas como las mostradas en la figura 2.12.

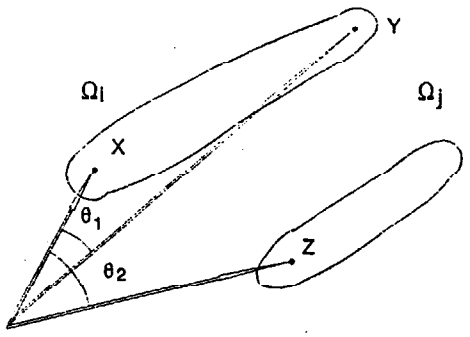


Figura 2.12: Ejemplo de clases adecuadas para la medida de semejanza 2.71.

Se puede observar como, con una semejanza como esta y para las muestras que aparecen en la figura, se cumple que:

$$S(X,Y)=\cos\theta_1=\frac{X^tY}{\|X\|\,\|Y\|}$$
$$S(X,Z)=\cos\theta_2=\frac{X^tZ}{\|X\|\,\|Z\|} \qquad [2.72]$$
$$S(X,Y)>S(X,Z)$$

En ciertos casos, las formas se representan mediante vectores de características con componentes binarias, es decir valuadas en 0 o 1, lo que quiere decir que, si el valor del elemento i del vector es x_i=1, esto indica que la forma posee la propiedad i, mientras que si es 0, carece de ella. En estos casos, una función de semejanza como la 2.71 presenta una interpretación geométrica interesante. Así, el numerador de 2.71 representa el número de atributos que poseen comunes X e Y, mientras que el producto de normas del denominador representa la media geométrica del número de atributos poseidos por uno de los vectores multiplicada por la del otro. Por tanto, la semejanza en este caso puede interpretarse como una medida de los atributos comunes que poseen ambos vectores.

Una variación binaria de la medida anterior, utilizadas en aplicaciones de taxonomía (clasificación de plantas y animales) o en nosología (clasificación de enfermedades infecciosas), es la denominada *medida de Tanimoto*, que viene dada por:

$$S(X,Y)=\frac{X^tY}{X^tX+Y^tY-X^tY} \qquad [2.73]$$

Se deja como ejercicio al lector, el dar una interpretación de esta medida.

2.3.2.- REGLA DE LA DISTANCIA MINIMA

Sea un conjunto de c clases $\{\Omega_1,\Omega_2,\ldots,\Omega_c\}$, donde cada una de las Ω_i resulte representada por un vector de características $\mathbf{Z}_i$, que denominaremos vector prototipo, o ***prototipo*** a secas, de la clase. Sea a su vez un vector de una forma incógnita **X**, que pretendemos clasificar. La clasificación de **X** según la *regla de la mínima distancia a los prototipos* se puede expresar como:

$$\mathbf{X}\in\Omega_i \text{ si: } D(\mathbf{X},\mathbf{Z}_i)<D(\mathbf{X},\mathbf{Z}_j) \quad \forall i,j=1,2,...,c;\ i\neq j \qquad \textbf{[2.74]}$$

Donde D representa al funcional de distancia definido para el espacio de representación.

La regla anterior se puede escribir alternativamente de la siguiente manera:

$$X\in\Omega_i \ st. \quad D(X,Z_i)=\min_{\forall j=1,2,...,c}\{D(X,Z_j)\} \qquad \textbf{[2.75]}$$

La fase de aprendizaje de un sistema con clasificador según la regla de decisión de la distancia mínima consistirá en obtener, a partir de las muestras de aprendizaje, los c prototipos $\mathbf{Z}_i$ que representen a las clases correspondientes. Un vector propotipo muy utilizado es el centroide o vector medio de la clase.

La clasificación por regla de distancia mínima es un caso de clasificación por función discriminante lineal. Así, sea el caso de $E=\mathbb{R}^n$ y métrica Euclídea. Para comprobarlo, partamos de la expresión del cuadrado de la distancia euclídea, cualitativamente análoga a la distancia a secas, y desarrollemos la expresión, es decir:

$$D^2(X,Z_i) = \|X-Z_i\|^2 = \|X\|^2-2Z_i^tX+\|Z_i\|^2 \qquad \textbf{[2.76]}$$

Si definimos:

$$\varphi_i(X)=Z_i^tX-\frac{1}{2}\|Z_i\|^2 \qquad \textbf{[2.77]}$$

El funcional de distancia queda como:

$$D^2(X,Z_i)=\|X\|^2-2\varphi_i(X) \qquad \textbf{[2.78]}$$

Como la norma del vector de la forma incógnita es independiente de la clase i, de 2.75 y 2.78 se puede deducir que, la minimización de la distancia es equivalente a la de maximización del funcional 2.77, con que la regla de clasificación se puede expresar como:

$$X\in\Omega_i \ si: \quad \varphi_i(X)=\max_{\forall j=1,2,...,c}\{\varphi_j(X)\} \qquad \textbf{[2.79]}$$

Que presenta la misma estructura que la regla de clasificación por discriminante lineal de la expresión 2.26. Si además, comparamos el funcional definido en 2.77 con la expresión 2.10, se pueden establecer las equivalencias:

$$\omega_i = Z_i$$
$$\omega_0 = -\frac{1}{2}\|Z_i\|^2 \qquad [2.80]$$

Lo que demuestra la afirmación mencionada: el proceso de clasificación por regla de distancia euclídea mínima a los prototipos de las clases es un caso particular de clasificación basado en función discriminante lineal.

Además, las superficies de decisión que separan las clases, son hiperplanos perpendiculares a los segmentos que unen los puntos del espacio de características que representan a los prototipos correspondientes. Además, dichos hiperplanos bisectan dicho segmento en su punto medio. La demostración de esta afirmación se deja al lector.

Al conjunto de regiones definidas por las fronteras asociadas al clasificador de mínima distancia se las denomina *regiones de Voronoi* (de orden 0).

La equivalencia entre la regla de distancia mínima a los prototipos y la regla basada en discriminantes lineales, tambien se presenta en el caso de que se utilice como métrica la distancia de Mahalanobis. Se propone tambien como ejercicio la demostración de esta afirmación.

2.3.3.- REGLA DEL VECINO MAS PROXIMO

El clasificador de distancia mínima puede generalizarse para permitir más de un prototipo por clase. Así, sea un conjunto de c clases $\{\Omega_1,\Omega_2,\ldots,\Omega_c\}$, donde cada una de las Ω_i resulte representada por un conjunto de vectores prototipos $\{Z_i^1,Z_i^2,\ldots,Z_i^{Ni}\}$, de manera que Z_i^l representa al prototipo l-ésimo de la clase i-ésima. Se dice que una forma incógnita **X** resulta clasificada en la clase i-ésima según la *regla de clasificación del vecino más próximo* (NN rule) si:

$$D(X,Z_i^p) = \min_{l=1,2,\ldots,N_j;\ j=1,2,\ldots,c} \{D(X,Z_j^l)\} \qquad [2.81]$$

Es decir, si en la clase i-ésima existe al menos un prototipo Z_i^p que sea, dentro del conjunto de los prototipos de todas las clases, el más próximo a la muestra incógnita.

Este esquema de clasificación presenta varias ventajas a priori. Por un lado permite, si se desea, definir reglas de clasificación sin esquemas de aprendizaje, ya que es

posible definir como conjunto de prototipos para la clase, al propio conjunto de muestras de aprendizaje. Sin embargo, esta estructura de clasificador es de naturaleza exhaustiva y, por tanto, si el número de muestras de aprendizaje es elevado, el costo computacional en la toma de decisión tambien lo es. Por otro lado, si se realiza una selección de prototipos adecuada, se pueden definir reglas de clasificación en estructuras de clases con formas más complejas de las que permite el clasificador por distancia mínima a prototipo.

Si la métrica definida para realizar la clasificación en el espacio es euclídea, y por analogía por el caso descrito en el apartado anterior, el funcional de distancia puede escribirse como:

$$[D_j^l(X,Z_j^l)]^2=\|X\|^2-2(Z_j^l)^tX+\|Z_j^l\|^2=\|X\|^2-2\varphi_j^l(X) \qquad \textbf{[2.82]}$$

Donde $\varphi_i^l(\mathbf{X})$ tiene estructura de discriminante lineal en $\mathbf{X}$ y, por tanto se puede reescribir la regla de clasificación como:

$$X\in\Omega_i si:\ \varphi_i^D(X)=\max_{l=1,2,...,N_j;\ j=1,2,...c}\{\varphi_j^l(X)\} \qquad \textbf{[2.83]}$$

Como se observa, para cada clase i existen N_i discriminantes asociados, y las superficies de decisión entre cada dos clases no seran en este caso hiperplanos como en el caso de la distancia mínima, sino que dicha superficie estará constituida por diversos hiperplanos, constituyendo una superficie hiperpoliédrica. Por este motivo, el discriminante obtenido según la regla NN se le denomina *función discriminante lineal a intervalos* (piecewise-linear).

El esquema de clasificación según la regla del vecino más próximo puede modificarse en el sentido de que la regla no suministre el prototipo más cercano a la muestra incógnita, sino el conjunto de prototipos más cercanos. Este tipo de regla es la denominada *regla de los K-vecinos más próximos* (K-NN rule), la cual suministra los K prototipos más próximos y a continuación, según un criterio de mayoría entre los K resultados, obtener la clasificación de la muestra incógnita.

Esta regla es útil en ciertas situaciones en que las muestras de clases diferentes se encuentran muy próximas. La regla NN suministra resultados más fiables que la K-NN sólo si las distancias entre muestras de la misma clase son más pequeñas que las distancias entre muestras de diferentes clases.

2.4.- LA DECISION COMO PROBLEMA ESTADISTICO PARAMETRICO

En este apartado abordaremos el problema de la definición de reglas de decisión desde una aproximación estadística. En este caso, se considera a los vectores de características como variables aleatorias n-dimensionales y a las clases de formas, distribuidas según densidades de probabilidad. La solución será obtener reglas de clasificación óptimas en el sentido de minimizar determinadas tasas relacionadas con la clasificación errónea.

2.4.1.- DECISION EN BASE A PROBABILIDADES A PRIORI Y POSTERIORI

Sea que nos planteamos definir una regla de clasificación de formas entre dos clases Ω_1 y Ω_2 partiendo de un vector de medidas **X**. Sea que es conocida la probabilidad a priori de que, una muestra pertenezca a una de las clases $P(\Omega_1)$ o a otra $P(\Omega_2)$. Ambas probabilidades están ligadas por la relación $P(\Omega_1)+P(\Omega_1)=1$.

Si hay que decidir la clasificación del vector incógnita **X** en una de las clases sin analizarlo, la regla de decisión posible en base a los datos disponibles es la denominada *regla de decisión en base a las probabilidades a priori*:

$$\begin{array}{l} \textit{Si } P(\Omega_1)>P(\Omega_2) \textit{ Entonces } X\in\Omega_1 \\ \textit{Si } P(\Omega_1)<P(\Omega_2) \textit{ Entonces } X\in\Omega_2 \end{array} \qquad [2.84]$$

Con este esquema, siempre se tomará la misma decisión, independientemente de **X** y la *probabilidad de error* vendrá definida por la menor de las probabilidades a priori.

Ahora bien, normalmente en los problemas de Reconocimiento de Formas es posible la observabilidad del vector **X**, con lo cual se puede disponer de más información para la toma de decisión. Sea que al diseñar se disponen, o es posible estimar, las densidades de probabilidad de las clases: $p(\mathbf{X}/\Omega_1)$, que es la probabilidad condicional de que una forma incógnita, perteneciendo a la clase Ω_1, poseea en su vector de medidas el valor **X** y $p(\mathbf{X}/\Omega_2)$, lo correspondiente para Ω_2. Así, para un valor dado de **X**, la diferencia entre estas dos densidades de probabilidad recoge la discriminación entre las clases, siempre que **X** haya sido seleccionada correctamente y contenga toda la información relevante relativa al problema.

Sea además p(**X**) la probabilidad de que el vector de características posea el valor **X**, independientemente de su clase de pertenencia. En base a las probabilidades a priori, las densidades de probabilidad de clase y la probabilidad de que el vector posea un valor concreto, es posible determinar las probabilidades a posteriori $P(\Omega_i/\mathbf{X})$, es decir las probabilidades de que, poseyendo la forma incógnita a **X** por valor de su vector de características, pertenezca a la clase Ω_i. Dicha determinación se puede realizar a partir del teorema de Bayes:

$$P(\Omega_i/X)=\frac{p(X/\Omega_i)P(\Omega_i)}{p(X)};\ \forall i=1,2 \qquad [2.85]$$

Las probabilidades a posteriori de este problema biclásico están ligadas por la expresión:

$$P(\Omega_1/X)+P(\Omega_2/X)=1 \qquad [2.86]$$

Con las probabilidades obtenidas en 2.85 podemos definir la denominada *regla de decisión en base a las probabilidades a posteriori*, que se puede expresar:

$$\begin{aligned} &Si\ P(\Omega_1/X)>P(\Omega_2/X)\ \ Entonces\ X\in\Omega_1 \\ &Si\ P(\Omega_1/X)<P(\Omega_2/X)\ \ Entonces\ X\in\Omega_2 \end{aligned} \qquad [2.87]$$

Podemos sustituir las probabilidades a posteriori según 2.85. Se puede observar como p(**X**) es factor común a ambas probabilidades a posteriori, no dependiendo de i, por lo que la regla la podemos simplificar de la siguiente manera:

$$\begin{aligned} &Si\ p(X/\Omega_1)P(\Omega_1)>p(X/\Omega_2)P(\Omega_2)\ \ Entonces\ X\in\Omega_1 \\ &Si\ p(X/\Omega_1)P(\Omega_1)<p(X/\Omega_2)P(\Omega_2)\ \ Entonces\ X\in\Omega_2 \end{aligned} \qquad [2.88]$$

Si ocurre que las densidades de probabilidad de las dos clases son iguales, la decisión se encuentra en el valor de probabilidades a priori.

Si observamos la estructura, tanto de la regla de clasificación 2.87 como de su equivalente 2.88 y la comparamos con la expresión 2.1, los términos se corresponden perfectamente al concepto de función discriminante. Así para la 2.88 podemos poner:

$$d_i(X)=p(X/\Omega_i)P(\Omega_i);\ \ \forall i=1,2 \qquad [2.89]$$

Con lo que la regla de decisión en base a probabilidades a posteriori se puede expresar en base a las funciones discriminantes probabilísticas como:

$$\begin{array}{l} Si\ d_1(X)>d_2(X)\ \ Entonces\ X\in\Omega_1 \\ Si\ d_1(X)<d_2(X)\ \ Entonces\ X\in\Omega_2 \end{array} \qquad [2.90]$$

La desigualdad de 2.87 se suele expresar despejando a un lado de la desigualdad las densidades de probabilidad y a otro, las probabilidades a priori. Así, se define como *relación de verosimilitud* l(**X**) (llikelihood ratio) al cociente:

$$l(X)=\frac{p(X/\Omega_1)}{p(X/\Omega_2)} \qquad [2.91]$$

Y como *valor umbral* (threshold value) θ a:

$$\theta=\frac{P(\Omega_2)}{P(\Omega_1)} \qquad [2.92]$$

Con lo que la regla de decisión queda como:

$$\begin{array}{l} Si\ l(X)>\theta\ \ Entonces\ X\in\Omega_1 \\ Si\ l(X)<\theta\ \ Entonces\ X\in\Omega_2 \end{array} \qquad [2.93]$$

Podemos analizar a continuación las tasas de error en el clasificador de la expresión 2.89. La función de probabilidad condicional de error cometido para los valores de **X** será:

$$P(\varepsilon/X)=\begin{cases} P(\Omega_1/X)\ \ si\ se\ decide\ clase\ \Omega_2 \\ P(\Omega_2/X)\ \ si\ se\ decide\ clase\ \Omega_1 \end{cases} \qquad [2.94]$$

Se observa como el error, en cada caso, viene definido por la menor de las probabilidades a posteriori paxa un **X** dado. Por ello a este esquema de clasificación se le denomina tambien *regla bayesiana de mínimo error.*

Denominemos Γ_1 a la región del espacio de representación donde se cumple que $P(\Omega_1/\mathbf{X})>P(\Omega_2/\mathbf{X})$ y Γ_2 a la región donde se cumple $P(\Omega_1/\mathbf{X})<P(\Omega_2/\mathbf{X})$. La probabilidad media de error que se comete al efectuar una clasificación es:
Es decir, la probabilidad promedio de error consta de dos términos: un primero correspondiente a clasificación errónea de muestras de Ω_1 y otro correspondiente a la clasificación errónea de muestras de Ω_2.

$$P(\varepsilon)=\int_{\Gamma} P(\varepsilon/X)P(X)dX=\int_{\Gamma}\min_{i}[P(\Omega/X)]P(X)dX=$$
$$\int_{\Gamma_2} P(\Omega_1/X)P(X)dX+\int_{\Gamma_1} P(\Omega_2/X)P(X)dX= \qquad [2.95]$$
$$P(\Omega_1)\int_{\Gamma_2} p(X/\Omega_1)dX+P(\Omega_2)\int_{\Gamma_1} p(X/\Omega_1)dX=P(\Omega_1)\varepsilon_1+P(\Omega_2)\varepsilon_2$$

El análisis anterior del clasificador bayesiano de mínimo error para el caso biclásico se puede extender fácilmente al caso multiclásico. Así, si el problema que se plantea es clasificar una muestra incógnita entre c clases, la regla puede escribirse, a partir de las probabilidades a posteriori:

$$Si\ P(\Omega_i/X)=\max_{\forall j=1,2,...,c}\{P(\Omega_j/X)\}\ Entonces\ X\in\Omega_i \qquad [2.96]$$

Así si, definimos la función discriminante probabilística asociada a cada clase como:

$$d_i(X)=p(X/\Omega_i)P(\Omega_i);\ \forall i=1,2,...,c \qquad [2.97]$$

La regla puede tambien escribirse de la siguiente manera:

$$Si\ d_i(X)=\max_{\forall j=1,2,...c}\{d_j(X)\}\ Entonces\ X\in\Omega_i \qquad [2.98]$$

La probabilidad de error, que será mínima para cada decisión como hemos visto, de asignar **X** a la clase Ω_i será:

$$P_i(\varepsilon/X)=\sum_{\forall j\neq i} P(\Omega_j/X)=1-P(\Omega_i/X) \qquad [2.99]$$

En base a las cuales, la regla puede expresarse en la siguiente manera:

$$Si\ P_i(\varepsilon/X)=\min_{\forall j=1,2,...,c}\{P_j(\varepsilon/X)\}\ Entonces\ X\in\Omega_i \qquad [2.100]$$

Un esquema funcional del clasificador bayesiano de mínimo error según la expresión 2.98 se muestra en la figura 2.13.

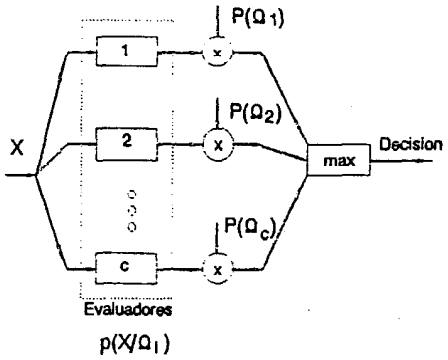


Figura 2.13: Esquema Funcional del Clasificador Bayesiano de mínimo error.

Del mismo modo que se definen funciones discriminantes porbabilísticas, asociadas a cada una de las clases de un problema dado, tambien se definen superficies de decisión, como frontera de separación entre dos clases, según lo descrito en la ecuación 2.3. Así, la superficie de decisión bayesiana entre las clases i y j será:

$$d_{ij}(X) = d_i(X) - d_j(X) = p(X/\Omega_i)P(\Omega_i) - p(X/\Omega_j)P(\Omega_j) = 0 \qquad [2.101]$$

Que se corresponden con las hipersuperficies en el espacio de representación por la igualdad entre las densidades de probabilidad de las clases en consideración, contrapesadas por las correspondientes probabilidades a priori. Dos ejemplos de ello se muestran en las figuras 2.14a y 2.14b, que se corresponden a las fronteras de decisión en un caso unidimensional triclásico y en uno bidimensional biclásico.

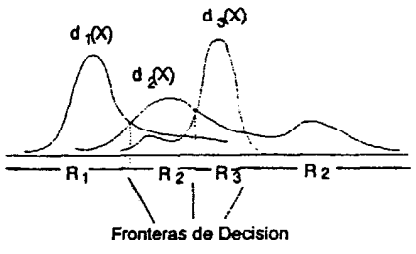


Figura 2.14a: Ejemplo de Fronteras de Decisión en un Caso triclásico monodimensional

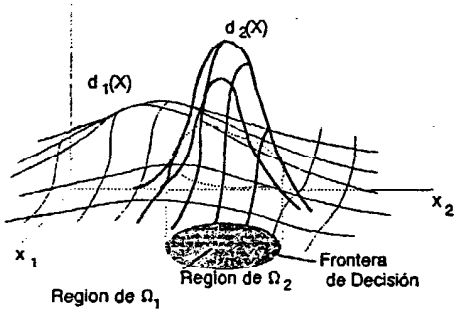


Figura 2.14b: Ejemplo de Fronteras de Decisión en un Caso Bidimensional Biclásico

Por último, hay que hacer notar que muchas veces, a efectos de simplificar la forma de la función discriminante, la ecuación 2.97 se suele expresar logarítmicamente, es decir:

$$d_i(X)=\ln p(X/\Omega_i)+\ln P(\Omega_i);\ \forall i=1,2,...,c \qquad [2.102]$$

Lo que no afecta en absoluto a la forma de la regla de decisión y hace más tratable el manejo de las densidades de probabilidad más comunes.

2.4.2.- CLASIFICACION Y TEORIA DE JUEGOS

El problema de la toma de decisión en un sistema de Reconocimiento de Formas puede tratarse como un juego estadístico, cuyos contendientes son el clasificador, que llamaremos jugador B y otro jugador, que denominaremos A y podemos hacer corresponder con el concepto "naturaleza". La tarea del clasificador es pues, encontrar una decisión óptima que minimice su costo o riesgo promedio como jugador A. Un juego se caracteriza por una serie de reglas con una estructura formal que gobiernan el comportamiento de los jugadores.

Un juego se denomina de *suma cero* (zero sum) si lo que gana un jugador es igual a lo que pierde el otro. Formalmente, un juego **G** se define como una tripleta:

$$G=(Y,Z,L) \qquad [2.103]$$

Donde **Y,Z** son conjuntos arbitrarios, cuyos elementos son las estrategias, respectivamente, de los jugadores A y B, y **L** es una función numérica acotada, definida en el espacio producto cartesiano **Y*Z** de pares (y,z), con $y\varepsilon Y$ y $z\varepsilon Z$, denominada *función de pérdida o ajuste*.

Un juego se denomina *finito* si los conjuntos de estrategias **Y** y **Z** son finitos:

$$Y=(y_1,y_2,...,y_m) \quad Z=(z_1,z_2,...,z_n) \qquad [2.104]$$

En este caso, **Y*Z** es el conjunto de pares:

$$Y*Z=[(y_1,z_1),(y_1,z_2),...,(y_1,z_n),...,(y_m,z_1),(y_m,z_2),...,(y_m,z_n)] \qquad [2.105]$$

Y la función de pérdida **L** da lugar a una matriz, denominada matriz de pérdida o ajuste, de dimensiones m*n, donde el término:

$$L_{ij}=L(y_i,z_j) \qquad [2.106]$$

Representa la pérdida en la que se incurre si el jugador B escoge la estrategia z_j cuando el A escoge la y_i. Por convención, una pérdida positiva representa una pérdida real, y una pérdida nula o negativa se considera una ganancia.

En el caso que nos ocupa, las estrategias del conjunto **Y** son los "estados" de la naturaleza, y se corresponden con las clases de formas Ω_i, mientras que las estrategias **Z** del jugador B son las decisiones del clasificador. Asumimos que el número de decisiones posibles para cada jugador es igual al de posibles clases.

Cada vez que se *juega*, la naturaleza selecciona una estrategia Ω_i, de acuerdo con la probabilidad a priori $P(\Omega_i)$, pero el clasificador no conoce dicha estrategia, solo conoce el vector de características incógnita **X**. La tarea del clasificador consiste en determinar, en base a la información de que dispone, las estrategia de A, es decir, la clase a la que pertenece **X**.

Antes de continuar es interesante que hagamos algunas reflexiones acerca de las diferencias entre este juego clasificador-naturaleza y un juego normal:

1.- La naturaleza no es un *oponente inteligente*, es decir, no escoge deliberadamente sus estrategias para maximizar las pérdidas del clasificador.

2.- Hay posibilidad de "espiar" a la naturaleza diseñando experimentos que nos permitan adquirir información acerca de las *técnicas* que utiliza la naturaleza para seleccionar sus estrategias en la tarea de diseño del clasificador.

2.4.3.- CLASIFICADOR BAYESIANO DE MINIMO RIESGO

Supongamos un problema de clasificación entre c clases planteado como un juego de suma cero como el descrito en el punto anterior. La naturaleza selecciona la clase Ω_k y produce a la entrada del clasificador una forma **X**. Para la asignación de **X** a cada una de las posibles clases, se puede determinar la probabilidad a posteriori correspondiente $P(\Omega_j/\mathbf{X})$, como quedó expuesto en el punto 2.4.1.

Perteneciendo **X** a la clase i, como hemos dicho, si el clasificador decide que pertenece a la clase j, incurre en una pérdida L_{ij}. Ahora bien, como existen c clases, el vector incógnita puede pertenecer a cualquiera de ellas, por lo que la *pérdida promedio esperada* o *riesgo* de asignar la muestra incógnita a la clase j será:

$$r_j(X)=\sum_{i=1}^{c} L_{ij}P(\Omega_i/X) \qquad [2.107]$$

Podemos pues diseñar un clasificador que calcule los riesgos asociados a tomar cada una de las c decisiones para la muestra incógnita **X**, y que tome una decisión en base a que el riesgo de la misma sea el mínimo, es decir, una regla como:

$$Si\ r_j(X)=\min_{\forall l=1,2,\ldots,c}\{r_l(X)\}\ Entonces\ X\in\Omega_j \qquad [2.108]$$

Esta estructura es la correspondiente al denominado *clasificador bayesiano de riesgo mínimo*. Sustituyendo la expresión 2.85 en la ecuación 2.101, el riesgo esperado resulta:

$$r_j(X)=\frac{1}{p(X)}\sum_{i=1}^{c} L_{ij}p(X/\Omega_i)P(\Omega_i) \qquad [2.109]$$

Como p(**X**) es común a todas las expresiones de riesgo, se puede eliminar de la expresión anterior al aplicar la regla de 2.108.

Por último hacemos notar que la estructura de la regla 2.108 es tal que se puede asimilar el funcional de riesgo a una función discriminante, análogamente a lo comentado para el clasificador bayesiano de mínimo error.

A continuación se incluyen análisis complementarios de los clasificadores biclásico y multiclásico según el esquema del mínimo riesgo.

a) ANALISIS DEL CLASIFICADOR BAYESIANO BICLASICO

Siendo c=2, los riesgos de escoger Ω_1 y Ω_2 son respectivamente:

$$\begin{aligned} r_1(X) &= L_{11}p(X/\Omega_1)P(\Omega_1)+L_{21}p(X/\Omega_2)P(\Omega_2) \\ r_2(X) &= L_{12}p(X/\Omega_1)P(\Omega_1)+L_{22}p(X/\Omega_2)P(\Omega_2) \end{aligned} \qquad [2.110]$$

El clasificador asignará **X** a Ω_1 si $r_1(X)<r_2(X)$, es decir si:

$$(L_{21}-L_{22})p(\mathbf{X}/\Omega_2)P(\Omega_2)<(L_{12}-L_{11})p(\mathbf{X}/\Omega_1)P(\Omega_1) \qquad [2.111]$$

O lo que es lo mismo, si:

$$\frac{P(\mathbf{X}/\Omega_1)}{p(\mathbf{X}/\Omega_2)} > \frac{L_{21}-L_{22}}{L_{12}-L_{11}}\frac{P(\Omega_2)}{P(\Omega_1)} \qquad [2.112]$$

Así, como según 2.91, el antecedente de la desigualdad es la denominada *relación de verosimilitudes*, y si al consecuente lo denominamos, análogamente a lo expuesto en el punto 2.4.1, como *valor umbral* θ, es decir:

$$\theta = \frac{L_{21}-L_{22}}{L_{12}-L_{11}}\frac{P(\Omega_2)}{P(\Omega_1)} \qquad [2.113]$$

La regla de decisión, respetando estas equivalencias, se puede escribir idénticamente igual a la 2.93. Como normalmente $L_{21}>L_{22}$ y $L_{12}>L_{11}$, es decir: *la pérdida por asignación incorrecta es mayor que la pérdida por asignación correcta*, el vector incógnita **X** se asignará a la clase Ω_1 si la relación de verosimilitudes supera un cierto valor umbral, lógicamente siempre positivo e independiente de la observación de **X**.

b) ANALISIS DEL CLASIFICADOR BAYESIANO MULTICLASICO

En un caso general con c clases, la regla de clasificación 2.108 se puede escribir:

$$Si \sum_{k=1}^{c} L_{kj}p(X/\Omega_k)P(\Omega_k) < \sum_{q=1}^{c} L_{qi}p(X/\Omega_q)P(\Omega_q) \;;\quad \forall j=1,2,...,c;\ j\neq i \qquad [2.114]$$

$$Entonces\ X\in\Omega_j$$

Con argumentos similares a los del caso biclásico podemos expresar esta ecuación en función de *relaciones de verosimilitudes* $l_{ij}(\mathbf{X})$ y *valores umbrales* θ_{ij}, cuyas expresiones son, respectivamente:

$$l_{ij}(X)=\frac{p(X/\Omega_i)}{p(X/\Omega_j)} \;;\quad \theta_{ij} = \frac{L_{ji}-L_{jj}}{L_{ij}-L_{ii}}\frac{P(\Omega_j)}{P(\Omega_i)} \qquad [2.115]$$

Sin embargo, el caso multiclásico se explica mejor utilizando una función de pérdida específica. En muchos problemas de Reconocimiento de Formas, la pérdida es nula

para decisiones correctas, y un valor fijo distinto de cero para decisiones erróneas. Así, por ejemplo, podemos definir:

$$L_{ij}=1-\delta_{ij} \qquad \textbf{[2.116]}$$

Donde δ_{ij} representa la delta de Kronecker. A esta función se la denomina *función de pérdida simétrica o cero-uno*. Sustituyendo la expresión anterior en la del riesgo esperado se obtiene:

$$r_j(X)=\sum_{i=1}^{c}(1-\delta_{ij})p(X/\Omega_i)P(\Omega_i)=\sum_{i=1}^{c}p(X/\Omega_i)P(\Omega_i)-\sum_{i=1}^{c}\delta_{ij}p(X/\Omega_i)P(\Omega_i)= \qquad \textbf{[2.117]}$$
$$p(X)-p(X/\Omega_j)P(\Omega_j)$$

Así, este clasificador bayesiano de mínimo riesgo asignará una forma **X** a Ω_i si:

$$p(X)-p(X/\Omega_i)P(\Omega_i) < p(X)-p(X/\Omega_j)P(\Omega_j);\quad \forall j=1,2,...,c;\quad j\neq i \qquad \textbf{[2.118]}$$

O, lo que es lo mismo, si:

$$p(X/\Omega_i)P(\Omega_i) > p(X/\Omega_j)P(\Omega_j);\quad \forall j=1,2,...,c;\quad j\neq i \qquad \textbf{[2.119]}$$

Que es exactamente la regla de decisión 2.98, es decir: la regla de decisión basada en las probabilidades a posteriori es la misma que la del clasificador bayesiano de mínimo riesgo con función de pérdida cero-uno.

2.4.4.- ESTUDIO DE CASO: DISTRIBUCION NORMAL

La estructura de los clasificadores bayesianos resulta determinada, en principio, por la forma de las densidades condicionales $p(\mathbf{X}/\Omega_i)$. De las diferentes funciones estudiadas, ninguna ha recibido tanta atención como la densidad normal multivariante, fundamentalmente debido a su tratabilidad analítica. Sin embargo, este modelo resulta apropiado para una situación muy común en los problemas de Reconocimiento de Formas: el caso en el que, los vectores de características **X** para una clase Ω_i pertenecen a un dominio contínuo de valores, y corresponden a versiones, afectadas por ruido, de un vector prototipo μ_i. Esta situación corresponde a aquellos casos en los que, el extractor de características se haya diseñado de manera que se extraigan características cuyo valor sea diferente para muestras de diferentes clases y similares para muestras de la misma clase.

En este punto vamos a analizar la densidad normal multivariante, concentrándonos fundamentalmente en lo correspondiente a los problemas de clasificación.

a) DISTRIBUCION NORMAL UNIVARIANTE

La densidad de probabilidad univariante (es decir, unidimensional) presenta la forma:

$$p(x) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad [2.120]$$

Siendo:

$$\mu = E[x] = \int_{-\infty}^{\infty} x p(x) dx$$

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \qquad [2.121]$$

La densidad normal univariante resulta completamente especificada por dos parámetros: la media μ y la varianza σ^2. Por ello, normalmente se suele expresar una cierta densidad de probabilidad normal en forma reducida como $N(\mu,\sigma^2)$.

Las muestras distribuidas según la densidad normal se suelen agrupar alrededor de la media, con una dispersión alrededor de ella proporcional a la desviación típica σ,

ubicándose aproximadamente el 95% de las muestras de la población en el intervalo $|x-\mu| \leq 2\sigma$.

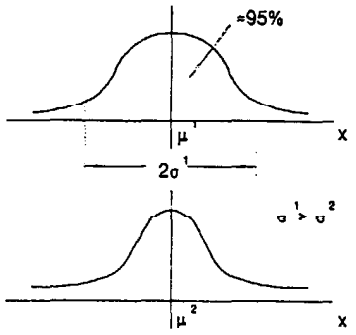


Figura 2.15: Gráficas de dos Distribuciones Normales Univariantes

b) DISTRIBUCION NORMAL MULTIVARIANTE

La densidad de probabilidad normal multivariante tiene la forma:

$$p(\boldsymbol{X}) = N(\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\, e^{\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right]} \qquad [2.122]$$

Siendo **X** un vector de características n-dimensionales, **μ** el vector media y Σ la matriz de covarianzas de variables, de dimensión n*n. Análogamente al caso univariante, la densidad normal multivariante se suele representar en forma reducida como $p(\mathbf{X})=N(\boldsymbol{\mu},\Sigma)$, y resulta completamente definida por n+n(n+1)/2 parámetros que son: los elementos del vector de medias y los elementos independientes de la matriz de covarianzas, que es una matriz simétrica y definida positiva.

Las muestras que constituyen una población normal tienden a situarse en una nube o agrupamiento (cluster), cuyo centro queda determinado por el vector de medias y cuya forma viene definida por la matriz de covarianzas. El lugar de los puntos de densidad de probabilidad constante constituyen hyperelipsoides del espacio de representación, centrados en el punto definido por el vector de medias y para los cuales la forma cuadrática:

Como la regla de decisión a utilizar es la 2.98, el término no dependiente de i es

$$(X-\mu)^t\Sigma^{-1}(X-\mu) \qquad [2.123]$$

Es constante. Si observamos, la misma se corresponde con la *Distancia de Mahalanobis*, introducida en el punto 2.3.1, por tanto, se puede decir que: *los puntos de igual densidad de probabilidad se encuentran a la misma distancia de Mahalanobis de la media*. Además, los ejes principales de estos hiperelipsoides son los autovectores de la matriz de covarianzas, y las longitudes de sus ejes están definidas por los autovalores.

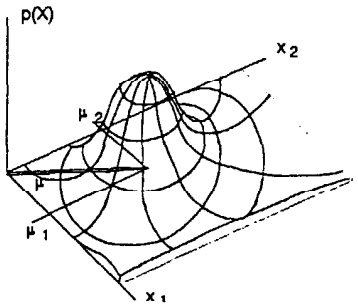


Figura 2.16a: Ejemplo de Densidad Normal Bivariante

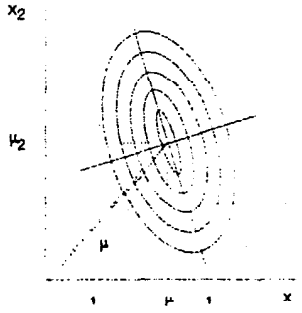


Figura 2.16b: Diagrama de Dispersiones en el Plano de Características, indicando Curvas de Isodensidad

c) FUNCIONES DISCRIMINANTES Y DENSIDAD DE PROBABILIDAD NORMAL

Abordamos en este punto el diseño y análisis de un clasificador bayesiano de mínimo error en un problema multiclásico (c clases) y multivariante (dimensión n). Sea que las probabilidades a priori de las clases son $\{P(\Omega_i);i=1,2,...,c\}$ conocidas, y que las densidades de probabilidad de las mismas se rigen por ley normal, es decir:

$$p(X/\Omega_i)=N(\mu_i,\Sigma_i)=\frac{1}{\sqrt{(2\pi)^n|\Sigma_i|}}e^{[-\frac{1}{2}(X-\mu_i)^t\Sigma_i^{-1}(X-\mu_i)]} \qquad [2.124]$$

Dada la naturaleza exponencial de la función de densidad, podemos definir la función discriminante asociada a cada clase según la expresión 2.102, con lo que obtenemos:

$$d_i(X)=\ln P(\Omega_i)+\ln p(X/\Omega_i)= \\ -\frac{1}{2}(X-\mu_i)^t\Sigma_i^{-1}(X-\mu_i)-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma_i|+\ln P(\Omega_i) \qquad [2.125]$$

común a todas las funciones discriminantes, por lo que podemos eliminarlo. Con ello la función discriminante queda:

$$d_i(X)=-\frac{1}{2}(X-\mu_i)^t\Sigma_i^{-1}(X-\mu_i)-\frac{1}{2}\ln|\Sigma_i|+\ln P(\Omega_i) \qquad \textbf{[2.126]}$$

A continuación analizaremos la clasificación para diferentes casos particulares, relacionados con formas específicas de la matriz de dispersiones.

I) CASO DE CARACTERISTICAS ESTADISTICAMENTE INDEPENDIENTES CON IDENTICA VARIANZA

Este caso corrresponde a:

$$\Sigma_i=\sigma^2 I; \quad \forall i=1,2,...,c \qquad \textbf{[2.127]}$$

Donde I representa la matriz identidad de dimensión n*n. Geométricamente las muestras de las clases se situan en agrupamientos hiperesféricos de igual tamaño, alrededor del vector media de cada clase. Esta matriz de covarianzas tiene como determinante e inversa a:

$$|\Sigma_i|=\sigma^{2n} \;;\; \Sigma_i^{-1}=\left(\frac{1}{\sigma^2}\right)I \qquad \textbf{[2.128]}$$

Con ellos, la expresión 2.126 del discriminante resulta:

$$d_i(X)=-\frac{1}{2}(X-\mu_i)^t\left(\frac{1}{\sigma^2}\right)I(X-\mu_i)-n\ln\sigma+\ln P(\Omega_i)= \\ -\frac{1}{2\sigma^2}(X-\mu_i)^t(X-\mu_i)-n\ln\sigma+\ln P(\Omega_i) \qquad \textbf{[2.129]}$$

En el primer sumando aparece la distancia Euclídea y, por otro lado tambien hay un sumando que resulta independiente de la clase, por tanto, el discriminante puede ponerse como:

$$d_i(X) = -\frac{\|X-\mu_i\|^2}{2\sigma^2} + \ln P(\Omega_i) \qquad [2.130]$$

Si además, las probabilidades a priori de todas las clases son iguales, el segundo sumando puede eliminarse, con lo que la función discriminante resulta:

$$d_i(X) = -\|X-\mu_i\|^2 \qquad [2.131]$$

La regla 2.98 asigna la muestra a la clase que maximiza el discriminante, o lo que es lo mismo, a la que minimiza la distancia Euclídea de **X** a su media. Por tanto, en este caso, *la clasificación se realiza por el criterio de distancia mínima*. Por otro lado, la expresión 2.130 tiene naturaleza de discriminante lineal así que, con un razonamiento análogo al utilizado para la regla de la distancia mínima en el punto 2.3.2, dicho discriminante queda como:

$$d_i(X) = \omega_i^t X + \omega_{i0} \qquad [2.132]$$

Con:

$$\omega_i = \frac{1}{\sigma^2}\mu_i$$

Y:

$$\omega_{i0} = -\frac{1}{\sigma^2}\mu_i^t\mu_i + \ln P(\Omega_i)$$

La frontera de decisión $d_{ij}(\mathbf{X})=0$ entre dos clases i y j es, por tanto, un hiperplano ortogonal al vector que une las medias de ambas clases. Si σ^2 es pequeña, en relación a la distancia Euclídea entre ambas medias, la posición de la frontera de decisión es relativamente insensible a las probabilidades a priori de las clases $P(\Omega_i)$ y $P(\Omega_j)$. Un ejemplo correspondiente a este caso se muestra en la Figura 2.17.

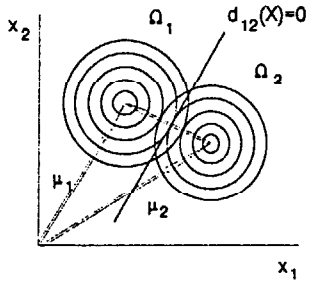


Figura 2.17: Ejemplo de Clases con Variables Estadísticamente Independientes e Idéntica Varianza

II) CASO DE CLASES CON IDENTICA MATRIZ DE COVARIANZAS

Que corresponde a:

$$\Sigma_i=\Sigma; \quad \forall i=1,2,....,c \qquad [2.133]$$

Geométricamente, las muestras se situan en agrupamientos hiperelipsoidales de igual tamaño y forma, estando centrado el agrupamiento de la clase Ω_i en la media de su clase μ_i. Las funciones discriminantes resultan:

$$d_i(X)=-\frac{1}{2}(X-\mu_i)^t\Sigma^{-1}(X-\mu_i)-\frac{1}{2}\ln|\Sigma|+\ln P(\Omega_i) \qquad [2.134]$$

El segundo sumando se puede eliminar al no depender de i. Además , si todas las probabilidades a priori de las clases son iguales, el discriminante se puede poner como:

$$d_i(X)=-(X-\mu_i)^t\Sigma^{-1}(X-\mu_i) \qquad [2.135]$$

Con lo que la regla de clasificación en base al máximo valor del discriminante se puede sustituir por la de asignar a aquella clase a la que la muestra posea mínima distancia de Mahalanobis a su media. La ecuación 2.134 tiene naturaleza de discriminante lineal, lo que se puede demostrar por simple desarrollo de la expresión. Por tanto, las fronteras de decisión serán tambien en este caso hiperplanos aunque en general no ortogonales a los vectores que unen las medias. Si las probabilidades a priori de las clases son iguales, el hiperplano corta a dicho vector en su punto medio. Una ilustración de este caso se muestra en la Figura 2.18.

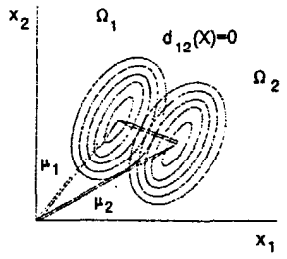


Figura 2.18: Ejemplo de Dos Clases con Idéntica Matriz de Covarianzas

III) CASO DE MATRIZ DE COVARIANZAS ARBITRARIA

En el caso más general las matrices de covarianzas son diferentes para cada clase y la expresión de la función discriminante es la 2.126, que desarrollada nos permite obtener:

$$d_i(X)=-\frac{1}{2}X^t\Sigma_i^{-1}X+\frac{1}{2}X^t\Sigma_i^{-1}\mu_i+\frac{1}{2}\mu_i^t\Sigma_i^{-1}X-\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i-\frac{1}{2}\ln|\Sigma_i|+\ln P(\Omega_i) \quad [2.136]$$

Que es la expresión de una función discriminante cuadrática, lo que se puede observar si la comparamos con la expresión 2.36 de dicho discriminante, que para cada clase es:

$$d_i(X)=X^tW_iX+\omega_i^tX+\omega_{i0} \quad [2.137]$$

Donde:

$$W_i=-\frac{1}{2}\Sigma_i^{-1}$$
$$\omega_i=\Sigma_i^{-1}\mu_i \quad [2.138]$$
$$\omega_{i0}=-\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i-\frac{1}{2}\ln|\Sigma_i|+\ln P(\Omega_i)$$

Las superficies de decisión son hipercuadráticas, como se discutión en el punto 2.2.3.

2.5.- REFERENCIAS

[ANDE-73] Anderberg M. R., **Cluster Analysis for Applications**, Academic Press, New York, 1973.

[BANK-90] Banks S., **Signal Processing and Pattern Recognition**, Prentice Hall, New York, 1990.

[BOW-92] Bow S., **Pattern Recognition and Image Preprocessing**, Marcel Dekker Inc., New York, 1992.

[BREI-84] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., **Classification and Regression Trees**, Wadsworth & Brooks/Cole Advanced Books & Sotfware, Pacific Grove, CA, 1984.

[CASA-87] Casacuberta F., Vidal E., **Reconocimiento Automático del Habla**, Marcombo, Barcelona, 1987.

[CHEN-93] Chen C. H., Pau L. F., Wang P. S. P., **Handbook of Pattern Recognition and Computer Vision**, World Scientific Pub. Co., Singapore, 1993.

[CHIE-78] Chien Y., **Interactive Pattern Recognition**, Marcel Dekker Inc., New York, 1978.

[CUAD-81] Cuadras C. M., **Métodos de Análisis Multivariante**, EUNIBAR, Barcelona, 1980.

[DUDA-73] Duda R. O., **Pattern Classification and Scene Analysis**, John Wiley & Sons, New York, 1973.

[ESCU-77] Escudero L. F., **Reconocimiento de Patrones**, Paraninfo, Madrid, 1977.

[FUKU-72] Fukunaga K., **Introduction to Statistical Pattern Recognition**, Academic Press, New York, 1972.

[HAND-81] Hand D. J., **Discrimination and Classification**, John Wiley & Sons, Chichester, 1981.

[HUBE-94] Huberty C. J., **Applied Disciminant Analysis**, John Wiley & Sons, Inc., New York, 1994.

[JAIN-88] Jain A. K., Dubes R. C., **Algorithms for Clustering Data**, Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[NILS-90] Nilsson N. J., **The Mathematical Foundations of Learning Machines**, Morgan Kaufmann Pub., San Mateo, California, 1990.

[SANC-78] Sanchez García M., **Modelos Estadísticos Aplicados a Tratamiento de Datos**, Centro de Cálculo de la Universidad Complutense, Madrid, 1978.

[SCHA-92] Schalkoff R., **Pattern Recognition. Statistical, Structural and Neural Approaches**, John Wiley & Sons, Inc., New York, 1992.

[SPAT-80] Spath H., **Cluster Analysis Algorithms for Data Reduction and Clasification of Objects**, Ellis Horwood Limited, Chichester, West Sussex, UK, 1980.

[TOU-74] Tou J. T., Gonzalez R. C., **Pattern Recognition Principles**, Addison Wesley, 1974.

[WEIS-91] Weiss S. M., Kulikowski C. A., **Computer Systems that Learn**, Morgan Kaufmann Pub. Inc., San Francisco, CA, 1991.

APENDICE: METODOS ESTADISTICOS PARAMETRICOS VERSUS METODOS GEOMETRICOS

Los métodos de clasificación estadísticos que se han expuesto permiten un aprendizaje adecuado y el diseño de reglas de clasificación con mínimo error, siempre que se conozcan las formas de las funciones de densidad de probabilidad asociadas, y que las mismas se puedan asociar a las funciones usuales (p.e. ley normal). En la práctica, estas asunciones son aplicables a un número relativamente reducido de casos, ya que, las funciones de densidad realmente observadas en los problemas prácticos se ajustan en pocas ocasiones a los modelos de funciones más usuales. Por ejemplo, es bastante frecuente encontrarse en problemas en los que las clases presentan distribuciones claramente multimodales (varios máximos), mientras que, todas las densidades paramétricas son unimodales. Además, hay que hacer notar que, los métodos paramétricos como la regla de clasificación bayesiana son conceptos estadísticos y, por tanto no es esperable un buen comportamiento, en general, en los casos en los que el conjunto de muestras de aprendizaje sea relativamente reducido.

Los problemas comentados conducen a que, siendo la aproximación estadística paramétrica más rigurosa, la misma sea sustituida por métodos geométricos en el diseño del clasificador y más en problemas de Reconocimiento de Formas en los haya un relativo control sobre la actividad del extractor de características, lo que nos permite diseñar esquemas de clasificación supervisando el buen cumplimiento el postulado tercero de Niemann en el espacio de medidas.

La aproximación estadística paramétrica, no obstante, presenta un alto interés teórico y sus resultados se utilizan frecuentemente como referencia para contrastar la bondad de otros métodos.

Tema 3

Aprendizaje Supervisado de Clasificadores

3.1.- INTRODUCCION

Este tema se dedica al estudio de procedimientos de naturaleza iterativa que permiten la determinación de reglas de decisión a partir del conjunto de muestras de aprendizaje. Como se estudió en el tema 2, una vez que se ha especificado un tipo de función de decisión para un problema de clasificación determinado, el objetivo que se plantea es la determinación de los coeficientes de dichas funciones. Los procedimientos que se estudian en este tema serán capaces de obtener solución para esos coeficientes, es decir *aprender*, siempre y cuando las clases sean separables mediante las funciones de decisión definidas.

3.2.- APRENDIZAJE DE FUNCIONES DE DECISION. PLANTEAMIENTO

Supongamos un problema de clasificación biclásica entre Ω_1 y Ω_2 en un espacio de representación bidimensional mediante una regla de decisión con frontera de decisión lineal, que en expresión generalizada es:

$$d(X)=a^t Y \qquad [3.1]$$

Si el vector de pesos **a** existe, las clases son *linealmente separables*. Supuesta dicha existencia, para las muestras controladas del conjunto de aprendizaje, que denominaremos tambien muestra de aprendizaje a secas, se debe cumplir que:

$$\begin{aligned} &Si\ X\in\Omega_1\ \ Entonces\ a^t Y>0 \\ &Si\ X\in\Omega_2\ \ Entonces\ a^t Y<0 \end{aligned} \qquad [3.2]$$

Supongamos que la muestra controlada **M** está constituida por cuatro vectores de medidas, dos de ellos pertenecientes a la clase Ω_1 y otras dos a la Ω_2. Para referirnos a ellas, cada vector llevará asociado un superíndice, que indica la clase a la que corresponde el vector, y un subíndice, que indica el número de orden dentro de la muestra controlada de su clase. Con esta nomenclatura, la muestra controlada **M** resulta:

$$M=M_{\Omega_1}\cup M_{\Omega_2} \quad Siendo: \begin{cases} M_{\Omega_1}=\{Y_1^1, Y_2^1\} \\ M_{\Omega_2}=\{Y_1^2, Y_2^2\} \end{cases} \qquad [3.3]$$

Para cada uno de los vectores de características de la muestra se deberá cumplir una u otra de las desigualdades 3.2, según su clase de pertenencia. Con el fin de

normalizar el sentido de dichas desigualdades, a los vectores de la muestra controlada de la clase Ω_2 los cambiamos de signo, con lo cual, el conjunto de desigualdades que debe cumplir el vector de pesos **a** para la muestra controlada es:

$$\begin{aligned} \omega_1 x_{11}^1 + \omega_2 x_{12}^1 + \omega_0 > 0 \\ \omega_1 x_{21}^1 + \omega_2 x_{22}^1 + \omega_0 > 0 \\ -\omega_1 x_{11}^2 - \omega_2 x_{12}^2 - \omega_0 > 0 \\ -\omega_1 x_{21}^2 - \omega_2 x_{22}^2 - \omega_0 > 0 \end{aligned} \qquad [3.4]$$

Si construimos una matriz Ψ en la que cada fila se corresponde con un vector de la muestra de aprendizaje, por lo cual resultará de dimensión m*(n+1), siendo m el número total de vectores de la muestra (en el caso particular que analizamos m=4), el conjunto de desigualdades 3.4 se puede expresar de la siguiente manera:

$$\Psi \boldsymbol{a} > \mathbf{0} \qquad [3.5]$$

Donde **0** representa el vector nulo de dimensión m.

Si existe un **a** que satisface 3.5, se dice que las desigualdades son *consistentes*. En Reconocimiento de Formas, esto corresponde la situación mencionada anteriormente de clases linealmente separables.

El problema del aprendizaje lo podemos, por tanto, definir como el de *encontrar un vector de pesos* **a** *tal que las desigualdades generadas en base a la muestra de aprendizaje sean consistentes*. Hay que observar como, en el conjunto de inecuaciones 3.5, los coeficientes son las coordenadas de los vectores de la muestra, mientras que, las incógnitas son las coordenadas del vector de pesos. Esta consideración nos es útil para analizar a continuación qué posibles valores puede adquirir el vector de pesos en un problema de clases linealmente separables concreto.

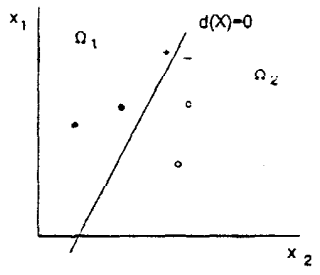


Figura 3.1: Ilustración Geométrica del Espacio de Características

Hasta este punto hemos manejado el concepto de **espacio de características**, es decir, el espacio n-dimensional cuyos elementos son los vectores de medidas **X** y por tanto, tiene por coordenadas de su sistema referencial a las componentes de dichos vectores. En este espacio, **a** se corresponde con el vector de coeficientes del hiperplano de separación (p.e. figura 3.1).

En la situación de aprendizaje del vector de pesos, la muestra controlada es un conjunto de puntos en posición fija del espacio, mientras que el vector de pesos **a** representa una recta cuya posición está a priori indefinida. Para analizar la situación resulta más adecuado interpretar las soluciones del vector **a** desde otro esquema de representación. Así se introduce el concepto de **espacio de pesos**, entendido como aquel espacio (n+1)-dimensional en el que las coordenadas del sistema de referencia vienen dadas por los coeficientes del vector de pesos, a priori parámetros no determinados, mientras que los vectores de la muestra de aprendizaje son los coeficientes que intervienen en las desigualdades 3.4. Con esta representación, cada solución al vector de pesos se corresponde con un punto del espacio de pesos.

Cada una de las desigualdades de 3.4 constituye una región acotada por un hiperplano que pasa por el origen, es decir, cada una se cumplirá para el conjunto de puntos de una región del hiperespacio, o dicho de otra manera, por el conjunto infinito de valores del vector de pesos que abarca esa región. El cumplimiento de todas las desigualdades se dará en aquella región intersección de las definidas por cada una de las desigualdades. Dicha región, denominada **región de decisión en el espacio de pesos**, será en general una parte del hiperespacio de pesos limitada por un poliedro cónico convexo, como se ilustra para el ejemplo concreto que nos ocupa, en la figura 3.2. Por ello, si el vector de pesos solución existe, la solución en general no es única sino que existen infinitos vectores de pesos solución.

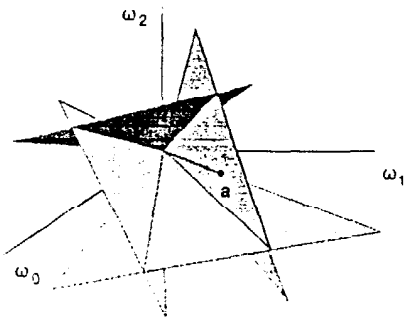


Figura 3.2: Configuración en el Espacio de Pesos Correspondiente al Ejemplo de la Figura 3.1

Resulta pues interesante imponer restricciones adicionales, como por ejemplo, las que nos pernmitan obtener soluciones que no correspondan a puntos en los límites de la región de decisión, lo que se puede conseguir buscando, en vez de un vector de pesos que converja en el límite inferior de las desigualdades ($\Psi\mathbf{a}>\mathbf{0}$), una solución a partir de un conjunto de desigualdades con un margen de seguridad respecto al límite inferior ($\Psi\mathbf{a}>\mathbf{b}>\mathbf{0}$). Tambien se puede seleccionar el vector de pesos solución con otra condición, como puede ser, la de que maximice la distancia desde las muestras de aprendizaje al hiperplano de separación de clases.

Básicamente, la aproximación a la búsqueda de una solución a la desigualdad 3.5 puede ser determinista o estadística. La aproximación determinista, es decir la que engloba los métodos que determinan el vector de pesos sin asumir nada en lo referente a las propiedades estadísticas de las clases, constituye el objetivo de este tema.

3.3.- PROCEDIMIENTOS BASADOS EN EL CONCEPTO DE DESCENSO SEGUN EL GRADIENTE

La determinación de un vector de pesos que cumpla las desigualdades 3.5 se puede realizar mediante procedimientos iterativos que, partiendo de un valor incial del vector de pesos, permita en un número finito de pasos acercarnos a un vector de pesos que sea solución de dichas desigualdades. Ahora bien, se precisa de algún mecanismo que nos permita controlar la evolución, en el sentido adecuado para alcanzar la solución, de dicho vector de pesos a lo largo de las iteraciones. Dicho mecanismo puede ser la utilización de una función criterio, tambien denominanda función objetivo, escalar del vector de pesos **a** actual (J(**a**)), que presente la particularidad de ser mínima si **a** es el vector solución.

El control de la evolución de **a** de iteración a iteración en búsqueda del mínimo de J(**a**) se puede realizar utilizando el denominado esquema de *descenso según el gradiente*.

Como recordatorio de las herramientas del análisis vectorial podemos decir que, dada una función escalar f(**Z**), que tiene como argumento a un vector **Z** de l componentes (es decir un vector tal que $\mathbf{Z}^t=\{z_1,z_2,\ldots,z_l\}$), se denomina *gradiente* $\nabla f(\mathbf{Z})$ de dicha función al vector:

$$\nabla f(\mathbf{Z}) = \begin{pmatrix} \frac{\partial f}{\partial z_1} \\ \frac{\partial f}{\partial z_2} \\ \cdot \\ \cdot \\ \frac{\partial f}{\partial z_l} \end{pmatrix} \qquad [3.6]$$

Es decir, el gradiente de una función escalar que tiene como argumento a un vector, es a su vez un vector, que presenta la interesante cualidad de que, cada componente refleja la *velocidad* de cambio de la función f en la dirección de la correspondiente componente de **Z**.

Una de las propiedades más interesantes del vector gradiente de una función escalar es que apunta en la dirección de la máxima velocidad de incremento de la función f cuando se incrementa su argumento. De la misma forma, el gradiente de f cambiado

de signo ($-\nabla f(Z)$) apunta en la dirección de la máxima velocidad de decremento de f. Utilizando esta propiedad, se pueden derivar esquemas iterativos de determinación del mínimo de una función.

La aproximación que se emplea para determinar un vector de pesos **a** que es solución del sistema de inecuaciones 3.5, es iterativa, como ya se adelantó. El proceso consiste en partir de un vector inicial de pesos **a**(1) que puede ser arbitrario. En cada iteración k+1 se obtiene el nuevo vector de pesos **a**(k+1) por corrección del correspondiente a la anterior iteración **a**(k) en base al valor del gradiente de la función objetivo J(**a**) según la expresión:

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \rho\,[\nabla J(\boldsymbol{a})]_{\boldsymbol{a}=\boldsymbol{a}(k)} \qquad [3.7]$$

Donde el coeficiente $\rho>0$ determina la magnitud de la corrección. Se puede observar como no se efectuará ninguna corrección en el vector de pesos cuando el gradiente de f sea nulo, es decir cuando f sea mínimo, caso que corresponderá, por las condiciones impuestas a la función criterio, a la situación de haber alcanzado una solución.

La ecuación 3.7 puede interpretarse geométricamente con ayuda de la figura 3.3. En ella y a efectos de claridad expositiva, suponemos que el vector **a** es unidimensional. Podemos observar como, si el gradiente es negativo en el paso k-ésimo, el vector de pesos **a**(k+1) se incrementa en la dirección positiva, es decir, acercándonos al mínimo de J(**a**). En el caso de que el gradiente sea positivo, la función criterio es creciente, luego, ocurrirá el efecto contrario de decrementarse el vector de pesos **a**(k+1). El proceso de modificación iterativa del vector de pesos concluirá, como se ha dicho anteriormente, solamente cuando se alcance el mínimo de la función criterio. Esto ocurrirá siempre que las desigualdades de la ecuación 3.5 sean *consistentes*.

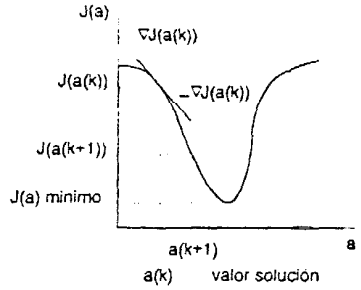


Figura 3.3: Ilustración Geométrica del Procedimiento de Descenso según el Gradiente

Evidentemente, el proceso de búsqueda de la solución del vector de pesos no solamente depende de la utilización de una función criterio adecuada, sino tambien la elección conveniente del coeficiente de corrección *c* incluido en la expresión recursiva 3.7 del procedimiento de descenso según el gradiente. En efecto, y con ayuda de la figura 3.3 se puede deducir que, si se elige un coeficiente corrector de valor muy pequeño, el proceso de convergencia puede ser muy lento, mientras que si se elige muy grande, el proceso de corrección puede oscilar, o incluso diverger.

En los puntos siguientes analizaremos dos procedimientos específicos de aprendizaje basados en este esquema: el procedimiento Perceptrón, que suministra una solución si las clases son linealmente separables pero que oscila indefinidamente en el caso de que no lo sean, y el procedimiento Ho-Kashyap, que además de encontrar dicha solución, si existe, "avisa" en el caso de que las clases no sean linealmente separables.

3.3.1.- PROCEDIMIENTO PERCEPTRON

El origen de los algoritmos de clasificación de formas puede datarse en los primeros desarrollos en el campo de la denominada *Biónica* (es decir el área dedicada a la aplicación de conceptos biológicos a las máquinas basadas en elementos y sistemas de naturaleza electrónica) relacionados con los problemas del aprendizaje en animales y máquinas.

Entre mediados de la década de los años cincuenta y principios de los sesenta, una clase de máquinas diseñadas por Rosenblatt y denominandas corrientemente *perceptrones*, parecieron ofrecer a muchos investigadores un modelo natural y potente de máquina de aprendizaje. Aunque hoy día se considera que las expectativas que se crearon en lo referente a las prestaciones del perceptron eran excesivamente optimistas, los conceptos matemáticos que surgieron de su desarrollo continuan jugando un papel de cierta relevancia en la teoría del Reconocimiento de Formas. Además, en los últimos años se ha suscitado un interés renovado por estos modelos en el marco de las Redes Neuronales.

El modelo básico del perceptrón capaz de clasificar una forma entre dos clases se muestra en la figura 3.4. La máquina está constituida por una capa S de *unidades sensoriales*, que pueden entenderse como el medio por el que la máquina recibe estímulos del exterior, que se conectan al módulo R generador de respuesta.

Cada unidad de salida R produce una respuesta que se determina a partir de una combinación lineal de los valores que alcanzan las unidades sensoriales. Así, la respuesta es:

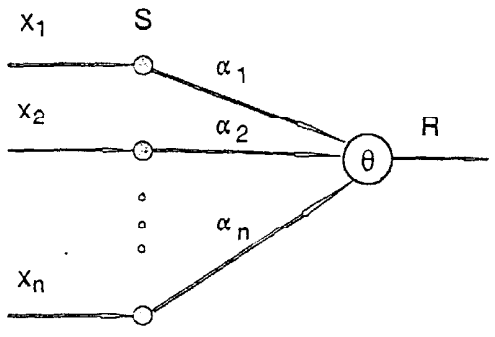


Figura 3.4: Esquema del Modelo Básico del Perceptrón

$$R = g(\eta) \qquad [3.8.a]$$

Pudiendo ser g un funcional de tipo signo, es decir:

$$g(\alpha) = sgn(\alpha) = \begin{cases} +1 & si\ \alpha > 0 \\ -1 & si\ \alpha < 0 \end{cases} \qquad [3.8.b]$$

U otrfuncional sin discontinuidades. Además, η es:

$$\eta = \sum_{i=1}^{n} \alpha_i x_i - \theta \qquad [3.8.c]$$

Como se puede observar, la unidad perceptrón implementa una regla de decisión basada en una discrminación lineal establecida por el signo del funcional η. En efecto, si establecemos las siguientes igualdades:

$$\begin{aligned} \omega_i &= \alpha_i \ \forall\ i=1,2,...n \\ \omega_0 &= -\theta \end{aligned} \qquad [3.8.d]$$

Obtenemos la expresión:

$$\eta = \omega^t X + \omega_0 = a^t Y = d(X) \qquad [3.8.e]$$

Ya analizada anteriormente. Además, es posible extender el perceptrón a un proceso de clasificación multiclásico, simplemente añadiendo tantas unidades de respuesta R como clases haya, estableciendo sus interconexiones a las unidades asociativas y realizando el proceso final de clasificación mediante una análisis de las salidas de las unidades R y una asignación a la clase cuya R asociada presente salida máxima. El modelo básico puede tambien extenderse al caso no lineal introduciendo el correspondiente preprocesador no lineal entre las unidades sensoriales y las de respuesta R, o bien, trantando el proceso como lineal por transformación lineal generalizada.

Otro aspecto muy interesante del perceptrón es su esquema de aprendizaje. El mismo es de naturaleza iterativa, de forma que el vector de pesos se va ajustando en iteraciones sucesivas, en las que se va comprobando la buena o mala clasificación de las muestras, una a una, del conjunto de aprendizaje y ajustando el valor del vector de pesos mediante un esquema de *premio-castigo*. En los puntos siguientes

nos centraremos más detalladamente en el análisis del procedimiento perceptrón considerado como uno de los derivados del concepto de descenso según el gradiente.

3.3.1.1.- FUNCION CRITERIO

La función criterio perceptrón para cada muestra del conjunto de aprendizaje puede escribirse como:

$$J(\boldsymbol{a},\boldsymbol{Y}) = \frac{1}{2}(|\boldsymbol{a}^t\boldsymbol{Y}| - \boldsymbol{a}^t\boldsymbol{Y}) \qquad [3.9]$$

Donde el primer sumando del consecuente de la expresión representa el valor absoluto del resultado de aplicar la función discriminante a la muestra Y correspondiente. Para la simplificación de las expresiones y del algoritmo perceptrón, aplicado al caso biclásico, las muestras Y de la clase Ω_2 se cambian de signo, para normalizarlas en el sentido mostrado en la expresión 3.4 y descrito en el apartado 3.2.

Así, si para un valor concreto del vector de pesos **a**, una muestra Y resulta mal clasificada, entonces $\mathbf{a}^t\mathbf{Y}<0$ con lo que $J(\mathbf{a},\mathbf{Y})=-\mathbf{a}^t\mathbf{Y}$, mientras que si resulta bien clasificada $\mathbf{a}^t\mathbf{Y}>0$ y por tanto $J(\mathbf{a},\mathbf{Y})=0$. La función criterio global $J_p(\mathbf{a})$ es:

$$J_p(\boldsymbol{a}) = \sum_{\forall \boldsymbol{Y}\in M} J(\boldsymbol{a},\boldsymbol{Y}) \qquad [3.10]$$

Presentará un mínimo igual a cero en el caso de que todas las muestras resulten bien clasificadas. El gradiente de la función criterio parcial será:

$$\nabla_{\boldsymbol{a}} J(\boldsymbol{a},\boldsymbol{Y}) = \frac{1}{2}[\boldsymbol{Y}\, sgn(\boldsymbol{a}^t\boldsymbol{Y}) - \boldsymbol{Y}]$$

Donde, por definición $\mathrm{sgn}(\mathbf{a}^t\mathbf{Y})$ es una función tal que:

$$sgn(\boldsymbol{a}^t\boldsymbol{Y}) = \begin{cases} 1 & si\ \boldsymbol{a}^t\boldsymbol{Y}>0 \\ -1 & si\ \boldsymbol{a}^t\boldsymbol{Y}\le 0 \end{cases} \qquad [3.12]$$

El hecho de introducir la condición de anulación del valor de la función discriminante con el de valor negativo en la expresión 3.12 se debe a que es deseable que se refleje en la función criterio parcial dicha condición, con la finalidad de realizar corrección en ese caso.

Si se analiza la ecuación 3.11, se observa que el gradiente es no nulo en el caso de que una muestra resulte mal clasificada y su valor en ese caso corresponde precisamente al de dicho vector muestra.

Sustituyendo 3.11 en la expresión recursiva de descenso según el gradiente 3.7 obtenemos:

$$a(k+1)=a(k)-c\nabla_a J(a,Y)=a(k)+\frac{\rho}{2}[Y(k)-Y(k)sgn[a^t(k)Y(k)]] \qquad [3.13]$$

Donde Y(k) representa a la muestra de aprendizaje considerada en el paso iterativo k-ésimo, $\rho>0$ define la magnitud de la corrección, como se dijo anteriormente y a(1) es un vector de pesos inicial arbitrario.

Sustituyendo 3.12 en 3.13 resulta la siguiente expresión de corrección:

$$a(k+1)=a(k)+\rho\begin{cases}0 & si\ a^t(k)Y(k)>0\\ Y(k) & si\ a^t(k)Y(k)\le 0\end{cases} \qquad [3.14]$$

Donde **0** representa al vector nulo de dimensión n+1.

3.3.1.2.- APRENDIZAJE POR PREMIO-CASTIGO

El algoritmo de aprendizaje del vector de pesos **a** según el procedimiento perceptrón es, como hemos dicho, de naturaleza iterativa y puede ser resumido como se muestra a continuación:

> *Dados dos conjuntos de muestras de aprendizaje pertenecientes a las clases* Ω_1 *y* Ω_2, *escogido un vector de pesos inicial* a(1) *que puede ser arbitrariamente escogido, definido un valor del factor de corrección* ρ *positivo, normalizadas las muestras de la clase* Ω_2 *en signo en la expresión generalizada, según se muestra en 3.4, y dada la expresión de corrección 3.14, el paso k-ésimo del algoritmo de aprendizaje es el siguiente:*
>
> $$Si\ [a^t(k)Y(k)\le 0]\quad Entonces\ a(k+1)=a(k)+\rho\,Y(k)$$
> $$En\ otro\ caso\ a(k+1)=a(k)$$

Es decir, el algoritmo modifica **a** si y solo si la muestra considerada en el paso k-ésimo ha resultado mal clasificada por el vector de pesos en este paso.

El algoritmo perceptrón, como puede observarse, es un procedimiento de aprendizaje por premio-castigo, donde, el premio se asigna en el caso de buena clasificación, y se corresponde con la ausencia de castigo, es decir, ausencia de corrección en a(k). En caso contrario, si la muestra resulta mal clasificada, la máquina resulta *castigada* modificando el valor de a(k). El procedimiento de corrección continúa hasta que, trás una pasada de todas las muestras por el algoritmo de apprendizaje, todas resultan correctamente clasificadas, es decir, no se efectúa ninguna corrección en a(k). En

este punto, el algoritmo ha alcanzado la convergencia en un resultado para el vector de pesos, pues esta condición corresponde al caso de que la función criterio global $J_p(\mathbf{Y})$ de la expresión 3.10 alcanza valor mínimo.

3.3.1.3.- CONVERGENCIA

Para un valor fijo de ρ, el **Teorema de Convergencia del Perceptrón** establece que si:

a) Las clases en consideración son linealmente separables

b) Cada muestra del conjunto de aprendizaje se "presenta" al procedimiento de aprendizaje tantas veces como sea necesario

Entonces el algoritmo perceptrón *converge en una solución en un número finito de pasos.*

La convergencia del algoritmo perceptrón puede demostrarse de diversas maneras, siendo las mostradas a continuación de las más concisas.

A efectos de simplificar la exposición de la demostración, sea $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m\}$ el conjunto de m muestras del conjunto de aprendizaje pertenecientes a las dos clases del problema, convenientemente normalizadas en signo como se indicó en el apartado 3.2. Sea que el vector de pesos solución al problema lo denominamos $\mathbf{a}^*$. Este vector presenta la propiedad:

$$(\mathbf{a}^*)^t \mathbf{Y}_i > 0 \quad \forall i=1,2,\ldots,m \qquad \textbf{[3.15]}$$

Expresión que se puede generalizar introduciendo un umbral T no negativo, de manera que si las clases son linealmente separables:

$$(\mathbf{a}^*)^t \mathbf{Y}_i > T \quad \forall i=1,2,\ldots,m \qquad \textbf{[3.16]}$$

De la discusión geométrica del apartado 3.2 se puede deducir que, la introducción de un umbral T en la expresión 2.16 equivale a establecer una "franja" a cada lado del hiperplano $\mathbf{a}^t(k)\mathbf{Y}(k)=0$, lo que da lugar a que cualquier muestra en esta región resulte incorrectamente clasificada. Además, un incremento de T provoca una disminución de la región de soluciones (figura 3.2) para **a** en el espacio de pesos.

Sea que por simplicidad asumimos que el factor de corrección $\rho=1$, lo que no implica pérdida de generalidad, ya que según la forma de la expresión 3.14 cualquier otro

valor de ρ puede asignarse a los vectores muestra como una constante de normalización. De 3.14 y 3.16 resulta:

$$a(k+1)=a(k)+\begin{cases} 0 & si\ a^t(k)Y(k)>T \\ Y(k) & si\ a^t(k)Y(k)\le T \end{cases} \qquad [3.17]$$

Con la intención tambien de simplificar la notación, sea que los indices k sólo se asocian a aquellos pasos en los que se produce corrección durante el proceso de aprendizaje, no contando los índices k correspondientes a muestras correctamente clasificadas. Por ello, readaptando la notación de índices la expresión anterior la escribimos:

$$a(k+1)=a(k)+Y(k) \qquad [3.18]$$

Que, como ocurrirá al haber corrección, se cumplirá para todo k:

$$a^t(k)Y_i(k)\le T \qquad [3.19]$$

La convergencia del algoritmo significa que, a partir de un cierto valor finito k_p del índice se cumplirá:

$$a(k_p)=a(k_p+1)=a(k_p+2)=... \qquad [3.20]$$

Con las simplificaciones y detalles comentados, una demostración de la convergencia es la siguiente.

DEMOSTRACION 1:

De la ecuación 3.18 se puede deducir que:

$$a(k+1)=a(1)+Y_i(1)+Y_i(2)+...+Y_i(k) \qquad [3.21]$$

Efectuando el producto escalar de $\mathbf{a}^*$ con ambos lados de la expresión anterior se obtiene:

$$a^t(k+1)a^* = a^t(1)a^*+Y_i^t(1)a^*+Y_i^t(2)a^*+...+Y_i^t(k)a^* \qquad [3.22]$$

Como a partir de la ecuación 3.16 se obtiene que cada producto escalar del vector muestra i-ésimo por el vector de pesos solución es mayor que el umbral T, entonces:

$$a^t(k+1)a^* \ge a^t(1)a^*+kT \qquad [3.23]$$

Utilizando la desigualdad de Cauchi-Schwartz[1] tenemos que:

$$[a^t(k+1)\ a^*]^2 \leq \|a(k+1)\|^2 \|a^*\|^2 \qquad [2.24]$$

De donde, despejando obtenemos:

$$\|a(k+1)\|^2 \geq \frac{[a^t(k+1)\,a^*]^2}{\|a^*\|^2} \qquad [3.25]$$

Sustituyendo 3.25 en 3.23 obtenemos la desigualdad:

$$\|a(k+1)\|^2 \geq \frac{[a^t(1)\,a^*+kT]^2}{\|a^*\|^2} \qquad [3.26]$$

Con un razonamiento alternativo podemos obtener una contradicción relacionada con el antecedente de la desigualdad. Así, de la expresión 3.18 podemos obtener:

$$\|a(j+1)\|^2 = \|a(j)\|^2 + 2a^t(j)\,Y_i(j) + \|Y_i(j)\|^2 \qquad [3.27]$$

Que se puede poner:

$$\|a(j+1)\|^2 - \|a(j)\|^2 = 2a^t(j)\,Y_i(j) + \|Y_i(j)\|^2 \qquad [3.28]$$

Si a continuación definimos:

$$Q = \max_{\forall i=1,2,\ldots,N} \|Y_i(j)\|^2 \qquad [3.29]$$

Y utilizamos la desigualdad 3.19 resulta el conjunto de desigualdades:

$$\|a(j+1)\|^2 - \|a(j)\|^2 \leq 2T+Q \qquad [3.30]$$

Que sumadas para todo j=1,2,...,k generan la nueva desigualdad:

---

[1] La desigualdad de Cauchy-Schwart establece que, para dos vectores μ y ξ se cumple que:

$$\|\mu\|^2 \|\xi\|^2 \geq (\mu^t\xi)^2$$

$$\| a(k+1) \|^2 \leq \| a(1) \|^2 + (2T+Q)k_p \qquad [3.31]$$

Comparando 3.31 con 3.26, se observa como las desigualdades entran en conflicto para un valor de k suficientemente grande. De ello se deduce que k no puede ser mayor que el valor k_p que es solución a:

$$\frac{[a^t(1)a^* + k_l T]^2}{\| a^* \|^2} = \| a(1) \|^2 + (2T+Q)k_l \qquad [3.32]$$

Ecuación que nos indica que k_p es finito, lo que implica que el algoritmo perceptrón, para el caso de clases linealmente separables, converge en un número finito de pasos.

c.q.d.

DEMOSTRACION 2:

Es posible tambien efectuar una demostración ligeramente diferente partiendo de la consideración de que el umbral T=0. Con esta condición, la expresión 3.23 resulta:

$$a^t(k+1)a^* \geq a^t(1)a^* + kh \qquad [3.33]$$

Donde:

$$h = \min_{\forall i=1,2,\ldots,N} [Y_i(j)\, a^*] \qquad [3.35]$$

Ya que $\mathbf{a}^*$ es un vector de pesos solución, por la ecuación anterior h será mayor que cero. Tambien, una vez que $\mathbf{a}^t(j)\mathbf{Y}_i(j) \leq 0$, la expresión 3.28 resulta:

$$\| a(j+1) \|^2 - \| a(j) \|^2 \leq \| Y_i(j) \|^2 = Q \qquad [3.35]$$

El resto de la demostración es equivalente. El límite en el número de pasos requeridos para convergencia con T=0 es el resultado de la solución de la ecuación:

$$\frac{[a^t(1)a^* + k_p h]^2}{\| a^* \|^2} = \| a(1) \|^2 + Qk_p \qquad [3.36]$$

Como comentario final a este punto hay que decir que, aunque las expresiones 3.32 y 3.37 establecen un límite en k_p, estas ecuaciones no pueden utilizarse para determinar el número de pasos requeridos para alcanzar la convergencia, ya que sería preciso conocer el vector solución $\mathbf{a}^*$. Además, y por otro lado, k_p depende tambien del vector de pesos inicial $\mathbf{a}(1)$.

3.3.1.4.- VARIACIONES DEL PROCEDIMIENTO

Del procedimiento perceptrón mostrado en los puntos anteriores se pueden formular diversas variaciones, dependiendo de como se seleccione el valor del factor de corrección ρ. Entre los tipos de corrección más usuales se encuentran: la *corrección fija*, la *corrección absoluta* y la *corrección fraccionaria*.

En el caso de la *corrección fija* ρ es una constante mayor que cero sin ninguna otra consideración adicional, como es el caso en lo analizado hasta ahora.

En el caso de la *corrección absoluta*, ρ se escoge de manera que sea suficientemente grande para asegurar que el vector muestra considerado resulte correctamente clasificado tras un solo ajuste del vector de pesos. En otras palabras, si **Y**(k) resulta mal clasificado por **a**(k), es decir si:

$$a^t(k)\,Y(k) \le 0 \qquad [3.37]$$

Se escoge ρ de manera que con el vector de pesos corregido **a**(k+1) se obtenga:

$$a^t(k+1)\,Y(k) > 0 \qquad [3.38]$$

Sustituyendo **a**(k+1) por su valor dado por la expresión 3.14, la ecuación anterior queda:

$$[a(k)+\rho\,Y(k)]^t Y(k) > 0 \qquad [3.39]$$

Teniendo en cuenta 3.37 y 3.39, un valor de ρ que cumple estas condiciones es:

$$\rho = ceil\left(\frac{|\,a^t(k)\,Y(k)\,|}{Y^t(k)\,Y(k)}\right) \qquad [3.40]$$

Donde *ceil*() es la función que asigna a ρ el valor entero más pequeño que es mayor que la cantidad resultado del cociente.

En el caso de *corrección fraccionaria* se escoge ρ de manera que el valor absoluto de la diferencia entre el valor actual del discriminante: $\mathbf{a}^t(k)\mathbf{Y}(k)$ y el valor para el vector de pesos corregido: $\mathbf{a}^t(k+1)\mathbf{Y}(k)$, sea una cierta fracción positiva λ del valor absoluto del discriminante actual, es decir:

$$|\mathbf{a}^t(k)\mathbf{Y}(k) - \mathbf{a}^t(k+1)\mathbf{Y}(k)| = \lambda |\mathbf{a}^t(k)\mathbf{Y}(k)| \quad [3.41]$$

Sustituyendo **a**(k+1) por su valor, dado por la expresión de corrección 3.14 y despejando ρ se obtiene:

$$\rho = \lambda \frac{|\mathbf{a}^t(k)\mathbf{Y}(k)|}{\mathbf{Y}^t(k)\mathbf{Y}(k)} \quad [3.42]$$

En este caso, el buen funcionamiento del algoritmo de corrección precisa que el vector inicial de pesos **a**(k) sea distinto del vector nulo. Respecto a este caso se puede concluir que si $\lambda>1$, cada muestra **Y**(k) resulta correctamente clasificada. Además, se ha demostrado que el algoritmo perceptrón converge para $0<\lambda<2$.

3.3.1.5.- CLASIFICADOR MULTICLASICO

Los problemas de clasificación multiclásico los analizamos en el apartado 2.2.2 del tema anterior, donde se consideraron tres opciones para la regla de clasificación. A continuación discutiremos la determinación de los vectores de pesos de las funciones de decisión para los tres casos utilizando el procedimiento perceptrón.

En el caso **a**, considerábamos que cada clase era separable del resto mediante una superficie de decisión. Existirán por tanto c funciones de decisión, cada una de las cuales se puede determinar utilizando el esquema del procedimiento perceptrón biclásico anteriormente descrito. Así, para determinar la frontera de decisión que separa la clase Ω_i del resto, basta considerar dos conjuntos de muestras de aprendizaje: el primero, constituido por las muestras de la clase i y el el segundo, constituido por las muestras de las restantes clases.

En el caso **b**, cada clase resultaba separable de cada una de las otras. En este caso, consiste en determinar c(c-1)/2 funciones de decisión, cada una de las cuales separa una Ω_i de una Ω_j. Tambien estas funciones se pueden determinar utilizando un procedimiento perceptrón biclásico, donde en cada función de decisión $d_{ij}(\mathbf{X})$, las muestras de aprendizaje de una clase son las de la clase i y las de la otra, las de la clase j.

En el caso **c** se asume la existencia de c funciones de decisión con la propiedad de que:

$$Si\ X\in\Omega_i\ Entonces:\ d_i(X) > d_j(X)\ \forall j=1,2,...,c;\ j\neq i \qquad [3.43]$$

A continuación mostraremos el procedimiento perceptrón para aprender los vectores de peso $\mathbf{a}_i$ de dichas funciones de decisión. Dicho algoritmo iterativo, ajustará en cada iteración las funciones de decisión que sea preciso, por lo que ahora no nos encontramos en situación de normalizar las muestras en signo como se ha hecho hasta ahora. En este sentido, es necesario reformular la expresión 3.14 de descenso según el gradiente sin incluir el cambio de signo en los vectores muestra **Y** correspondientes. La expresión recursiva de ajuste resulta:

$$a_i(k+1) = a(k) + \rho \begin{cases} 0 & si\ a^t(k)Y(k)>0\ \ e\ \ Y(k)\in\Omega_i \\ Y(k) & si\ a^t(k)Y(k)\leq 0\ \ e\ \ Y(k)\in\Omega_i \\ -Y(k) & si\ a^t(k)Y(k)\geq 0\ \ e\ \ Y(k)\notin\Omega_i \\ 0 & si\ a^t(k)Y(k)<0\ \ e\ \ Y(k)\notin\Omega_i \end{cases} \qquad [3.44]$$

En base a esta expresión el procedimiento resulta de la siguiente manera:

Definido ρ *y el conjunto de vectores de pesos iniciales* $\{a_i(1);i=1,2,...,c\}$, *en el paso* k *en el que* $Y(k)\varepsilon\Omega_i$ *se efectúan las siguientes operaciones:*

Si $(d_i[X(k)] > d_j[X(k)]\ \forall j=1,2,...,c;\ j\neq i)$ *Entonces* $a_j(k+1) = a_j(k)\ \forall j=1,2,...,c$

En otro caso si (*Existe algun* (*algunos*) *m tal* (*tales*) *que* $d_i[X(k)] \leq d_m[X(k)]$) *en*

$a_i(k+1) = a_i(k) + \rho\, Y(k)$;

$a_m(k+1) = a_m(k) - \rho\, Y(k)$;

$a_j(k+1) = a_j(k)\ \forall j=1,2,...,c;\ j\neq i;\ j\neq m$;

}

Si las clases son separables según el modelo **c** este procedimiento converge en los vectores solución en un número finito de pasos.

3.3.2.- PROCEDIMIENTOS DE ERROR CUADRATICO MINIMO

El procedimiento perceptrón y sus variantes convergen en una solución si las clases en consideración son separables por la superficie de decisión especificada. Sin embargo si las clases no son separables, el procedimiento oscilará indefinidamente. Como no es posible determinar a priori el número de pasos requeridos para la convergencia en el primer caso, tampoco es posible asegurar con certeza absoluta si un número elevado de pasos de aprendizaje implica que las clases no son separables. El procedimiento que se describe a continuación, además de ser convergente para clases separables, indica si las clases consideradas no son separables por la superficie especificada.

Por otro lado, en el procedimiento perceptrón se procede a probar la clasificación de las muestras una a una y corregir en base a las muestras mal clasificadas. En este apartado, por el contrario, vamos a considerar un procedimiento de aprendizaje de clasificadores biclásicos que involucra todas las muestras en cada paso de ajuste.

En este procedimiento, en vez de plantearse la búsqueda de un **a** que, para cada una de las m muestras de aprendizaje, asegure el cumplimiento de la desigualdad $\mathbf{a}^t\mathbf{Y}_i>0$ $\forall k=1,2,...,m$ (donde las muestras correspondientes a la clase Ω_2 se habrán multiplicado por -1 para normalizar el signo de las desigualdades), se pretende resolver $\mathbf{a}^t\mathbf{Y}_i=b_i$ $\forall k=1,2,...,m$, donde los b_i son un conjunto de constantes positivas arbitrariamente especificadas. Ambas formulaciones son absolutamente equivalentes en lo referente a la búsqueda de un vector de pesos solución, con la ventaja de sustituir el problema de búsqueda de solución a un sistema de inecuaciones, por el de solucionar un sistema de ecuaciones lineales.

El tratamiento de ecuaciones lineales simultáneas se simplifica utilizando notación matricial. Así, siendo Ψ la matriz de dimensión m*(n+1), introducida en el apartado 3.2, donde cada fila corresponde a una de las muestras del conjunto de aprendizaje $\mathbf{Y}_i$, y denominando **b** al vector m-dimensional constituido por los márgenes b_i asociados a cada muestra $\mathbf{Y}_i$, es decir:

$$\boldsymbol{b} = (b_1 \;\; b_2 \;\; . \;\; . \;\; . \;\; b_i \;\; . \;\; . \;\; . \;\; b_m)^t \qquad [3.45]$$

El problema de determinación de **a** es el de resolver el sistema:

$$\Psi \boldsymbol{a} = \boldsymbol{b} \qquad [3.46]$$

Si Ψ no fuese singular la solución sería $\mathbf{a}=\Psi^{-1}\mathbf{b}$, pero al ser una matriz rectangular de dimensiones m*(n+1) que normalmente posee más filas que columnas, existen más ecuaciones que incógnitas, **a** está sobredeterminado y no suele existir una solución

exacta. Sin embargo, se puede buscar un vector de pesos que minimice alguna función de error entre $\Psi\mathbf{a}$ y $\mathbf{b}$. Así, si definimos el vector error m-dimensional $\mathbf{e}$:

$$\boldsymbol{e} = \Psi\boldsymbol{a}-\boldsymbol{b} \qquad [3.47]$$

Se puede obtener una solución tratando de minimizar una función criterio escalar basada en este vector. Dicha función puede ser la magnitud de dicho vector error:

$$J_e(\boldsymbol{a}) = \|\Psi\boldsymbol{a}-\boldsymbol{b}\|^2 = \sum_{i=1}^{m} (\boldsymbol{a}^t\boldsymbol{Y}_i-b_i)^2 \qquad [3.48]$$

Que como se observa, equivale a minimizar la suma de errores cuadráticos entre los valores de los discriminantes para las muestras y sus márgenes correspondientes. Por este motivo, a los procedimientos que se basan en la minimización de esta función critero se les denomina de *error cuadrático mínimo*, lo cual es un problema clásico y se puede resolver por procedimiento de búsqueda según el gradiente.

3.3.2.1.- PROCEDIMIENTO HO-KASHYAP

Una opción para obtener un vector de pesos solución a partir de la expresión matricial 3.46 consiste en no predefinir $\mathbf{b}$, que por tanto estará indeterminado a priori, pero restringiendo a que sus componentes b_i posean valor positivo en todo momento:

$$b_i > 0 \ \forall i=1,2,...m \qquad [3.49]$$

En esta situación, con dos vectores de parámetros $\mathbf{a}$ y $\mathbf{b}$ a determinar, la función criterio basada en el error cuadrático dependerá de ambos, por lo que debemos expresarla como:

$$J_e(\boldsymbol{a},\boldsymbol{b}) = \|\Psi\boldsymbol{a}-\boldsymbol{b}\|^2 \qquad [3.50]$$

Los gradientes respecto a ambos vectores serán, respectivamente:

$$\nabla_a J_e(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{m} 2(\boldsymbol{a}^t\boldsymbol{Y}_i - \boldsymbol{b}_i)\boldsymbol{Y}_i = 2\Psi^t(\Psi\boldsymbol{a}-\boldsymbol{b})$$

$$\nabla_b J_e(\boldsymbol{a},\boldsymbol{b}) = -2(\Psi\boldsymbol{a}-\boldsymbol{b}) = -2\boldsymbol{e} \qquad [3.51]$$

Como **a** no posee restricciones, para minimizar respecto a dicho vector podemos igualar el primero de los gradientes anteriores a cero, con lo que resulta:

$$\Psi^t\Psi\boldsymbol{a} = \Psi^t\boldsymbol{b} \qquad [3.52]$$

Con ello hemos transformado el problema de resolver 3.46 a resolver 3.52, cuya ventaja estriba en que la matriz $\Psi^t\Psi$ es cuadrada de dimensiones (n+1)*(n+1) y generalmente no singular, por lo que posee inversa, pudiéndose obtener una solución para **a** simplemente como:

$$\boldsymbol{a} = (\Psi^t\Psi)^{-1}\Psi^t\boldsymbol{b} = \Psi^{\#}\boldsymbol{b} \qquad [3.53]$$

Esta ecuación liga a ambos vectores, siendo $\Psi^{\#}$ la denominada *pseudoinversa* de Ψ [1]. Existiendo esta ecuación de relación y dada la restricción de valores 3.49 para el vector de márgenes, el procedimiento de ajuste iterativo lo establecemos para **b**. La expresión será:

$$\boldsymbol{b}(k+1) = \boldsymbol{b}(k) - \rho\delta[\boldsymbol{b}(k)] \qquad [3.54]$$

Donde $\delta[\mathbf{b}(k)]$ representa la variación, dependiente del gradiente, a realizar sobre **b**(k) para obtener **b**(k+1). Como **b** está limitado a valores positivos de sus componentes, el ajuste iterativo debe evitar su convergencia a cero o su actualización a valores negativos. Ello obliga a formular una expresión modificada de la variación en el descenso según el gradiente:

$$\delta[\boldsymbol{b}(k)] = \frac{1}{2}[\nabla_b J_e - |\nabla_b J_e|] \qquad [3.55]$$

Donde $|\nabla_b J_e|$ es un vector cuyas componentes son el valor absoluto de las de $\nabla_b J_e$. La expresión puede interpretarse así: si la componente i (i=1,2,...,m) del vector error es positiva, se incrementa la componente i de **b**(k) en una cantidad igual a aquella

[1] La matriz pseudoinversa $\Psi^{\#}$ de una dada Ψ presenta algunas interesantes propiedades como:
a) Si Ψ es cuadrada y no singular: $\Psi^{\#}=\Psi^{-1}$
b) Se cumple que $\Psi^{\#}\Psi = I$, siendo I la matriz identidad, pero en general $\Psi\Psi^{\#}\neq I$
c) Si $\Psi^t\Psi$ es singular, la solución de $\Psi^t\Psi\mathbf{a}=\Psi^t\mathbf{b}$ no es única

para alcanzar la igualdad. Si por el contrario es negativa, y para respetar la restricción, la componente i de **b**(k) correspondiente no se modifica. Si definimos los vectores siguientes:

$$\begin{aligned} \boldsymbol{e}(k) &= \Psi \boldsymbol{a}(k) - \boldsymbol{b}(k) \\ \boldsymbol{e}^{+}(k) &= \frac{1}{2}[\boldsymbol{e}(k) + |\boldsymbol{e}(k)|] \end{aligned} \qquad \textbf{[3.56]}$$

Obtenemos que las componentes del vector $\mathbf{e}^{+}$(k) son tales que:

$$\forall i=1,2,\ldots,m: \begin{cases} \text{si } e_i(k) \geq 0 \text{ entonces } e_i^{+}(k)=0 \\ \text{si } e_i(k)<0 \text{ entonces } e_i^{+}(k)=e_i(k) \end{cases} \qquad \textbf{[3.59]}$$

De las expresiones anteriores podemos hacer la siguiente lectura: si todas las componentes de **e**(k) no son positivas, pero no todas nulas, las clases no son separables por la función especificada. Si, por otro lado, son todas nulas, se ha alcanzado la solución. En otro caso, la búsqueda de solución debe continuar. Así, podemos escribir el siguiente procedimiento:

> *Dados un factor de corrección* ρ, *un vector de márgenes inicial* **b**(1) *arbitrario, pero con la restricción de que sus componentes sean todas positivas, en el paso* k *las operaciones a efectuar son:*

$\boldsymbol{a}(k) = \Psi^{\#}\boldsymbol{b}(k)$
$\boldsymbol{e}(k) = \Psi\boldsymbol{a}(k) - \boldsymbol{b}(k)$
Si $\forall i=1,2,\ldots,m \;\; e_i(k)\neq 0$ *y no todas nulas entonces las clases no son separable*
En otro caso: Si $\forall i=1,2,\ldots,m \; e_i(k)=0$ *entonces alcanzada la solución*

En otro caso:
$$\begin{cases} \boldsymbol{e}^{+}(k) = \frac{1}{2}(\boldsymbol{e}(k)-|\boldsymbol{e}(k)|) \\ \boldsymbol{b}(k+1) = \boldsymbol{b}(k) + 2\rho\boldsymbol{e}^{+}(k) \\ \textit{Siguiente iteración} \end{cases}$$

Es decir, se observa como el procedimiento incluye un *test de separabilidad*, que nos indica si las clases no son separables mediante la función de decisión definida.

3.3.2.2.- COMENTARIOS AL PROCEDIMIENTO HO-KASHIAP

Cuando existe solución para las desigualdades $\Psi\mathbf{a}>\mathbf{0}$, el procedimiento Ho-Kashyap *converge en una, para valores del factor de corrección* $0<\rho\leq 1$. La demostración de esta afirmación se sale de los objetivos del curso y por ello no la incluimos. No obstante, el lector interesado puede encontrar demostraciones en **[DUDA-73]** y **[TOU-74]**.

Este procedimiento presenta una velocidad alta de convergencia debido a, por un lado, la modificación simultánea del vector de pesos y el vector de márgenes y, por otro, a la utilización de un esquema en el cual la adaptación de valores se efectúa para todas las muestras de aprendizaje a la vez. La desventaja del método se encuentra en la necesidad de efectuar la inversión de la matriz $\Psi^t\Psi$. Sin embargo, esto no presenta serias dificultades a menos que las muestras presenten una dimensionalidad muy elevada. Además, este proceso de inversión hay que efectuarlo sólo una vez para cada posible discriminante.

Por último, hacemos un comentario relativo a la existencia de inversa de $\Psi^t\Psi$. Dado que la matriz Ψ está constituida por las muestras expresadas en forma homogénea (es decir, vectores de dimensión n+1), el hecho de que $\Psi^t\Psi$ posea inversa depende, no solo de las muestras del conjunto de aprendizaje, sino también de la función de decisión escogida. Hay que tener en cuenta que si al menos n+1 de las muestras utilizadas para generar Ψ están *bien distribuidas* (lo que significa que al menos ese subconjunto de n+1 muestras no forman parte de un mismo hiperplano del espacio), se garantiza que $\Psi^t\Psi$ posea inversa.

3.4.- METODO DE LAS FUNCIONES POTENCIALES

Los procedimientos de aprendizaje anteriormente analizados permiten determinar funciones de decisión para clasificadores si las clases son linealmente separables en el espacio de representación utilizado. En este apartado nos centraremos en el estudio de la metodología de diseño basada en el concepto de función potencial, conceptualmente diferente a las basadas en la técnica de descenso según el gradiente, que permite el diseño de funciones de decisión generales.

El método se basa en el establecimiento de una analogía entre los problemas de clasificación y ciertos problemas de la física. Para exponer los fundamentos del mismo, supongamos esa analogía entre la función discriminante y el potencial en un problema electrostático.

Sea un problema de clasificación de formas en un entorno biclásico Ω_1, Ω_2. Cada muestra de aprendizaje $\mathbf{X}_i$ se hace corresponder con un punto en el espacio n-dimensional de representación. Si a cada una de las citadas muestras le suponemos asociada una cierta carga q_i que será positiva si la muestra pertenece a Ω_1 y negativa si pertenece a Ω_2, por la analogía suponemos que en el espacio existirá una distribución de potencial que será máximo en los puntos en que está situada cada muestra y decrecerá según no alejemos de ella. A cada carga $\mathbf{X}_i$, por tanto, le podemos asociar una función potencial $U(\mathbf{X},\mathbf{X}_i)$, que reflejará el potencial generado en cada punto $\mathbf{X}$ del espacio por una carga unitaria situada en la posición $\mathbf{X}_i$.

Por tanto, en cada punto del espacio de representación existirá un potencial generado por las contribuciones de las funciones potenciales situadas en las posiciones de las muestras del conjunto de aprendizaje. Dicha distribución de potencial la podemos asimilar a una cierta función discriminante, de manera que:

$$d(X) = \sum_{i=1}^{m} q_i U(X,X_i) \qquad [3.60]$$

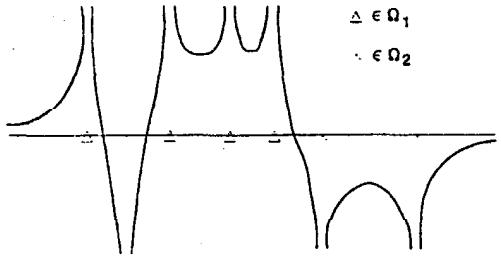


Figura 3.5: Ejemplo de Función Potencial como Función Discriminante en un problema Unidimensional Biclásico

El lugar de los puntos tales que $d(\mathbf{X})=0$ constituye la frontera de decisión entre las dos clases, como se puede observar en el ejemplo de la figura 3.5.

3.4.1.- PROCEDIMIENTO DE APRENDIZAJE BICLASICO

Para obtener un discriminante en un problema n-dimensional biclásico se parte de un conjunto de aprendizaje con m muestras y de una cierta función potencial $U(\mathbf{X},\mathbf{X}_i)$ y se asocia a las muestras de la clase Ω_1 "carga" de valor +1 y a las de Ω_2 "carga" de valor -1. El procedimiento de aprendizaje es iterativo de prueba-corrección de la función discriminante que se ajustará, muestra a muestra, hasta que el conjunto de aprendizaje al completo resulte bien clasificado. Así si k es el índice de la iteración actual, $d_{k-1}(\mathbf{X})$ el discriminante obtenido en la iteración anterior, $d_k(\mathbf{X})$ el resultado del ajuste en la iteración actual y q_k el valor de la "carga" para la muestra actual $\mathbf{X}[k]$, la expresión de ajuste resulta:

$$d_k(X) = d_{k-1}(X) + q_k U(X,X[k]) \qquad [3.61]$$

Donde los valores de q_k se definen según la siguiente regla:

$$q_k = \begin{cases} 0 \ si \ X[k] \in \Omega_1 \ y \ d_{k-1}(X[k]) > 0 \\ 0 \ si \ X[k] \in \Omega_2 \ y \ d_{k-1}(X[k]) < 0 \\ 1 \ si \ X[k] \in \Omega_1 \ y \ d_{k-1}(X[k]) \leq 0 \\ -1 \ si \ X[k] \in \Omega_2 \ y \ d_{k-1}(X[k]) \geq 0 \end{cases} \qquad [3.62]$$

Basándonos en 3.61, el procedimiento puede escribirse:

Partiendo de una función discriminante inicial $d_0(\mathbf{X})=0$. *En la iteración* k *hacer:*:

Si ($\mathbf{X}(k)\varepsilon\Omega_1$) *Entonces*: {

Si ($d_{k-1}(\mathbf{X}[k])>0$) *Entonces*: $d_k(\mathbf{X})=d_{k-1}(\mathbf{X})$;

En caso contrario: $d_k(\mathbf{X})=d_{k-1}(\mathbf{X})+U(\mathbf{X},\mathbf{X}[k])$;

};

Si ($\mathbf{X}(k)\varepsilon\Omega_2$) *Entonces*: {

Si ($d_{k-1}(\mathbf{X}[k])<0$) *Entonces*: $d_k(\mathbf{X})=d_{k-1}(\mathbf{X})$;

En caso contrario: $d_k(\mathbf{X})=d_{k-1}(\mathbf{X})-U(\mathbf{X},\mathbf{X}[k])$;

};

Las iteraciones continuarán hasta que las m muestras de aprendizaje resulten bien clasificadas.

3.4.2.- GENERACION DE FUNCIONES POTENCIALES

Una cuestión importante al utilizar este método de aprendizaje es la definición de qué tipos de funciones utilizar como potenciales. La definición del método que hicimos al principio de este apartado nos determinaba unas pautas que deben cumplir estas funciones:

I) Deben presentar valores máximos en los puntos $\mathbf{X}=\mathbf{X}_i$ donde se sitúan las muestras

II) Deben ser funciones contínuas y decrecer monótonamente según se incrementa la distancia $\mathbf{X}$-$\mathbf{X}_i$.

III) Si $U(\mathbf{Z},\mathbf{X}_i)=U(\mathbf{Z},\mathbf{X}_j)$, $\mathbf{X}_i$ y $\mathbf{X}_j$ deben tener el mismo grado de similaridad a $\mathbf{Z}$.

Clásicamente se utilizan dos tipos de funciones potenciales $U(\mathbf{X},\mathbf{X}_i)$: las obtenidas en base a series truncadas de expansiones ortonormales, que denominaremos *Tipo 1*, y las expresadas en base a funciones simétricas de las dos variables $\mathbf{X}$ y $\mathbf{X}_i$, que denominaremos de *Tipo 2*. A continuación haremos una breve reseña de ambas.

3.4.2.1.- FUNCIONES POTENCIALES DE TIPO 1

Las funciones potenciales de este tipo son series truncadas de p términos, de la forma:

$$U(X,X_i) = \sum_{k=1}^{p} \varphi_k(X)\varphi_k(X_i) \qquad [3.63]$$

Donde $\varphi_k(\mathbf{X})$ son funciones ortonormales, de los vectores n-dimensionales $\mathbf{X}$, sobre la región de definición de las formas. A continuación hacemos una ligera repaso de conceptos relacionados con dichas funciones ortonormales.

a) FUNCIONES ORTONORMALES UNIVARIANTES

Se define *producto interno* de dos funciones f(x) y g(x) en el intervalo [a,b]=a≤x≤b a:

$$(f,g) = \int_a^b f(x)g(x)dx \qquad [3.64]$$

Se denomina *norma* de una función ***f(x)*** al producto interno de la función por si misma:

$$(f,f) = \int_a^b f(x)f(x)dx \qquad [3.65]$$

Una función cuya norma es unitaria se dice que está *normalizada*. Por tanto, resulta sencillo normalizar cualquier función, simplemente dividiéndola por su norma.

Dos funciones ***f(x)*** y ***g(x)*** se denominan *ortogonales* respecto a una función de peso ***u(x)*** en el intervalo [a,b] si:

$$\int_a^b u(x)f(x)g(x)dx = 0 \qquad \textbf{[3.66]}$$

Un conjunto de funciones $\phi_1(x)$, $\phi_2(x)$,... que son ortogonales dos a dos en [a,b] se denomina *conjunto ortogonal*. Así, la *condición de ortogonalidad* de cualquier conjunto de funciones se suele escribir:

$$\int_a^b u(x)\phi_i(x)\phi_j(x)dx = A_{ij}\delta_{ij} \qquad \textbf{[3.67]}$$

Donde δ_{ij} es la delta de Kronecker. Si $A_{ii}=1$ $\forall i$, el conjunto de funciones es *ortonormal*, y la condición anterior es en dicho caso la de *ortonormalidad* del sistema.

Resulta corriente absorber ***u(x)*** dentro de de las funciones del conjunto ortogonal. Además, si un sistema de funciones $\phi_1^*(x)$, $\phi_2^*(x)$,... es ortogonal en un intervalo [a,b], es posible obtener un conjunto ortonormal $\{\phi_i(x)\}$ en el mismo intervalo, absorviendo además ***u(x)*** dentro de las funciones de dicho conjunto, de la siguiente manera:

$$\phi_i(x) = \sqrt{\frac{u(x)}{A_{ii}}}\,\phi_i^*(x) \qquad \textbf{[3.68]}$$

Donde A_{ii} se puede obtener a partir de 3.67 con i=j:

$$A_{ii} = \int_a^b u(x)[\phi_i^*(x)]^2 dx \qquad \textbf{[3.69]}$$

Por otro lado, un conjunto de funciones $f_1(x),f_2(x),...,f_m(x)$ se dice que es *linealmente independiente* si no existe ningún conjunto de coeficientes $c_1,c_2,...,c_m$ que no sean todos nulos tales que:

$$c_1 f_1(x)+c_2 f_2(x)+...+c_m f_m(x) = 0 \qquad \textbf{[3.70]}$$

Se cumpla $\forall x$. Las funciones de un conjunto ortogonal son todas linealmente independientes.

Por último, un conjunto de funciones se denomina *completo* si cualquier función contínua a tramos puede aproximarse de manera todo lo próxima que se desee a partir de una combinación lineal de funciones del conjunto. Un ejemplo de conjunto ortonormal completo está constituido por las funciones del bien conocido desarrollo en serie de Fourier, que permite aproximar cualquier función contínua a tramos como una combinación lineal de funcionales de dicho desarrollo.

Los conjuntos de funciones ortonormales unidimiensionales en diferentes intervalos han sido clásicamente objeto de estudio de las matemáticas operacionales. Sin embargo, los conjuntos de funciones ortonormales multivariantes, es decir, definidas en un espacio n-dimensional, no han sido estudidas tan exahustivamente debido a su propia complejidad. A continuación mostraremos un método que permite generar conjuntos completos de funciones ortonormales multidimensionales a partir de conjuntos completos de funciones ortonormales unidimensionales.

b) GENERACION DE FUNCIONES ORTONORMALES MULTIVARIANTES

Dado un sistema completo de funciones ortonormales de una variable $\phi_1(x)$, $\phi_7(x)$,... en el intervalo $a \le x \le b$, Se puede construir un sistema ortonormal completo de funciones de dos variables $\{\varphi_k(\mathbf{X})\}$, siendo $\mathbf{X}=(x_1,x_2)^t$, de la siguiente manera **[COUR-55]**:

$$\begin{aligned}
\varphi_1(x_1,x_2) &= \phi_1(x_1)\phi_1(x_2) \\
\varphi_2(x_1,x_2) &= \phi_1(x_1)\phi_2(x_2) \\
\varphi_3(x_1,x_2) &= \phi_2(x_1)\phi_1(x_2) \\
\varphi_4(x_1,x_2) &= \phi_2(x_1)\phi_2(x_2) \\
\varphi_5(x_1,x_2) &= \phi_1(x_1)\phi_3(x_2) \\
&\vdots
\end{aligned} \qquad \textbf{[3.71]}$$

Es decir, la regla de construcción se basa simplemente en tomar parejas de funciones del conjunto unidimensional y multiplicarlas, después de sustituir la variable x por x_1 y x_2. El orden en que se toman las funciones unidimensionales no es relevante, basta con que se preserve el de las variables.

Las funciones $\varphi_k(\mathbf{X})$ así obtenidas constituyen un conjunto ortonormal en la región cuadrada $a \le x_1 \le b$, $a \le x_2 \le b$. La extensión de este procedimiento a cualquier caso n-dimensional es directa; consiste en multiplicar entre si grupos de n funciones del conjunto unidimensional despues de la sustitución de x por $x_1,x_2,x_3,x_4,....x_n$. Así, si las

funciones del conjunto inicial son ortonormales en el intervalo $a \le x \le b$, las multidimensionales generadas son tambien ortonormales en el hipercubo definido por $\{a \le x_j \le b \ \forall j=1,2,...,n$.

Por ejemplo, las funciones de un conjunto ortonormal multivariante con n=4 $\varphi_k(\mathbf{X})=\varphi_k(x_1,x_2,x_3,x_4)$ se generan de la siguiente manera:

$$\begin{aligned}
\varphi_1(X) &= \phi_1(x_1)\phi_1(x_2)\phi_1(x_3)\phi_1(x_4) \\
\varphi_2(X) &= \phi_1(x_1)\phi_1(x_2)\phi_1(x_3)\phi_2(x_4) \\
\varphi_3(X) &= \phi_1(x_1)\phi_1(x_2)\phi_2(x_3)\phi_1(x_4) \\
\varphi_4(X) &= \phi_1(x_1)\phi_1(x_2)\phi_2(x_3)\phi_2(x_4) \\
\varphi_5(X) &= \phi_1(x_1)\phi_2(x_2)\phi_1(x_3)\phi_1(x_4) \\
&\vdots
\end{aligned} \qquad [3.72]$$

3.4.2.2.- FUNCIONES POTENCIALES DE TIPO 2

Las funciones de este tipo son simétricas respecto a las dos variables $\mathbf{X}$ y $\mathbf{X}_i$, presentan valor máximo para $\mathbf{X}=\mathbf{X}_i$ y decrecen monótonamente según se incrementa la distancia $\mathbf{X}-\mathbf{X}_i$, como se ha comentado anteriormente. Ejemplos de estas funciones son:

$$\begin{aligned}
U(X,X_i) &= e^{-\alpha\|X-X_i\|^2} \\
U(X,X_i) &= \frac{1}{1+\alpha\|X-X_i\|^2} \\
U(X,X_i) &= \left|\frac{sen[\alpha\|X-X_i\|^2]}{\alpha\|X-X_i\|^2}\right|
\end{aligned} \qquad [3.73]$$

Donde α es una constante positiva y $\|\mathbf{X}-\mathbf{X}_i\|$ indica la norma o módulo del vector $\mathbf{X}-\mathbf{X}_i$.

Gráficas en casos unidimensionales y bidimensionales de estas funciones se pueden encontrar en la figura 3.6.

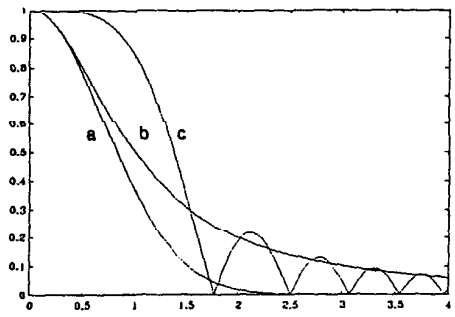


Figura 3.6.: Gráfica de las Funciones Potenciales de la expresión 3.73 en Caso Unidimensional.

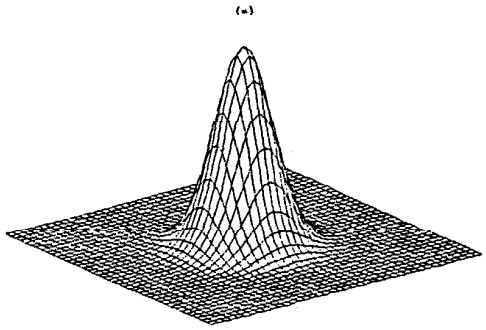


Figura 3.7a: Gráfica de la Función Potencial (a) de la expresión 3.73

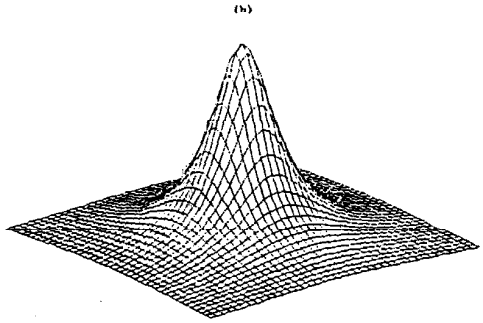


Figura 3.7b: Gráfica de la Función Potencial (b) de la expresión 3.73 en caso Bidimensional

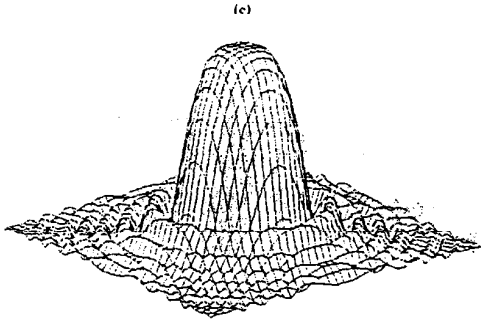


Figura 3.7c: Gráfica de la Función Potencial (c) de la expresión 3.73 en caso Bidimensional

3.4.3.- PROCEDIMIENTO DE APRENDIZAJE MULTICLASICO

En un problema con c clases, el procedimiento de aprendizaje de las funciones de decisión $\{d_i(\mathbf{X}), \forall i=1,2,...,c\}$ según el método de las funciones potenciales, y con una regla de decisión como la siguiente:

$$Si\ d_I(X) = \max_{\forall j=1,2,...,c} d_j(X)\ Entonces\ X \in \Omega_i \qquad [3.74]$$

Se puede resumir como se indica a continuación:

Partiendo del conjunto de m *muestras de aprendizaje y definiendo las funciones de decisión iniciales* $d_i^{[0]}(\mathbf{X})$ *que por simplicidad pueden ser nulas, en el paso* k *(siendo la muestra* $\mathbf{X}[k] \varepsilon \Omega_i$*), las operaciones a realizar son las siguientes:*

$$SI\ \left(d_i^{[k-1]}(X[k]) = \max_{\forall j=1,2,...,c} d_j^{[k-1]}(X[k])\right)\ Entonces\ d_j^{[k]}(X) = d_j^{[k-1]}\ \forall j=1,2,...,c$$

En caso contrario {

$\forall I$ *tal que* $d_I^{[k-1]}(X[k]) \geq d_i^{[k-1]}(X[k])$ *hacer* $d_I^{[k]}(X) = d_I^{[k-1]}(X) - U(X,X[k])$

para i hacer $d_i^{[k]}(X) = d_i^{[k-1]} + U(X,X[k])$

para el resto hacer $d_j^{[k]}(X) = d_j^{[k-1]}(X)$

}

3.5.- PERCEPTRON MULTICAPA

En el apartado 3.3.1 hemos hecho referencia al perceptrón de una capa y hemos establecido su correspondencia con una regla de discriminación entre clases basada en discriminante lineal. Ahora bien, tambien es posible definir discriminaciones para clases no linealmente separables utilizando perceptrones multicapa, que son redes sin realimentaciones (feed-forward) con una o más capas de nodos entre la capa de entrada y la de salida. Estas capas adicionales contienen unidades ocultas, o nodos que no están conectados directamente a la entrada y a la salida. En la figura 3.8 se muestra una red perceptrón multiclásica de tres capas, con dos capas de unidades ocultas. A efectos de unificar la nomenclatura de los análisis que posteriormente haremos con los del apartado 3.3, los pesos de las conexiones de la red los hacemos corresponder con los elementos de los vectores de pesos en los discriminantes lineales.

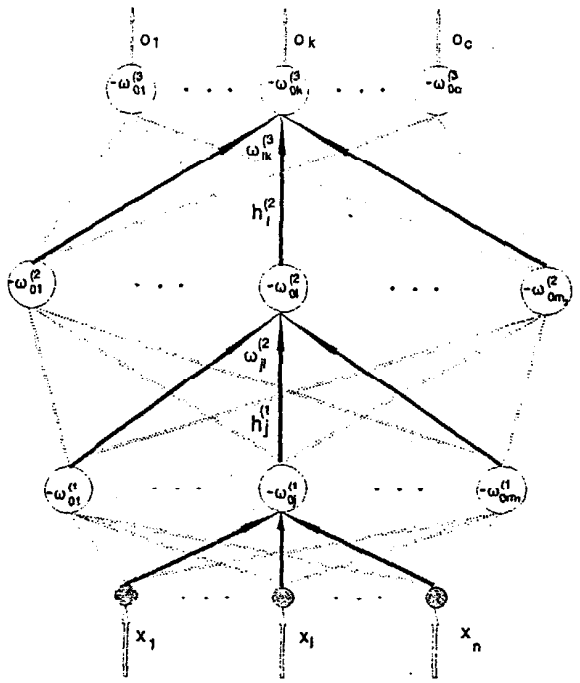


Figura 3.8: Perceptrón Multiclásico

3.5.1.- DESCRIPCION Y PROPIEDADES

Un perceptrón multicapa es una red con n entradas, luego, capaz de generar decisiones en un espacio de medidas n-dimensional, y c salidas, por lo que es posible discriminar entre c clases. La descripción analítica de su estructura es la siguiente. Cada una de las c salidas es:

$$d_k(X) = o_k = g_k^{(3}(z_k) \qquad [3.75]$$

Donde $g^{(3}_k$ es la función de activación nolineal asociada al nodo k de la capa de salida. Dicha función puede ser de tipo signo, como se ha dicho en 3.3.1 o una función sigmiodal como, por ejemplo:

$$g(r) = th(\beta r) = \frac{e^{\beta x}-e^{-\beta x}}{e^{\beta x}+e^{-\beta x}} = \frac{2}{1+e^{-2\beta r}}-1 \qquad [3.76]$$

El funcional presenta asíntotas horizontales en +1 y -1, La gráfica de la curva es la indicada en la figura 3.9.

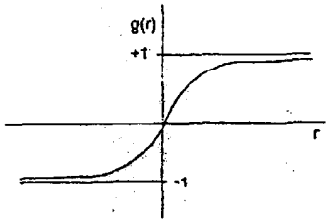


Figura 3.9: Función sigmoidal asintótica

Siguiendo con la descripción de la red, z_k en 3.75 es:

$$z_k = \sum_{l=1}^{m_2} \omega_{lk}^{(3} h_l^{(2} + \omega_{0k}^{(3} = \qquad [3.77]$$

$$= \sum_{l=0}^{m_2} \omega_{lk}^{(3} h_l^{(2} \quad ; \; siendo \; h_0^{(2} = 1$$

Donde los $\omega^{(3}_{lk}$ son los pesos de las conexiones de cada nodo l de la segunda capa oculta al nodo k de la capa de salida. A su vez, $h^{(2}_l$ se corresponde con la salida de la segunda capa oculta, cuya relación de entrada-salida es:

$$h_l^{(2} = g_l^{(2}(v_l^{(2}) \quad ; \quad v_l^{(2} = \sum_{j=0}^{m_1} \omega_{jl}^{(2} h_j^{(1} \qquad [3.78]$$

Siendo $\omega^{(2}_{jl}$ los pesos correspondientes a las conexiones de los nodos de la primera a la segunda capa oculta. Por último, la relación entrada-salida en la primera capa oculta es, análogamente:

$$h_j^{(1} = g_j^{(1}(v_j^{(1}) \;\; ; \;\; v_j^{(1} = \sum_{i=1}^{n} \omega_{ij}^{(1} x_i + \omega_{0j}^{(1} \qquad \textbf{[3.79]}$$

Se puede establecer la decisión de una muestra incógnita **X** entre las c clases utilizando la ya conocida regla de mayoría:

$$\textit{Para } \mathbf{X} \textit{ si: } d_\lambda(X) = \max_{\forall k = 1,2,...c} \{d_k(X)\} \;\; \textit{entonces } X \in \Omega_\lambda \qquad \textbf{[3.80]}$$

La capacidad del perceptrón multicapa para discriminar entre clases separadas por límites más complejos que el hiperplano se deriva de la presencia a la salida de cada nodo de la nolinealidad recogida en las funciones g. Las capacidades de perceptrones con una, dos y tres capas que hacen uso de nolinealidades de tipo signo se recogen en la tabla 3.1. En la segunda columna se indican los tipos de regiones limitadas por superficies de decisión que pueden generarse, las dos siguientes columnas presentan ejemplos de regiones de decisión que pueden formarse para tratar el denominado problema del "OR exclusivo" y los correspondientes a regiones engarzadas. La última columna recoge ejemplos de las superficies de decisión más generales que pueden obtenerse en cada caso.

Como se he mencionado anteriormente, un perceptrón de una capa genera superficies de decisión que son hiperplanos. Un perceptrón de dos capas es capaz de definir regiones convexas en el espacio de características, regiones que pueden ser, o abiertas o conchas convexas (convex hulls), entendiendo así a aquellas regiones con topología tal que cualquier línea que une dos puntos de los límites de la región, pasa solamente por otros puntos que son tambien de la región.

Las regiones convexas se forman como intersección de regiones semiplano definidas por cada nodo de la primera capa oculta del perceptrón multicapa. Cada nodo de la primera capa se comporta como un perceptrón de una capa, que separa puntos del espacio mediante un hiperplano, definido por su vector de peso y peso umbral. La combinación de las salidas de diversos nodos de la primera capa en un nodo de la segunda capa con su vector de pesos y peso umbral convenientemente definidos, permite definir una región convexa como intersección de los semiplanos definidos en la primera capa, lo que se corresponde a efectuar una operación de AND lógico en el nodo de salida. La región convexa tendrá tantos lados como nodos en la primera capa.

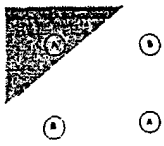
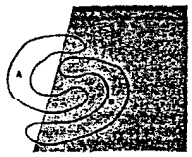
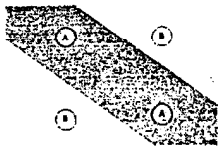
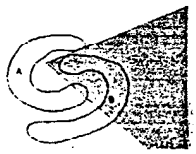
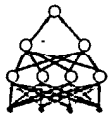
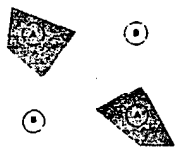
ESTRUCTURA	TIPOS DE REGIONES DE DECISION	PROBLEMA DEL OR EXCLUSIVO	CLASES CON REGIONES ENGARZADAS	FORMAS DE REGIONES MAS GENERALES
UNA CAPA 	REGION LIMITADA POR UN HIPERPLAN O			
DOS CAPAS 	REGIONES CONVEXAS ABIERTAS O CERRADAS			
TRES CAPAS 	ARBITRARIO (complejidad limitada por el número de nodos)			

Tabla 3.1

El análisis anterior suministra algunas pistas acerca de la selección del número de nodos a utilizar en un perceptrón de dos capas: el número de nodos debe ser suficiente para formar una región tan compleja como la requerida por el problema, pero no deben ser tantos que no resulte útil la estima de los pesos por un procedimiento de aprendizaje.

Un perceptrón de tres capas sin realimentación puede generar fronteras de decisión arbitrariamente complejas y puede separar clases engarzadas como se muestra en la tabla 3.1, lo que se puede mostrar con un pequeño análisis. El espacio de características se puede dividir en hipercubos (cuadrados en el caso de espacio de dimensión dos). Para generar cada hipercubo se requiere que en la primera capa oculta existan 2n nodos, uno por cada cara del hipercubo siendo n la dimensionalidad del espacio (así, para dimensión dos, se precisan cuatro nodos), así como un nodo en la segunda capa oculta para efectuar el AND lógico de los de la primera capa. Así, las salidas de los nodos de la segunda capa oculta serán de valor "alto" sólo si el vector de características de la entrada se corresponde a un punto en el interior del hipercubo. Los hipercubos así definidos se asignan a la región de la clase correspondiente conectando la salida correspondiente a la segunda capa oculta al nodo de salida correspondiente a su clase, que realiza una operación de OR lógico. El análisis anterior puede genralizarse a la utilización de regiones convexas cualesquiera, con lo cual será posible generar regiones no convexas e incluso desconectadas como se muestra en la tabla 3.1.

Un problema añadido, del que podemos dar unas reglas mínimas, es el número de nodos que debe poseer el perceptrón de tres capas. De entrada, son validos en este sentido, los comentarios del párrafo anterior referentes a la primera capa son válidos. Para la segunda capa debe ser mayor que uno, si las regiones de una misma clase se encuentran desconectadas o las clases se encuentran engarzadas y no se pueden formar a partir de un área convexa. En el peor caso debe ser igual al número de regiones desconectadas.

Las discusiones anteriores se han centrado, por simplicidad, en análisis intuitivos de perceptrones multicapa con nolinealidades de tipo signo en los nodos. Lo anterior es transportable tambien al caso de nolinealidades reflejadas en funciones continuas sigmoidales, aunque el comportamiento de estas redes es más complejo, ya que las hipersuperficies de decisión son curvas suaves en vez segmentos de línea recta, con lo que el análisis resulta por ello más dificil. Sin embargo, presentan la ventaja de que pueden entrenarse con un nuevo procedimiento denominado de *propagación hacia atrás* que describiremos a continuación.

3.5.2.- APRENDIZAJE POR RETROPROPAGACIÓN

El aprendizaje por retropropagación (back-propagation) utiliza el método de descenso según el gradiente para determinar iterativamente los pesos de la red a partir de un conjunto de aprendizaje, constituido por p muestras de entrada, que ordenamos de la siguiente manera $\{\mathbf{X}[1], \mathbf{X}[2],...\mathbf{X}[\mu],...\}$, y las salidas respectivas, vectores que denominaremos $\{\xi[1],\xi[2],...\xi[\mu],...\}$. La estructura de red en la que nos centraremos

es del tipo de la figura 3.8, con nolinealidades continuas y diferenciables en los nodos. La función criterio a minimizar es el error cuadrático entre las salidas de la red, para la configuración actual de pesos, y las salidas esperadas para el conjunto de aprendizaje, es decir:

$$J[\mu] = \frac{1}{2}\sum_{k=1}^{c}(\xi_k[\mu] - o_k[\mu])^2 \qquad [3.81]$$

Esta función será mínima cuando la diferencia entre salidas esperadas y salidas reales sea mínima. La corrección de pesos por descenso según el gradiente se puede expresar como:

$$\Delta\omega_{ij}^{(b}[\mu] = \omega_{ij}^{(b}[\mu+1] - \omega_{ij}^{(b}[\mu] = -\rho\left[\frac{\partial J[\mu]}{\partial\omega_{ij}^{(b}}\right]_{\omega_{ij}^{(b} = \omega_{ij}^{(b}[\mu]} \qquad [3.82]$$

Donde b=1,2,...,p_h, siendo p_h el número de capas de la red. Para determinar la expresión de corrección de cada peso se procede por capas. Para simplificar las expresiones, no incluimos en ellas el número de orden μ de la muestra. Para una red de tres capas:

I) Pesos de las conexiones de la segunda capa oculta a la salida:

$$\Delta\omega_{ik}^{(3} = -\rho\frac{\partial J}{\partial o_k}\frac{\partial o_k}{\partial\omega_{ik}^{(3}} = \rho(\xi_k - o_k)\frac{dg_k^{(3}}{dz_k}h_i^{(2} = \rho\delta_k^{(3}h_i^{(2} \qquad [3.83]$$

Donde hemos definido:

$$\delta_k^{(3} = (\xi_k - o_k)\frac{dg_k^{(3}}{dz_k} \qquad [3.84]$$

II) Para las conexiones de la primera a la segunda capa oculta, derivamos respecto a los pesos $\omega^{(2}_{ji}$, que se encuentran empotrados en la red más profundamente que los anteriores. Utilizando la regla de encadenamiento obtenemos:
Donde, para mantener analogía con 3.83 , hemos definido:

$$\Delta\omega_{jl}^{(2} = -\rho\sum_{k=1}^{c}\frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial h_l^{(2}}\frac{\partial h_l^{(2}}{\partial \omega_{jl}^{(2}} =$$

$$= \rho\frac{dg_l^{(2}}{dv_l^{(2}}\left[\sum_{k=1}^{c}\delta_k^{(3}\omega_{lk}^{(3}\right]h_j^{(1} = \qquad [3.85]$$

$$= \rho\delta_l^{(2}h_j^{(1}$$

$$\delta_l^{(2} = \frac{dg_l^{(2}}{dv_l^{(2}}\sum_{k=1}^{c}\delta_k^{(3}\omega_{lk}^{(3} \qquad [3.86]$$

III) Para las conexiones de la entrada a la primera capa oculta, igual que en el caso anterior, la variación de los pesos puede obtenerse por encadenamiento:

$$\Delta\omega_{ij}^{(1} = -\rho\sum_{k=1}^{c}\frac{\partial J}{\partial z_k}\sum_{l=0}^{m_2}\frac{\partial z_k}{\partial h_l^{(2}}\frac{\partial h_l^{(2}}{\partial h_j^{(1}}\frac{\partial h_j^{(1}}{\partial \omega_{ij}^{(1}} =$$

$$= \rho\frac{dg_j^{(1}}{dv_j^{(1}}\left[\sum_{l=0}^{m_2}\delta_l^{(2}\omega_{jl}^{(2}\right]x_i = \qquad [3.87]$$

$$= \rho\delta_j^{(1}x_i$$

Donde:

$$\delta_j^{(1} = \frac{dg_j^{(1}}{dv_j^{(1}}\sum_{l=0}^{m_2}\delta_l^{(2}\omega_{jl}^{(2} \qquad [3.88]$$

Las ecuaciones 3.83, 3.85 y 3.87 presentan la misma estructura, solo variando en los coeficientes 3.84, 3.86 y 3.88, por lo que, en general, con un número arbitrario de capas, la regla de corrección de pesos por propagación hacia atrás tiene la forma:

$$\Delta \omega_{pq}^{(b} = \rho \delta_{salida}^{(b} h_{entrada}^{(b-1} \qquad [3.89]$$

Donde: *salida* y *entrada* se refieren a los dos extremos p y q de la conexión considerada y $h^{(b-1}_{entrada}$, a la entrada de activación de la unidad correspondiente al peso considerado, que puede ser la salida de otra unidad o una entrada a la red, según corresponda.

Las ecuaciones 3.86 y 3.88 nos permiten determinar los "errores" δ para una cierta unidad oculta a partir de los correspondientes δ de las unidades de salida o_k a los que se encuentra conectada. Como podemos ver por las expresiones, las entradas a la red se propagan "hacia adelante" en la misma hasta la salida, pero los errores δ se propagan "hacia atrás" desde salida a entrada en el proceso de aprendizaje. Esta forma de propagación de las modificaciones es la causa del nombre que se da al procedimiento de aprendizaje.

Las funciones de activación g a utilizar deben ser funciones diferenciables, dada la naturaleza de las expresiones de corrección de pesos, siendo válidas expresiones sigmoidales como las 3.76, donde en el caso de la exponencial, el parámetro β suele definirse típicamente a 1 ó 1/2. Normalmente las derivadas de estas funciones de activación suelen expresarse en términos de las mismas funciones, con lo que para las recogidas en 3.76 tiene la forma:

$$\frac{dg(r)}{dr} = \beta[1 - g^2(r)] \qquad [3.90]$$

3.5.3.- PROCEDIMIENTO DE APRENDIZAJE

El procedimiento de aprendizaje por retropropagación utiliza el esquema iterativo de ajuste de pesos por descenso según el gradiente para el ajuste de pesos que se aplica paso a paso, tomando una muestra del conjunto de aprendizaje cada vez.

Definidas:

a) La muestra de aprendizaje, que ordenamos como se indicó en el párrafo anterior (μ=1,2,...)

b) La estructura de la red en cuanto a número de capas y número de nodos por capa

c) La función de activación g(r) y el factor de corrección ρ

El procedimiento comienza *inicializando todos los pesos y pesos umbrales de las conexiones de la red a pequeños valores aleatorios.* En este punto se entra en el esquema iterativo de ajuste, que para el paso μ en una red con tres capas se puede resumir como sigue:

1) *Tomar el vector de entrada* **X**[μ] *y el de salida* ξ[μ] *de entre los del conjunto de aprendizaje.*

2) *Propagar la señal generada por el vector de entrada hacia adelante en la red, capa a capa hasta la salida, utilizando las expresiones* 3.79, 3.78, 3.77 y 3.75 *en este orden.*

3) *Calcular los* δ *de la capa de salida, comparando las salidas obtenidas por la red con las esperadas según la muestra de aprendizaje, aplicando la expresión* 3.84. *Propagar a continuación hacia atrás el cálculo de los* δ *de las capas interiores por aplicación de las ecuaciones* 3.86 *y* 3.88 *en este orden.*

4) *Calcular las correcciones de los pesos de las conexiones en las diferentes capas aplicando las expresiones de corrección* 3.83, 3.85 *y* 3.87, *el nuevo peso de cada conexión se calcula según* 3.82, *es decir, según la expresión:*

$$\omega_{pq}[\mu+1] = \omega_{pq}[\mu]+\Delta\omega_{pq}[\mu]$$

El proceso se repite cíclicamente sobre cada muestra del conjunto de aprendizaje, presentandose dicho conjunto tantas veces como sea necesario, hasta que los pesos se estabilicen.

3.5.4.- COMENTARIOS ADICIONALES

El procedimiento de propagación hacia atrás mostrado anteriormente precisa de la definición del factor de corrección ρ, que define la "velocidad" de aprendizaje. No existe una demostración que defina el rango de valores que debe poseer dicho coeficiente para asegurar la convergencia, por lo que, en la práctica se suele escoger un valor pequeño para asegurar que no se produzcan oscilaciones en los valores de los ajustes, que implican oscilaciones en el valor de la función criterio. Este hecho, unido a la propia estructura de la red y del procedimiento de aprendizaje, hace que el número de presentaciones del conjunto de muestras de aprendizaje al procedimiento sea relativamente elevado (generalmente más de 100 presentaciones de todo el conjunto).

Anteriormente se afirmó que un perceptrón de tres capas es capaz de generar funciones de decisión de cualquier grado de complejidad. Esta afirmación viene avalada por el **Teorema de Kolmogorov acerca de la Existencia de Transformaciones en Redes Neuronales** que establece: *Dada cualquier función continua* $f:[0,1]^n \rightarrow \mathbf{R}^c$, $f(\mathbf{X})=\mathbf{o}$, *se puede implementar con una red neuronal de tres capas sin realimentaciones, que contenga* n *elementos en la primera capa*, 2n+1 *en la segunda y* c *en la de salida, utilizando nolinealidades crecientes contínuamente en los nodos.* Desafortunadamente, el teorema no establece nada acerca de como seleccionar los pesos o nolinealidades de los nodos, ni tampoco nada acerca de cual es la sensibilidad de la función de salida a las variaciones de los pesos y funciones internas.

3.6.- REFERENCIAS

[ABRA-68] Abramowitz M., Stegun I. A. (eds.), **Handbook of Mathematical Functions**, Doves Pub. Inc., New York, 1968.

[BOW-84] Bow S., **Pattern Recognition. Applications to Large Data-Set Problems**, Marcel Dekker, Inc., New York, 1984.

[BOW-92] Bow S., **Pattern Recognition and Image Preprocessing**, Marcel Dekker, Inc., New York, 1992.

[CHEN-93] Chen C. H., L. F. Pau, P. S. P. Pau (eds.), **Handbook of Pattern Recognition & Computer Vision**, World Scientific, Singapore, 1993.

[CHIE-78] Chien Y., **Interactive Pattern Recognition**, Marcel Dekker, Inc., New York, 1978.

[COUR-55] Courant R., Hilbert D., **Methods of Mathematical Physics, vol. 1**, Interscience Pub., New York, 1955.

[DUDA-73] Duda R., Hart P., **Pattern Classification and Scene Analysis**, Wiley, New York, 1973.

[FREE-91] Freeman J. A., Skapura D. M., **Neural Networks. Algorithms, Applications, and Programming Techniques**, Addison-Wesley Pub. Co., Reading, Massachusetts, 1991.

[GALL-94] Gallant S. I., **Neural Networks Learning and Expert Systems**, The MIT Press, Cambridge, Massachusetts, 1994.

[HAND-81] Hand D. J., **Discrimination and Classification**, John Wiley & Sons, Chichester, 1981.

[HAYK-94] Haykin S., **Neural Networks. A Comprehensive Foundation**, McMillan College Pub. Co., New York, 1994.

[HERT-91] Hertz J., Krogh A., Palmer R. G., **Introduction to the Theory of Neural Computation**, Addison-Wesley Pub. Co., Redwood City, CA, 1991.

[HETC-90] Hetch-Nielsen R., **Neurocomputing**, Addison-Wesley Pub. Co., Reading, Massachusetts, 1990.

[LIPP-87] Lippmann R. P., *An Introduction to Computing with Neural Nets*, IEEE ASSP Magazine, pp. 4-22, April, 1987.

[MINS-88] Minsky M. L., Papert S. A., **Perceptrons. An Introduction to Computational Geometry**, Expanded Edition, The MIT Press, Cambridge, Massachusetts, 1988.

[NILS-90] Nilsson N., **Learning Machines**, Morgan Kaufmann Publishers, San Mateo, California, 1990.

[RUME-86] Rumelhart D. E., Hinton G. E., Williams R. J., *Learning Internal Representations by Error Propagation*, en Rumelhart D. E., MacClelland J. L. (eds.), **Parallel Distributed Processing. Explorations in the Microestructure of Cognition, vol. 1: Foundations**, The MIT Press, Cambridge, Massachusetts, 1986.

[SCHA-92] Schalkoff R., **Pattern Recognition. Statistical, Structural and Neural Approaches**, John Wiley & Sons, Inc., 1992.

[TOU-74] Tou J., Gonzalez R., **Pattern Recognition Principles**, Addison-Wesley, Reading, Massachusetts, 1973.

[WEIS-91] Weiss S., Kulikowski C. A., **Computer Systems that Learn**, Morgan Kaufmann Pub., Inc., San Francisco, CA, 1991.

3.5.- COMENTARIOS ADICIONALES

Los procedimientos presentados a lo largo del tema emplean muestras de aprendizaje para generar los coeficientes de las funciones de decisión. Para obtener funciones de decisión suficientemente robustas o generales, los vectores de la muestra de aprendizaje deben ser representativos de las clases en consideración.

Una cuestión importante referente al problema de aprendizaje es la siguiente: ¿cuántas muestras deben escogerse para obtener clasificadores suficientemente robustos?. La respuesta intuitiva es: tantas como sea posible. La solución práctica, sin embargo, no puede ser en general esa, ya que es una regla que choca frontalmente con la economía y comodidad del diseño. En un esquema determinista como es el caso que nos ocupa, Coves **[COVE-65]** determinó que el número total de vectores de la muestra de aprendizaje para resolver un problema biclásico debe ser como mínimo dos veces la dimensionalidad del espacio para los procedimientos perceptrón y de mínimo error cuadrático, de manera que se pueda asegurar suficiente robustez en el clasificador diseñado.