Reconocimiento de Patrones

J. Kittler (revisado y ampliado por GTI-IIE) Revisión:0.9, Fecha: 1/09/2002

Notas del seminario de Reconocimiento de Patrones de Grupo de Tratamiento de Imágenes del Instituto de Ingeniería Eléctrica, basado en las notas del curso del Prof. J. Kittler en la Univ. de Surrey.

Índice

1. Modelo de Sistema de Reconocimiento de Patrones

1.1. Introducción

El objetivo del procesamiento e interpretación de datos sensoriales es lograr una descripción concisa y representativa del universo observado. La información de interés incluye nombres, características detalladas, relacionamientos, modos de comportamiento, etc. que involucran a los elementos del universo (objetos, fenómenos, conceptos)

Estos elementos se perciben como patrones y los procesos que llevan a su comprensión son llamados *procesos perceptuales*. El etiquetado (clasificación, asignación de nombres) de esos elementos es lo que se conoce como *reconocimiento de patrones*. Por lo tanto, el reconocimiento de patrones es una herramienta esencial para la interpretación automática de datos sensoriales.

El sistema nervioso humano recibe aproximadamente 10^9 bits de datos sensoriales por segundo y la mayoría de esta información es adquirida y procesada por el sistema visual. Análogamente, la mayoría de los datos a ser procesados automáticamente aparecen en forma de imágenes.

El procesamiento de imágenes de escenas complejas es un proceso en múltiples niveles que se ilustra en la figura 1 mostrando la participación relativa de los dos tipos de metodologías necesarias:

- Reconocimiento de patrones basado en atributos.
- Reconocimiento de patrones basado en la estructura.

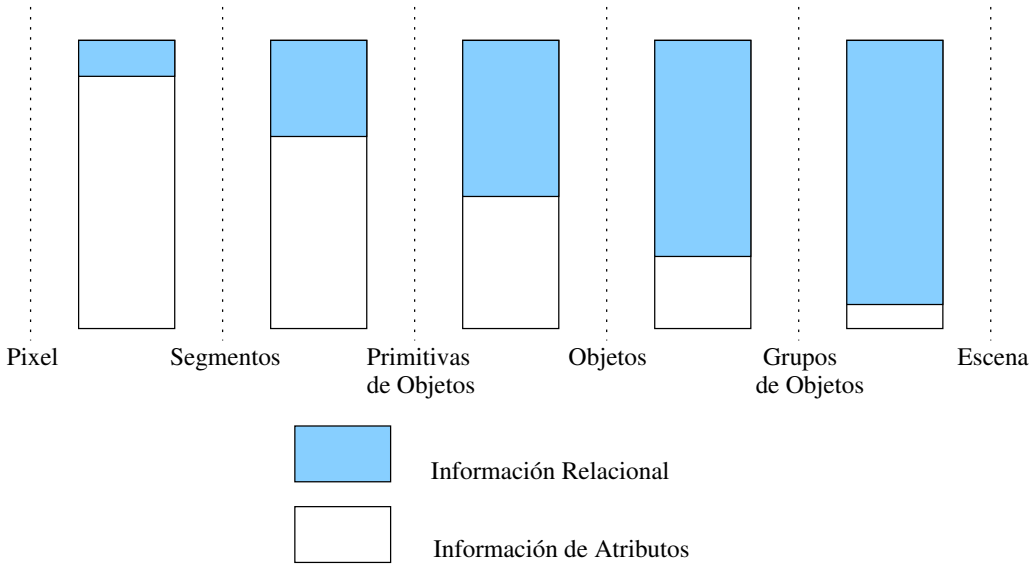

Figura 1: Distintos tipos de información usada en diferentes niveles de procesamiento.

1.2. Modelo de Sistema de Reconocimiento de Patrones

Los procesos perceptuales del ser humano pueden ser modelados como un sistema de tres estados:

- adquisición de datos sensoriales
- extracción de características
- toma de decisiones

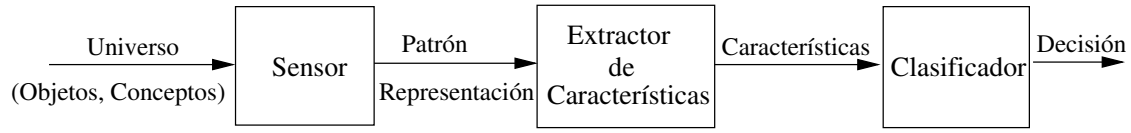

Figura 2: Etapas en un sistema de reconocimiento de patrones.

Por lo tanto es conveniente dividir el problema del reconocimiento automático de una manera similar

Sensor Su propósito es proporcionar una representación feasible de los elementos del universo a ser clasificados. Es un sub-sistema crucial ya que determina los límites en el rendimiento de todo el sistema.

Idealmente uno debería entender completamente las propiedades físicas que distinguen a los elementos en las diferentes clases y usar ese conocimiento para diseñar el sensor, de manera que esas propiedades pudieran ser medidas directamente. En la práctica frecuentemente esto es imposible porque:

- no se dispone de ese conocimiento
- muchas propiedades útiles no se pueden medir directamente (medición no intrusiva)
- no es económicamente viable

Extracción de Características Esta etapa se encarga, a partir del patrón de representación, de extraer la información discriminatoria eliminando la información redundante e irrelevante. Su principal propósito es reducir la dimensionalidad del problema de reconocimiento de patrones.

Clasificador El la etapa de toma de decisiones en el sistema. Su rol es asignar a la categoría apropiada los patrones de clase desconocida a priori.

1.3. Modelo del Proceso de Generación de Patrones

1.3.1. Modelo Probabilístico

Las diferencias en los patrones de una misma clase puede deberse a ruido, deformaciones, variabilidad biológica, etc. Por lo tanto debemos asumir esta variabilidad en los patrones y el proceso asociado a la generación de patrones puede ser descripto adecuadamente mediante un modelo probabilístico.

De lo anterior podemos asumir que cada patrón

$$\mathbf{x} = [x_1, x_2, \cdots, x_n]$$

es un vector aleatorio n-dimensional perteneciente a una de m posibles clases ω_i $i = 1, \cdots, m$ donde cada clase ω_i tiene una probabilidad de ocurrencia *a priori* igual a $P(\omega_i)$. La distribución de probabilidad del vector patrón $\mathbf{x}$ de la clase ω_i se caracteriza por la función densidad de probabilidad condicional para la i-esima clase $p(\mathbf{x}|\omega_i)$.

1.3.2. Relaciones Básicas

Notar que las probabilidades a priori de las clases suman uno, o sea

$$\sum_{i=1}^{m} P(\omega_i) = 1$$

La densidad conjunta o función de densidad de probabilidad no condicional $p(\mathbf{x})$ vienen dada por

$$p(\mathbf{x}) = \sum_{i=1}^{m} p(\mathbf{x}|\omega_i) P(\omega_i).$$

En la práctica nos interesa calcular la probabilidad *a posteriori* (una vez observado el patrón $\mathbf{x}$) para cada clase ω_i, la cual viene dada por la *Fórmula de Bayes* que relaciona probabilidades condicionales según:

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{m} p(\mathbf{x}|\omega_j)P(\omega_j)} \ .$$

1.3.3. Ejemplo: Un problema de reconocimiento de caracteres

- $m = 26$ número de diferentes caracteres excluyendo los dígitos.
- $n = 8$ número de medidas
- $x_i \ i = 1 \cdots n$ distancia entre el centro de gravedad y el punto de intersección más lejano en la semirrecta con origen en este centro y formando un ángulo $\frac{(i-1)\pi}{4}$ con el eje $0\vec{x}$.
- $P(\omega_i)$ probabilidad a priori de la ocurrencia del i-ésimo carácter en un lenguaje dado.

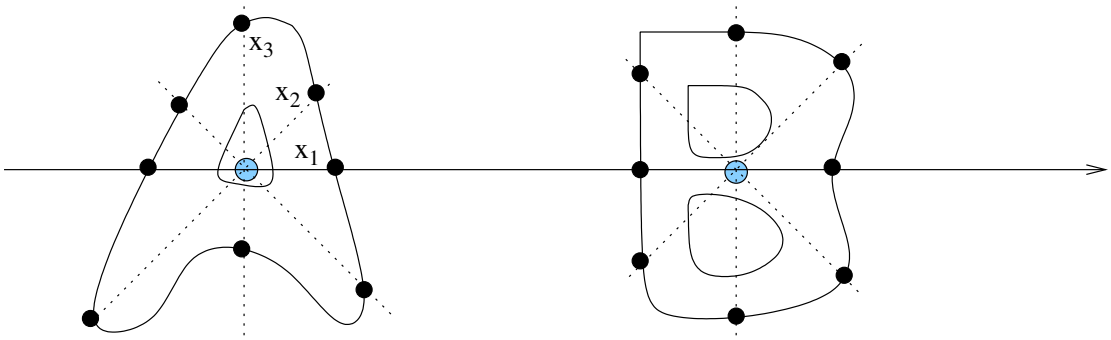

Figura 3: Atributos en reconocimiento de caracteres.

1.4. Reglas de Decisión Estadística

1.4.1. Regla del Mínimo Costo

Dadas las características del modelo probabilístico adoptado para el proceso de generación de patrones; cómo decidimos a que clase asignar el patrón $\mathbf{x}$ observado?

Para resolver este problema, definamos un costo de decisión $\rho_{ij} \ 1 \leq i, j \leq n$ asociado con la decisión de asignar a la clase ω_j un patrón $\mathbf{x}$ que pertenece a la clase ω_i.

Notar que

- ρ_{ii} costo de una decisión correcta, en general se define como 0
- ρ_{ij} es en general diferente de ρ_{ji}
- $\rho_{ij} \geq 0$

Ejemplo: Verificación de Firmas

En una aplicación de detección de firmas falsas tendríamos en principio dos clases

$$\begin{cases} \omega_1 \ \leftrightarrow \ \text{la firma es auténtica} \\ \omega_2 \ \leftrightarrow \ \text{la firma ha sido falsificada} \end{cases}$$

Claramente en este contexto podemos cometer 2 tipos de errores de clasificación que sin embargo implican costos muy diferentes; de modo que se deberá cumplir $\rho_{21} \gg \rho_{12}$.

Denotaremos con Γ_j la región del espacio de observación Ω tal que nuestra regla de decisión asigna

$$\mathbf{x} \rightarrow \omega_j \quad \forall\, \mathbf{x} \in \Gamma_j$$

o sea que la región Γ_j está asociada con la clase ω_j.

El costo medio de clasificar un patrón $\mathbf{x} \in \Gamma_j$ como perteneciente a la clase ω_j es

$$r_j(\mathbf{x}) = \sum_{i=1}^{m} \rho_{ij} P(\omega_i|\mathbf{x})$$

Por lo tanto el costo para toda la región Γ_j se obtiene integrando sobre todos los valores posibles con sus correspondientes probabilidades de observación:

$$R_j = \int_{\Gamma_j} r_j(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Finalmente el costo total de nuestro sistema de decisión viene dado por

$$R = \sum_{j=1}^{m} R_j = \sum_{j=1}^{m} \int_{\Gamma_j} \left(\sum_{i=1}^{m} \rho_{ij} P(\omega_i|\mathbf{x}) \right) p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

De modo que podemos concluir que el costo total será minimizado si el espacio de observación se particiona de manera tal que si $\mathbf{x} \in \Gamma_j$ se tenga

$$\sum_{i=1}^{m} \rho_{ij} P(\omega_i|\mathbf{x}) \leq \sum_{i=1}^{m} \rho_{ik} P(\omega_i|\mathbf{x}) \quad \forall\, k \neq j$$

Hemos llegado por lo tanto a la llamada *Regla de Decisión de Mínimo Costo de Bayes*, que establece

$$\text{Asignar } \mathbf{x} \rightarrow \omega_j \quad \Longleftrightarrow \quad \sum_{i=1}^{m} \rho_{ij} p(\mathbf{x}|\omega_i) P(\omega_i) = \min_{1 \leq k \leq m} \sum_{i=1}^{m} \rho_{ik} p(\mathbf{x}|\omega_i) P(\omega_i)$$

donde en en la última ecuación hemos usado la identidad de Bayes: $P(\omega_i|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$.

1.4.2. Regla del Mínimo Error

Consideramos ahora un modelo de costos cero-uno, o sea

$$\begin{cases} \rho_{ii} = 0 & \forall\, 1 \leq i \leq m \\ \rho_{ij} = 1 & \forall\, 1 \leq i, j \leq m,\ i \neq j \end{cases}$$

En este caso el lado derecho de la regla del mínimo costo queda

$$\sum_{i=1}^{m} \rho_{ik} P(\omega_i|\mathbf{x}) = \sum_{\substack{1 \leq i \leq m \\ i \neq k}} P(\omega_i|\mathbf{x}) = 1 - P(\omega_k|\mathbf{x})$$

y la regla de decisión correspondiente resulta:

$$\text{Asignar } \mathbf{x} \rightarrow \omega_j \quad \Longleftrightarrow \quad P(\omega_j|\mathbf{x}) = \max_{1 \leq k \leq m} P(\omega_k|\mathbf{x})$$

Observar que si asumimos que se asigna $\mathbf{x} \rightarrow \omega_j$, la probabilidad condicional de error $\epsilon(\mathbf{x})$ viene dada por

$$\epsilon(\mathbf{x}) = 1 - P(\omega_j|\mathbf{x})\,.$$

Por lo tanto el error condicional será mínimo si la clase ω_j se elije usando la última regla de decisión. En este caso el error medio

$$e = \int_{\Omega} \epsilon(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

se conoce como *error de Bayes*.

Integrando en cada región de decisión este error se puede descomponer como

$$e = 1 - \sum_{j=1}^{m} \Big(\int_{\Gamma_j} P(\omega_j|\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} \Big) = 1 - \sum_{j=1}^{m} P(\omega_j) \Big(\int_{\Gamma_j} p(\mathbf{x}|\omega_j) \mathrm{d}\mathbf{x} \Big) .$$

2. Diseño de un Sistema de Reconocimiento de Patrones

2.1. Problemas en el Diseño de un Sistema de Reconocimiento de Patrones

Si conociéramos totalmente las características estadísticas del modelo en el proceso de generación de los patrones, esto es si supiéramos $P(\omega_i),\ p(\mathbf{x}|\omega_i)\ \forall\, i$; entonces podríamos diseñar el sistema de reconocimiento de patrones (SRP) óptimo mediante aplicación directa de la teoría de decisión de Bayes. Sin embargo en la práctica surgen los siguientes problemas:

- el modelo no se puede conocer totalmente y/o
- la complejidad del SRP a diseñar está restringida por consideraciones económicas (hardware, tiempo)

Por lo general la base de conocimiento disponible para el diseño de un SRP es un conjunto de entrenamiento constituido por observaciones, ya sea etiquetadas o no.

En el primer caso asumimos que para cada patrón o vector de observaciones $\mathbf{x}_i\ \ i = 1, 2, \cdots, N$ en el conjunto de entrenamiento, un experto asigna una etiqueta con la clase correcta γ_i. El diseño de un sistema basado en un conjunto de datos clasificados de antemano se conoce como *aprendizaje supervisado*.

Si no se dispone de conocimiento experto sobre el conjunto de datos, o si el etiquetado de los patrones de entrenamiento es impracticable por razones prácticas; entonces el problema de diseño implica la necesidad de una primera etapa de análisis de los datos. Este proceso primario de análisis se conoce como una etapa de *aprendizaje no supervisado*.

Por lo tanto, en el caso general, dado un conjunto de entrenamientos el diseño del SRP implica:

1. Inferencia del modelo a partir de los datos (aprendizaje).
2. Desarrollo de reglas de decisión prácticas.
3. Simulación y evaluación del rendimiento del sistema.

2.2. Reglas de Decisión para Clases con Distribución Normal (Gaussiana)

Supongamos que las clases responden a distribuciones normales, o sea que las densidades de probabilidad condicionales de cada clase tienen la forma

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\mathbf{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^T\big) \quad \forall\, i = 1, \cdots, m\,.$$

siendo

$$\begin{cases} \boldsymbol{\mu}_i = E\left[\mathbf{x}|\omega_i\right] & \text{vector medio para la } i\text{-ésima clase} \\ \mathbf{\Sigma}_i = \mathrm{Cov}[\mathbf{x}|\omega_i] = E\left[(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T|\omega_i\right] & \text{matriz de covarianza para la } i\text{-ésima clase} \end{cases}$$

Si tomamos logaritmo a ambos lados de la regla de Bayes del mínimo error obtenemos:

$$\text{Asignar } \mathbf{x} \to \omega_j \iff$$
$$\log P(\omega_j) - \frac{1}{2}\left(n\log(2\pi) + \log|\mathbf{\Sigma}_j| + (\mathbf{x}-\boldsymbol{\mu}_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)^T\right) =$$
$$\max_{1\le k\le m}\left[\log P(\omega_k) - \frac{1}{2}\left(n\log(2\pi) + \log|\mathbf{\Sigma}_k| + (\mathbf{x}-\boldsymbol{\mu}_k)\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)^T\right)\right] \quad (1)$$

Multiplicando por 2, sacando para afuera de la expresión el signo de menos y agrupando los términos que no dependen de $\mathbf{x}$ resulta:

$$\text{Asignar } \mathbf{x} \to \omega_j \iff (\mathbf{x}-\boldsymbol{\mu}_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)^T - C_j = \min_{1\le k\le m}\left[(\mathbf{x}-\boldsymbol{\mu}_k)\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)^T - C_k\right]$$

siendo

$$C_k = 2\log P(\omega_k) - (n\log(2\pi) + \log|\mathbf{\Sigma}_k|)$$

Observar que en el caso particular en que las probabilidades a priori son uniformes y las matrices de correlación tienen determinante constante en las distintas clase, de modo que

$$P(\omega_k) = \frac{1}{m}\ \forall 1 \le k \le m \quad \text{y} \quad \log|\mathbf{\Sigma}_k| = \log|\mathbf{\Sigma}_{k'}|\ \forall 1 \le k \ne k' \le m$$

la regla de Bayes del mínimo error se simplifica a:

$$\text{Asignar } \mathbf{x} \to \omega_j \iff (\mathbf{x}-\boldsymbol{\mu}_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)^T = \min_{1\le k\le m}\left[(\mathbf{x}-\boldsymbol{\mu}_k)\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)^T\right]$$

Si definimos para cada clase una norma a partir de su matriz de covarianza como sigue

$$\forall \mathbf{x} \in \Omega \ \Rightarrow\ \|\mathbf{x}\|_k^2 = \mathbf{x}\mathbf{\Sigma}_k^{-1}\mathbf{x}^T \quad 1 \le k \le m$$

podemos escribir el criterio anterior como:

$$\text{Asignar } \mathbf{x} \to \omega_j \iff \|\mathbf{x}-\boldsymbol{\mu}_j\|_j = \min_{1\le k\le m}\|\mathbf{x}-\boldsymbol{\mu}_k\|_k$$

De modo que asignamos un patrón $\mathbf{x}$ arbitrario a la clase ω_j tal que su vector medio $\boldsymbol{\mu}_j$ este más cercano en la distancia inducida por la norma asociada a esa clase.

2.2.1. Caso Particular 1

Supongamos que las probabilidades a priori y las covarianzas son constantes en las clases

$$\mathbf{\Sigma}_i = \mathbf{\Sigma}_j = \mathbf{\Sigma} \quad \text{y} \quad P(\omega_i) = P(\omega_j) \quad \forall 1 \le i \ne j \le m$$

Definimos la norma y distancias asociadas a $\mathbf{\Sigma}$ como antes:

$$\|\mathbf{x}\|^2 = \mathbf{x}\mathbf{\Sigma}^{-1}\mathbf{x}^T \quad \Rightarrow \quad d(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\| \quad \forall\, \mathbf{x},\mathbf{x}' \in \Omega$$

Esta distancia se conoce como *distancia de Mahalanobis* asociada a la covarianza $\mathbf{\Sigma}$. Notar que en el caso particular que $\mathbf{\Sigma} = \mathbf{I}$ está distancia coincide con la distancia cuadrática Euclidiana.

Por lo tanto la regla del mínimo error se reduce a asignar el patrón $\mathbf{x}$ a la clase cuyo vector medio sea el más cercano según la distancia de Mahalanobis (*Regla de decisión por media más cercana*):

$$\text{Asignar } \mathbf{x} \to \omega_j \iff d(\mathbf{x},\boldsymbol{\mu}_j) = \min_{1\le k\le m} d(\mathbf{x}-\boldsymbol{\mu}_k)$$

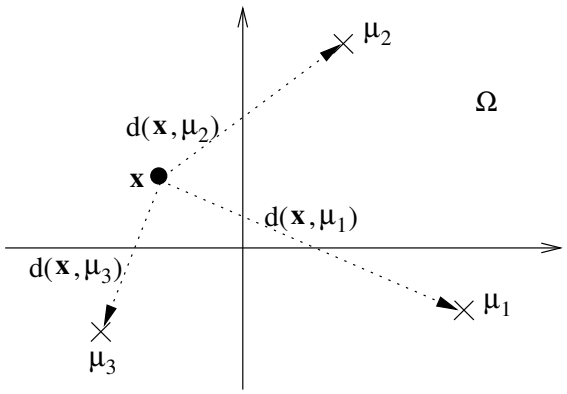

Figura 4: Regla de asignación por media más cercana.

2.2.2. Caso Particular 2

Ahora analizaremos el problema general de decisión entre dos clases

$$\Rightarrow \quad m = 2, \quad \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$$

Aplicando la regla de Bayes vemos que la ecuación

$$f(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{\mu}_1\|_1 - \|\mathbf{x} - \boldsymbol{\mu}_2\|_2 + C_2 - C_1 = 0$$

define la superficie de separación, conocida como *superficie discriminante*, entre las regiones asociadas a cada una de las clases ω_1 y ω_2. En general esta superficie es cuadrática ya que su ecuación resulta:

$$\underbrace{\mathbf{x}^T\big(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\big)\mathbf{x}}_{\text{cuadrático}} - \underbrace{2\mathbf{x}^T\big(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\big)}_{\text{lineal}} + \underbrace{\big(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + C_2 - C_1\big)}_{\text{constante}} = 0\,.$$

2.2.3. Caso Particular 3

En este caso asumimos que tenemos 2 clases con la misma covarianza

$$\Rightarrow \quad m = 2, \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$$

Observando la ecuación para el caso anterior, vemos que se anula el término cuadrático y por lo tanto la superficie discriminante resulta lineal como función del patrón vectorial $\mathbf{x}$. La superficie de separación es un *hiperplano* definido por:

$$\mathbf{x}^T\boldsymbol{\Sigma}\big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\big) + \text{cte} = \mathbf{x}^T\mathbf{w} + \text{cte} = 0 \qquad \text{siendo} \quad \mathbf{w} = \boldsymbol{\Sigma}\big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\big)\,.$$

El resultado es un clasificador lineal se puede implementar como se muestra en la figura 5.

Esta estructura es idéntica a la de una importante familia de máquinas lineales usadas en sistemas de decisión, entre las cuales podemos mencionar al *Perceptrón*.

2.2.4. Inferencia de los parámetros

La inferencia de los parámetros $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$ involucrados en las reglas de decisión es directa a partir de los conjuntos de entrenamiento

$$\{\mathbf{x}_{ij} \quad i = 1, \cdots, m;\ j = 1, \cdots, N_i\} \quad \text{con} \quad \mathbf{x}_{ij} \in \omega_i\ .$$

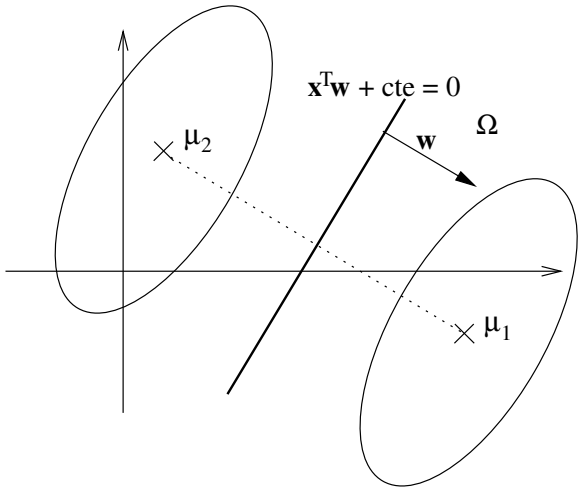

Figura 5: Regla de asignación por media más cercana.

El patrón vectorial medio para la clase ω_i se puede estimar como

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$$

en tanto la matriz de covarianza se estima con

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_i\right)\left(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_i\right)^T$$

2.3. Evaluación del Desempeño de un Sistema de Clasificación

Supongamos que disponemos de un conjunto de patrones vectoriales $\mathbf{x}_j \;\; j = 1, \cdots, N$ con sus correspondientes clases verdaderas γ_j conocidas. Es importante notar que este conjunto de patrones debería tener independencia estadística con el conjuntos de patrones de entrenamiento del sistema.

Sean β_j las etiquetas asignadas a cada $\mathbf{x}_j$ por el sistema de reconocimiento de patrones que estamos evaluando, e introduzcamos las variables aleatorias $\eta(\mathbf{x}_j)$ definidas según

$$\eta(\mathbf{x}_j) = \begin{cases} 0 & \text{si } \gamma_j = \beta_j \\ 1 & \text{si } \gamma_j \neq \beta_j \end{cases}$$

Notar que el valor esperado de $\eta(\mathbf{x})$ para un patrón $\mathbf{x} \in \Omega$ elegido al azar es

$$E\left[\eta\right] = \int_\Omega \left(1 \cdot \epsilon(\mathbf{x}) + 0 \cdot \left[1 - \epsilon(\mathbf{x})\right]\right) p(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_\Omega \epsilon(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} = e$$

en tanto la varianza resulta

$$E\left[(\eta - e)^2\right] = \int_\Omega \epsilon(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} - e^2 = e(1 - e) \; .$$

Podemos también deducir esto observando η es una variable aleatoria de Bernoulli tal que la probabilidad $p\{\eta = 1\}$ es igual a la probabilidad media e de error del sistema.

Si hacemos N observaciones independientes de η y definimos la nueva variable aleatoria $\hat{e}$ como

$$\hat{e} = \frac{1}{N} \sum_{j=1}^{N} \eta_j$$

encontramos que $\hat{e}$ tiene distribución Binomial con valor medio e (como se ve calculando la esperanza) y por lo tanto $\hat{e}$ es un estimador insesgado del error medio e.

La desviación estándar de este estimador se calcula como

$$\sigma_{\hat{e}} = \sqrt{\mathrm{Var}\left[\hat{e}\right]} = \left(E\left[\hat{e}^2\right] - e^2\right)^{\frac{1}{2}}$$

Sustituyendo $\hat{e}$ y desarrollando resulta

$$\mathrm{Var}\left[\hat{e}\right] = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} E\left[\eta_j \eta_k\right] - e^2 = \frac{1}{N} E\left[\eta^2\right] + \frac{N-1}{N} e^2 - e^2 = \frac{1}{N} e(1-e)$$

de donde la desviación del estimador será

$$\sigma_{\hat{e}} = \sqrt{\frac{e(1-e)}{N}} \leq \frac{1}{2\sqrt{N}} \,.$$

Resumiendo, hemos visto que podemos estimar el error medio de clasificación e presentándole a nuestro sistema de clasificación un conjunto de patrones que pertenecen a clases conocidas. El error se estima contando el número de discrepancias entre la clase verdadera y la etiqueta de clase asignada por el sistema, y dividiendo finalmente este resultado entre el número de muestras en la prueba.

Notar que si el error medio del sistema es pequeño, digamos de $\approx 1\,\%$, vamos a necesitar de un número grande de muestras de prueba para verificar este valor de desempeño con una razonable confianza relativa.

3. Apendizaje no supervisado

Es común encontrarse con situaciones en las que el sistema de clasificación de patrones debe diseñarse partiendo de un conjunto de patrones de entrenamiento $\{x_j\,; \quad j = 1, 2, \cdots, N\}$ para los cuales no conocemos sus etiquetas de clase γ_i.

Estas situaciones se presentan cuando no disponemos del conocimiento de un experto o bien cuando el etiquetado de cada muestra individual es impracticable. Esto último ocurre por ejemplo en el caso de aplicaciones con sensores remotos, como ser imágenes satelitales de terrenos donde sería muy costoso o imposible recoger información real del tipo de suelo sensado en cada punto de las imágenes. En estos casos el proceso de diseño requiere una primera etapa de análisis de las estructuras presentes en los datos de entrenamiento.

3.1. Aprendizaje no supervisado y análisis de agrupamientos

Dado un conjunto de entrenamiento suficientemente grande podemos inferir la función densidad de probabilidad conjunta $p(\mathbf{x})$ y recordando que

$$p(\mathbf{x}) = \sum_{i=1}^{m} P(\omega_i) p(\mathbf{x}|\omega_i)$$

podemos deducir que si la densidad conjunta es multimodal cada uno de los modos debería corresponderse con la distribución condicional de cada una de las clases presentes. Por lo tanto identificando estos modos en $p(\mathbf{x})$ sería en principio posible particionar el espacio de observación en regiones disjuntas $\Gamma_i\,, i = 1, \cdots, m$ asociadas con cada una de las clases presentes.

Si las distribuciones condicionales de cada clase son normales cabria la posibilidad de recuperar los parámetros de cada distribución a partir del conjunto de entrenamiento. A partir de esto podríamos seguir con el diseño del clasificador como se vio en la sección anterior. Sin embargo podemos conformarnos con recobrar

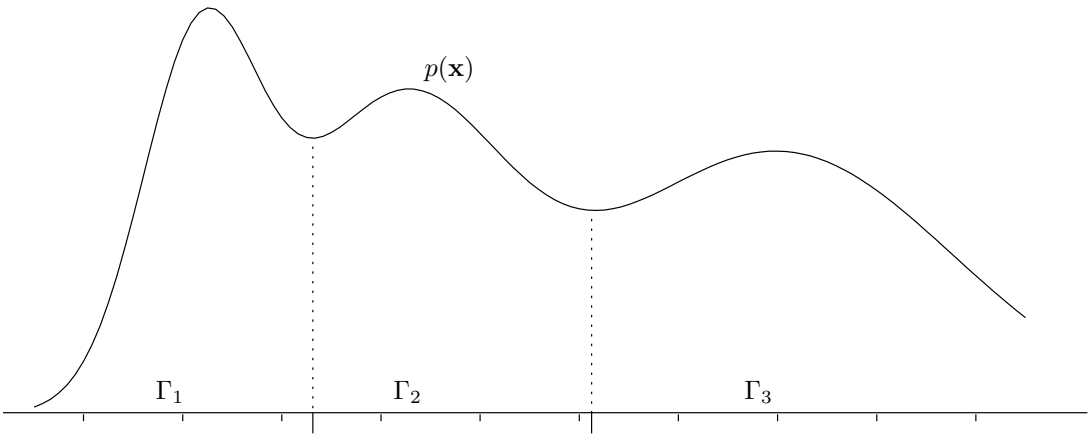

Figura 6: Distribución conjunta multimodal y regiones asociadas a cada clase..

directamente las regiones Γ_i lo cual es suficiente para nuestros intereses ya que esto puede usarse directamente para la clasificación de nuevos datos simplemente usando el criterio:

$$\text{Asignar } \mathbf{x} \text{ a } \omega_j \iff \mathbf{x} \in \Gamma_j$$

Un opción alternativa seria usar este u otro criterio para clasificar los patrones en el conjunto de entrenamiento y luego usar estas etiquetas para diseñar el sistema de reconocimiento de patrones usando un aprendizaje supervisado. En la práctica ocurre que determinar explicitamente las regiones Γ_i implicaría estimar la función de densidad conjunta y luego analizarla en un espacio de dimensión n lo que generalmente es impracticable por su complejidad computacional. Además como vimos, solo necesitamos de un método indirecto que nos permita etiquetar automáticamente los patrones de entrenamiento. Entonces lo que queremos es alguna forma de hacer una partición del conjunto de entrenamiento en clases con una misma etiqueta y esto es lo que se conoce como métodos de agrupamiento o *clustering*.

Intuitivamente podemos anticipar que las modas en la función de densidad conjunta $p(\mathbf{x})$ estarán asociadas a regiones con alta densidad de muestras en el espacio de observación. El proposito de las técnicas de agrupamiento será justamente detectar y agrupar estos *enjambres* de puntos.

3.2. Medidas de Similitud y Criterios de Agrupamiento

El propósito de los métodos de agrupamiento será analizar y extraer la estructura presente en un conjunto de patrones o muestras de entrenamiento. Diremos que un conjunto de datos está bien estructurado si contiene varios enjambres de patrones cercanos entre si, o sea regiones de alta densidad, separados por otras regiones relativamente vacias o con poca densidad.

Vemos que los puntos de un mismo agrupamiento apareceran más proximos entre ellos que a puntos en otros agrupamientos. Esta observación nos lleva a concluir que si queremos decidir si un punto $\mathbf{x}$ pertenece o no a un agrupamiento necesitaremos una medida de proximidad o similitud. Se han sugerido y estudiado un gran número de tales medidas, pero probablemente las más comunmente usadas son las medidas de distancia y en particular la distancia Euclideana.

La afinidad de un punto a un agrupamiento se puede determinar ya sea midiendo su similitud con otros puntos en el agrupamiento o bien con un modelo definido para el agrupamiento. El ejemplo más sencillo de esto último es representar un agrupamiento i por su vector medio μ_i; en este caso la afinidad entre un puntox y el agrupamiento se puede cuantificar con la distancia Euclideana al cuadrado

$$d(\mathbf{x}, \mu_i) = \left[(\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i)\right]$$

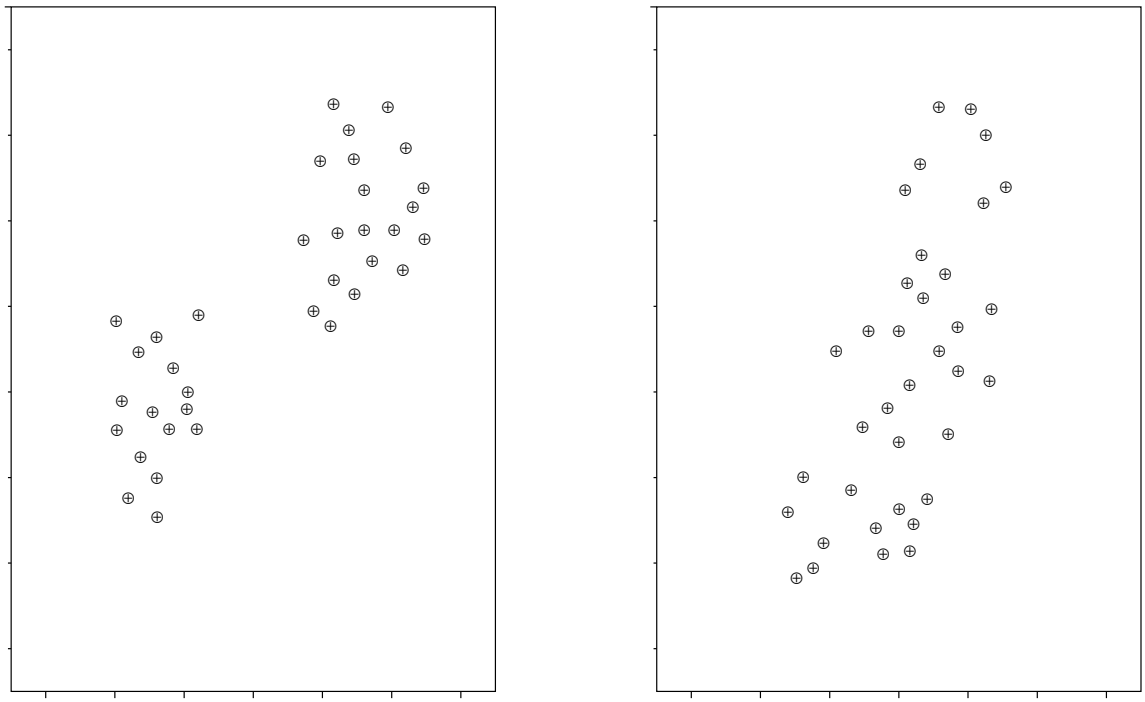

Figura 7: Datos structurados vs. no estructurados.

Pero para particionar un conjunto de puntos en agrupamientos de una manera óptima no nos alcanza con una medida de afinidad o similitud sino que además necesitamos algun *criterio de agrupamiento* que nos permita definir cuantitativamente cuando una partición es mejor que otra. Obviamente tanto el criterio de agrupamiento que definamos tanto como el algoritmo de agrupamiento asociado, estarán intimamente relacionados con la medida de similitud usada y se definirán a partir de esta.

En la siguiente sección veremos algunos ejemplos de métodos de agrupamiento que se basan en los conceptos anteriores.

3.3. Algoritmo de k-medias (k-means).

Supondremos que el conjunto de datos $\mathbf{X}$ contiene k agrupamientos y que cada uno de estos subconjuntos $\mathbf{X}_i$ puede representarse adecuadamente con su valor medio μ_i. Como se menciona anteriormente, en este caso podemos usar la distancia Euclideana como una medida de similitud. Se deduce que un criterio de agrupamiento adecuado en este caso es considerar la suma total sobre el conjunto de entrenamiento de la distancia cuadrática de cada punto al vector valor medio de su agrupamiento.

El objetivo del algoritmo de agrupamiento será encontrar entre todas las particiones de $\mathbf{X}$ en k conjuntos $\{\mathbf{X}_i\ ;\quad i = 1, 2, \cdots, k\}$ aquella que minimice el criterio de agrupamiento elegido.

Dicho formalmente, queremos encontrar los agrupamientos $\{\mathbf{X}_i\}$ que minimizan la función

$$J = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k} \sum_{j=1}^{N_i} d(\mathbf{x}_{ij}, \mu_i) \quad \text{siendo} \quad \mathbf{x}_{ij} \in \mathbf{X}_i,\ N_i = \#\mathbf{X}_i$$

entre todas las posibles particiones de $\mathbf{X}$ en k subconjuntos.

Un algoritmo para minimizar J puede deducirse considerando el efecto de un cambio minimal o atómico en la configuración de agrupamientos, que consiste en sacar un punto $\mathbf{x}$ que este en el agrupamiento $\mathbf{X}_l$ para pasarlo a otro agrupamiento $\mathbf{X}_r$.

Claramente esta reasignación afectara solo a los agrupamientos l y r cuyos valores medios pasarán a ser

$$\bar{\mu}_l = \mu_l + \frac{1}{N_l - 1}(\mu_l - \mathbf{x}) \quad \text{y} \quad \bar{\mu}_r = \mu_r - \frac{1}{N_r + 1}(\mu_r - \mathbf{x})$$

respectivamente.

Para deducir la primera ecuación calculamos el valor medio de $\mathbf{X}_i$ antes y despues de la reasignación

$$\mu_l = \frac{1}{N_l}\sum_{j=1}^{N_l} \mathbf{x}_j \qquad \bar{\mu}_l = \frac{1}{N_l - 1}\sum_{j=1}^{N_l - 1} \mathbf{x}_j = \frac{1}{N_l - 1}\left(\sum_{j=1}^{N_l} \mathbf{x}_j - \mathbf{x}\right)$$

donde hemos asumido que el punto reasignado es el último en la sumatoria. De aqui resulta que

$$(N_l - 1)\bar{\mu}_l = N_l\,\mu_l - \mathbf{x} \quad \Rightarrow \quad \bar{\mu}_l = \frac{N_l}{N_l - 1}\,\mu_l - \frac{1}{N_l - 1}\,\mathbf{x} \quad \Rightarrow \quad \bar{\mu}_l = \mu_l + \frac{1}{N_l - 1}(\mu_l - \mathbf{x})$$

y análogamente se verifica la segunda identidad.

Por lo tanto para calcular el cambio global en el valor de J bastará calcular los cambios en las contribuciones de J_l y J_r. Para el nuevo agrupamiento l-esimo tendremos

$$\begin{aligned}
\bar{J}_l &= \sum_{j=1}^{N_l - 1} d(\mathbf{x_j}, \mathbf{mu}_l) = \sum_{j=1}^{N_l - 1} (\mathbf{x}_j - \mu_l)^T(\mathbf{x}_j - \mu_l) = \\
&= \sum_{j=1}^{N_l} \big(\mathbf{x}_j - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big)^T \big(\mathbf{x}_j - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big) - \big(\mathbf{x} - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big)^T \big(\mathbf{x} - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big) = \\
&J_l - \frac{2}{N_l - 1}(\mu_l - \mathbf{x}) \underbrace{\sum_{j=1}^{N_l}(\mathbf{x}_j - \mu_l)}_{0} + \frac{N_l}{(N_l - 1)^2}(\mu_l - \mathbf{x})^T(\mu_l - \mathbf{x}) + \frac{N_l^2}{(N_l - 1)^2}(\mu_l - \mathbf{x})^T(\mu_l - \mathbf{x})
\end{aligned}$$

de donde luego de agrupar concluimos que

$$\bar{J}_l = J_l - \frac{N_l}{N_l - 1}(\mu_l - \mathbf{x})^T(\mu_l - \mathbf{x}) = J_l - \frac{N_l}{N_l - 1}\,d(\mathbf{x}, \mu_l)$$

y análogamente para el agrupamiento r se obtiene

$$\bar{J}_r = J_r + \frac{N_r}{N_r - 1}(\mu_r - \mathbf{x})^T(\mu_r - \mathbf{x}) = J_r + \frac{N_r}{N_r - 1}\,d(\mathbf{x}, \mu_r)$$