CAPÍTULO 10

CLASIFICACIÓN II: RECONOCIMIENTO ESTADÍSTICO

10.1 INTRODUCCIÓN

En el capítulo anterior se introdujeron una serie de técnicas bajo la perspectiva de las redes neuronales incluyendo algunas relativas a agrupamiento. En este capítulo se ahonda más en el tema de los agrupamientos, si bien desde una perspectiva estadística. Las técnicas estudiadas son las que se exponen a continuación, cuya ampliación se puede encontrar en el capítulo 14 del libro de referencia, así como en Duda y col. (2001) o Balasto y col. (2006):

- *agrupamiento borroso* (*fuzzy clustering, fuzzy k-Means*)
- *clasificación paramétrica*: *clasificador Bayesiano*
- *clasificación no paramétrica*: *ventana de Parzen*

10.2 AGRUPAMIENTO BORROSO

El objetivo de la técnica de agrupamiento conocida como *Agrupamiento borroso o Fuzzy Clustering* consiste en dividir n objetos $x \in X$ caracterizados por p propiedades en c "clusters" o grupos. Supongamos el conjunto de datos $X = \{x_1, x_2, ..., x_n\} \in \Re^p$ un subconjunto del espacio real p-dimensional $\Re^p$. Cada $x_k = \{x_{k_1}, x_{k_2}, ..., x_{k_p}\} \in \Re^p$ se denomina vector de características, x_{k_j} es la j-ésima característica de la observación x_k.

Por ejemplo, imaginemos que la intención es clasificar caracteres alfanuméricos utilizando como características los siete momentos invariantes de Hu. En este caso, cada carácter vendría identificado por los siete momentos, siendo $p = 7$.

Puesto que los elementos de un clúster deben ser tan similares entre sí como sea posible y a la vez deben ser tan diferentes a los elementos de otros clústeres como también sea posible, el proceso se controla por el uso de medidas de similitud basadas en distancias. Así la similitud o la diferencia entre dos puntos x_k y x_l puede interpretarse como la distancia entre esos puntos.

Una distancia entre dos objetos x_k y x_l es una función que toma valores reales $d : XxX \rightarrow R^+$ cumpliendo:

$$d(x_k, x_l) = d_{kl} \geq 0 ; \; d_{kl} = 0 \Leftrightarrow x_k = x_l ; \; d_{kl} = d_{lk} \text{ y } \; d_{kl} \leq d_{kj} + d_{jl} \quad (10.1)$$

Cada partición del conjunto $X = \{x_1, x_2, ..., x_n\}$ puede enfocarse desde dos perspectivas: fuzzy y no fuzzy. Una partición no fuzzy se conoce en terminología inglesa como "crisp". Si se desea realizar una partición del conjunto X en c clústeres tendremos $S_i \{i = 1, ..., c\}$ subconjuntos. A partir de esta consideración se define lo que se conoce como grado de pertenencia μ_{ik} de cada objeto x_k al subconjunto S_i. En el caso de conjuntos "crisp" un objeto x_k se dice que pertenece a un S_i dado y no pertenece al resto. Esto se expresa con los valores discretos $\{0,1\}$ de la siguiente forma: $\mu_{ik} = 1$ para indicar pertenece y $\mu_{ik} = 0$ para expresar que no pertenece. Por el contrario, en el caso de conjuntos fuzzy se dice que un objeto puede pertenecer a diferentes subconjuntos y así se habla por ejemplo de que x_k pertenece a un conjunto S_i con grado de pertenencia μ_{ik} y a S_j con grado de pertenencia μ_{jk}. Como ejemplo, supongamos que se tienen tres conjuntos S_i, S_j y S_h, en este caso podríamos decir que el objeto x_k pertenece a los conjuntos con los siguientes grados de pertenencia $\mu_{ik} = 0.4$, $\mu_{jk} = 0.5$ y $\mu_{hk} = 0.1$. Los valores tomados corresponden al intervalo continuo [0,1].

Dado $X = \{x_1, x_2, ..., x_n\}$ y el conjunto V_{cn} de todas las matrices reales de dimensión c x n, con $2 \leq c < n$. Se puede obtener una matriz representando la partición de la siguiente manera $U = \{\mu_{ik}\} \in V_{cn}$. Tanto en el supuesto "crisp" como en el fuzzy se deben cumplir las siguientes condiciones:

1) $\mu_{ik} \in \{0,1\}$ *crisp* o $\mu_{ik} \in [0,1]$ *fuzzy* $\quad 1 \leq i \leq c; \; 1 \leq k \leq n$

$$2)\ \sum_{i=1}^{c}\mu_{ik}=1 \qquad 1\le k\le n \qquad (10.2)$$

$$3)\ 0<\sum_{k=1}^{n}\mu_{ik}<n \qquad 1\le i\le c$$

Para ilustrar los conceptos anteriores sea $X=\{x_1,x_2,x_3\}$ con él podríamos construir las siguientes particiones teniendo en cuenta que $c=2$.

"crisp"

$$U=\begin{bmatrix} x_1 & x_2 & x_3\\ 1 & 0 & 0\\ 0 & 1 & 1\end{bmatrix}$$

$$U=\begin{bmatrix} x_1 & x_2 & x_3\\ 0 & 1 & 0\\ 1 & 0 & 1\end{bmatrix}$$

"fuzzy"

$$U=\begin{bmatrix} x_1 & x_2 & x_3\\ 0.3 & 0.5 & 0\\ 0.7 & 0.5 & 1\end{bmatrix}$$

$$U=\begin{bmatrix} x_1 & x_2 & x_3\\ 0.9 & 0.4 & 0.2\\ 0.1 & 0.6 & 0.8\end{bmatrix}$$

La localización de un clúster S_i se representa por su centro $v_i=\{v_{i_1},v_{i_2},\ldots,v_{i_p}\}\in\Re^p$ con $i=1,\ldots,c$, alrededor del cual se concentran los objetos.

La definición básica de llevar a cabo el problema de la partición fuzzy para $m>1$ consiste en minimizar la siguiente función objetivo según la ecuación (10.3):

$$\min z_m(U;v)=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\|x_k-v_i\|_G^2 \qquad (10.3)$$

donde G es una matriz de dimensión $p\ x\ p$ que es simétrica y definida positiva. Así se puede definir una norma general del tipo,

$$\|x_k-v_i\|_G^2=(x_k-v_i)'G(x_k-v_i) \qquad (10.4)$$

Diferenciando la función objetivo para v_i (suponiendo constante U) y μ_{ik} (suponiendo constante v) y aplicando la condición de que $\sum_{i=1}\mu_{ik}=1$, se obtiene,

$$v_i = \frac{1}{\sum_{k=1}^{n} (\mu_{ik})^m} \sum_{k=1}^{n} (\mu_{ik})^m x_k \qquad i = 1,\ldots,c \quad (10.5)$$

$$\mu_{ik} = \frac{\left(\frac{1}{\|x_k - v_i\|_G^2} \right)^{2/m-1}}{\sum_{j=1}^{c} \left(\frac{1}{\|x_k - v_j\|_G^2} \right)^{2/m-1}} \qquad i = 1,\ldots,c; k = 1,\ldots,n \qquad (10.6)$$

El exponente m se conoce como peso exponencial y disminuye la influencia del ruido al obtener los centros de los clústeres, reduciendo la influencia de los valores pequeños de μ_{ik} (puntos lejos de v_i) frente a valores altos de μ_{ik} (puntos cercanos a v_i). Cuanto mayor sea $m > 1$ mayor es dicha influencia.

Uno de los más conocidos algoritmos para el agrupamiento "crisp" es el algoritmo c-medias o ISODATA.

De modo similar el algoritmo conocido como *c-medias fuzzy* para obtener los valores de las expresiones (10.5) y (10.6) es el siguiente:

1) Elegir c ($2 \le c \le n$), m ($1 < m < \infty$) y la matriz G de dimensión pxp siendo simétrica y definida positiva. Inicializar $U^{(0)}$ y poner $t = 0$.

2) Calcular los c centros fuzzy de los clústeres a partir de (10.5) $\{v_i^{(t)}\}$ utilizando $U^{(t)}$.

3) Calcular los nuevos grados de pertenencia de la matriz $U^{(t+1)}$ utilizando $\{v_i^{(t)}\}$ a partir de la condición (10.6) si $x_k \neq v_i^{(t)}$. De lo contrario

$$\mu_{jk} = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases} \qquad (10.7)$$

4) Elegir una norma matricial y calcular $\Delta = \|U^{(t+1)} - U^t\|_G$. Si $\Delta > \varepsilon$ poner $t = t + 1$ y regresar al paso 2), de lo contrario detener el proceso.

Por tanto, para llevar a cabo un proceso de agrupamiento es necesario definir: a) número de clústeres c; b) el peso exponencial m (siendo un valor típico de 2); c) la matriz G que induce la norma; d) el método para inicializar $U^{(0)}$ y e) el criterio de terminación $\Delta \leq \varepsilon$.

Especial mención requiere el estudio de la matriz G que determina la forma del clúster. Si se elige la norma Euclídea, entonces G es la matriz identidad I y la forma de los clústeres se asume que constituyen una hiperesfera. G también se puede elegir como una matriz diagonal con $G_D = \left[diag\left(\sigma_j^2\right)\right]^{-1}$ o la norma de Mahalanobis con $G_M = \left[\text{cov}(x)\right]^{-1}$, donde σ_j^2 denota la varianza de la característica j y cov la covarianza.

Existen diferentes medidas escalares para validar la partición, siendo el objetivo encontrar el número de clústeres que obtenga las mejores medidas:

a) *Coeficiente de partición* (*CP*): mide la cantidad de solapamiento entre los clústeres

$$CP = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\left(\mu_{ij}\right)^2 \qquad (10.8)$$

Cuanto más próximo sea este valor a la unidad tanto mejor será la clasificación de los datos llegando a ser una partición pura cuando toma el valor de la unidad.

b) *Coeficiente de entropía* (*CE*):

$$CE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\log\left(\mu_{ij}\right) \qquad (10.9)$$

En este caso, la mejor partición viene dada por el menor valor de este coeficiente.

c) *Índice de partición* (*SC*): es la razón entre la suma de compacidad y la separación de los clústeres. Se trata de una suma de las medidas de validez de cada clúster individual normalizada por la división de la cardinalidad fuzzy de cada clúster (N_i)

$$SC = \sum_{i=1}^{c}\frac{\sum_{j=1}^{n}\left(\mu_{ij}\right)^2\left\|x_j - v_i\right\|^2}{N_i\sum_{k=1}^{c}\left\|v_k - v_i\right\|^2} \quad \text{con } N_i = \sum_{j=1}^{n}\mu_{ij} \qquad (10.10)$$

Este índice es útil cuando se comparan diferentes particiones con el mismo número de clústeres. Valores bajos de *SC* indican mejores particiones.

d) Índice de separación (*S*): utiliza una distancia mínima para la validez de la separación

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^2 \left\|x_j - v_i\right\|^2}{n d_{\min}^2} \quad \text{con} \quad d_{\min} = \min_{i,j}\left\{\left\|v_i - v_j\right\|\right\} \qquad (10.11)$$

Cuanto más bajo sea el valor de *S* más compacta y separada es la partición.

e) Índice Xie-Beni (*XB*): utiliza una distancia mínima para la validez de la separación

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m \left\|x_j - v_i\right\|^2}{n d_{\min}^2} \quad \text{con} \quad d_{\min} = \min_{i,j}\left\{\left\|v_i - x_j\right\|\right\} \qquad (10.12)$$

Cuanto menor sea el valor de *XB* tanto mejor es la partición y por tanto los dos se encuentran mejor agrupados.

f) Índice de Dunn (*ID*): también utilizado para identificar conjuntos de clústeres que son compactos y bien separados,

$$ID = \min_{i\in c}\left\{\min_{j\in c, i\neq j}\left\{\frac{\min_{x\in C_i, y\in C_j}\left\|x-y\right\|}{\max_{k\in c}\left\{\min_{x,y\in C}\left\|x-y\right\|\right\}}\right\}\right\} \qquad (10.13)$$

Obsérvese que el valor del numerador se obtiene como resultado de computar las distancias entre x e y, que son patrones pertenecientes a clústeres distintos C_i y C_j (se trata de una distancia entre clústeres), mientras que en el denominador la distancia es entre patrones $\boldsymbol{x}$ e $\boldsymbol{y}$ que pertenecen al mismo clúster C. Este índice resulta ser computacionalmente hablando muy costoso.

g) Índice de Dunn Alternativo (*IDA*): propuesto para reducir el elevado coste computacional del anterior, para lo cual se hace uso de la desigualdad triangular $\left\|\left\|y - v_j\right\| - \left\|x - v_j\right\|\right\| \leq \left\|x - y\right\|$.

$$IDA = \min_{i\in c}\left\{\min_{j\in c, i\neq j}\left\{\frac{\min_{x\in C_i, y\in C_j}\left|\left\|x - v_j\right\| - \left\|y - v_j\right\|\right|}{\max_{k\in c}\left\{\min_{x,y\in C}\left\|x-y\right\|\right\}}\right\}\right\} \qquad (10.14)$$

10.3 CLASIFICADOR PARAMÉTRICO: BAYESIANO

10.3.1 Caso Normal Multivariable: media desconocida

Supongamos que las muestras siguen una distribución Normal con media **m** y matriz de covarianza C. Por simplicidad en este caso concreto el único parámetro desconocido es la media,

$$p(\mathbf{x}_i/\mathbf{m}) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mathbf{m})^t C^{-1}(\mathbf{x}_i - \mathbf{m})\right\} \quad (10.15)$$

$$\ln p(\mathbf{x}_i / \mathbf{m}) = -\frac{1}{2}\ln\left\{(2\pi)^d |C|\right\} - \frac{1}{2}(\mathbf{x}_i - \mathbf{m})^t C^{-1}(\mathbf{x}_i - \mathbf{m}) \quad (10.16)$$

la expresión anterior se obtiene identificando w con **m**

$$\nabla_{\mathbf{m}} \ln p(\mathbf{x}_i / \mathbf{m}) = C^{-1}(\mathbf{x}_i - \mathbf{m}) \quad (10.17)$$

La minimización del riesgo empírico supone que

$$\frac{1}{n}\sum_{i=1}^{n} C^{-1}(\mathbf{x}_i - \mathbf{m}) = 0 \quad (10.18)$$

multiplicando por C y despejando **m** se obtiene una estima para la misma, que resulta

$$\mathbf{m}^* = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \quad (10.19)$$

Éste es un resultado bastante satisfactorio, dice que la estima de máxima verosimilitud para la media desconocida de una distribución es exactamente la media aritmética de las muestras, es decir, la *media simple*.

Geométricamente, si consideramos las n muestras como una nube de puntos, la media simple es el centroide de la nube, que puede ser considerado el representante de dicha clase.

La media anteriormente obtenida tiene las propiedades estadísticas conocidas y siempre se tiene tendencia a utilizarla aunque no se tenga el conocimiento de que la misma es la solución de máxima verosimilitud.

10.3.2 Caso Normal Multivariable: media y matriz de covarianza desconocidas

En el caso general y más típico de una Normal multivariable, ni la media **m** ni la matriz de covarianza C son conocidas. Por tanto, esos parámetros desconocidos constituyen las componentes del vector de parámetros $w = \{w_1, w_2\}$. Consideremos el caso univariable con $w_1 = m$ y $w_2 = \sigma^2$, en cuyo caso

$$\ln p(x_i/w) = -\frac{1}{2}\ln 2\pi w_2 - \frac{1}{2w_2}(x_i - w_1)^2 \quad (10.20)$$

$$\nabla_w \ln p(x_i / w) = \begin{bmatrix} \dfrac{1}{w_2}(x_i - w_1) \\ -\dfrac{1}{2w_2} + \dfrac{(x_i - w_1)^2}{2w_2^2} \end{bmatrix} \quad (10.21)$$

La minimización de los datos de entrenamiento conduce ahora a las condiciones,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{w}_2}(x_i - \hat{w}_1) = 0 \qquad -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\hat{w}_2} + \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \hat{w}_1)^2}{2\hat{w}_2^2} = 0 \quad (10.22)$$

donde $\hat{w}_1$ y $\hat{w}_2$ son las estimas de máxima verosimilitud para w_1 y w_2, respectivamente. Sustituyendo $\hat{m} = \hat{w}_1$ y $\hat{\sigma}^2 = \hat{w}_2$ obtenemos las estimas de máxima verosimilitud para m y σ^2

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{w}_1)^2 \quad (10.23)$$

Aunque el análisis del caso multivariable es básicamente muy similar, se requiere mucha más manipulación. El resultado muy bien conocido en estadística es que las estimas de máxima verosimilitud para **m** y C están dadas por,

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \qquad C = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t \quad (10.24)$$

La expresión (10.24) nos dice que la estima de máxima verosimilitud para el vector media es la media simple. La estima de máxima verosimilitud para la matriz de covarianza es la media aritmética de las n matrices $(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t$. Puesto que la verdadera matriz de covarianza es el valor esperado de la matriz $(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t$, se obtiene un resultado muy satisfactorio.

10.3.3 Teoría de la decisión de Bayes: el clasificador Bayesiano

La teoría de la decisión de Bayes es un método estadístico clásico en clasificación de patrones. Se basa en el supuesto de que el problema de la decisión se enfoca en términos probabilísticos y que todas las probabilidades relevantes resultan conocidas.

Antes de proceder a la generalización del desarrollo de este método, vamos a ver un ejemplo. Consideremos el problema de diseñar un clasificador para separar dos clases de frutas en una cinta transportadora, melocotones M o albaricoques A. Alguien que observe la cinta tiene muy difícil predecir qué tipo de fruta aparecerá en la cinta, por lo que la secuencia de tipos de frutas se presenta aleatoria. Utilizando la terminología de la teoría de la decisión, diremos que según aparece cada tipo de fruta, pertenecerá a uno de los dos estados: o es un melocotón o es un albaricoque. Sea y el estado que designa la pertenencia a uno de los dos tipos de fruta, con $y = M$ e $y = A$ para melocotón y albaricoque respectivamente. Puesto que la pertenencia a uno de los dos estados es impredecible, consideramos a y una variable aleatoria.

Si la cosecha de melocotones ha sido más o menos igual que la de albaricoques, diremos que la siguiente fruta que va a aparecer tiene la misma probabilidad de ser melocotón o albaricoque. De forma más precisa, suponemos que existe alguna *probabilidad a priori* $P(y=M)$ de que la siguiente fruta sea un melocotón y alguna *probabilidad a priori* de que sea un albaricoque $P(y=A)$. Esas probabilidades a priori reflejan nuestro conocimiento a priori sobre cuál es la probabilidad de ver aparecer un melocotón o un albaricoque antes de que aparezca la próxima pieza de fruta. Ambas probabilidades $P(y=M)$ y $P(y=A)$ son no negativas y su suma es la unidad.

Supongamos por un instante que estamos forzados a tomar una decisión sobre el tipo de fruta que aparecerá a continuación. La única información son esas probabilidades a priori, por tanto si tenemos que tomar una decisión, parece razonable utilizar la siguiente *regla de decisión*, decidir M si $P(y=M) > P(y=A)$, en caso contrario decidir A.

No obstante, la mayoría de las veces no se toma una decisión con esa poca información. En el ejemplo podríamos usar la componente de color rojo x como una información más. Así, diferentes muestras de frutas darán diferentes valores de x, resulta natural expresar esta variabilidad en términos probabilísticos. En este sentido, consideremos a x una variable aleatoria continua cuya distribución depende del tipo de fruta. Sean $p(x / y = A)$ y $p(x / y = M)$ las funciones de *densidad de probabilidad condicionales* para x dado que el tipo de fruta sea A o M.

Supongamos que conocemos tanto las probabilidades a priori como estas últimas funciones de densidad de probabilidad. Suponer además que medimos la componente de color rojo de un albaricoque y descubrimos que vale x, la pregunta será ¿cómo influye

esta medida sobre nuestra actitud con relación al tipo de fruta de que se trata? La respuesta nos la proporciona la *regla de Bayes*:

$$p(y=A/x)=\frac{p(x/y=A)P(y=A)}{p(x)} \quad p(y=M/x)=\frac{p(x/y=M)P(y=M)}{p(x)} \qquad (10.25)$$

con $$p(x)=p(x/y=A)P(y=A)+p(x/y=M)P(y=M) \qquad (10.26)$$

La regla de Bayes muestra cómo la observación del valor x cambia las probabilidades a priori a las *probabilidades a posteriori* $p(y=A/x)$ y $p(y=M/x)$.

La ecuación (10.26) es una constante de normalización, que no necesita ser obtenida a la hora de tomar una decisión, puesto que la regla de decisión es una comparación de las magnitudes relativas de las probabilidades a posteriori. Una vez que se determinan esas probabilidades a posteriori, la siguiente regla (función discriminante) se utiliza para clasificar x.

$$fd(x)=\begin{cases} A & \text{si} \quad p(x/y=A)P(y=A)>p(x/y=M)P(y=M) \\ M & \text{de otro modo} \end{cases} \qquad (10.27)$$

Si ahora ampliamos el número de tipos de frutas hasta J y el número de atributos dados por el vector de atributos $\mathbf{x} = \{x_1, x_2,...,x_d\}$ estamos ante un problema de clasificación general. Utilizando el teorema de Bayes y considerando que tanto las probabilidades a priori $P(y=c_j)$ como las densidades condicionales para cada clase $p(\mathbf{x}/y=c_j)$ son conocidas o se pueden estimar, es posible determinar para una observación dada $\mathbf{x}$ la probabilidad de que esa observación pertenezca a una determinada clase. Estas probabilidades, llamadas probabilidades a posteriori, pueden usarse para construir una regla discriminante

$$p(y=c_j/\mathbf{x})=\frac{p(\mathbf{x}/y=c_j)P(y=c_j)}{p(\mathbf{x})} \qquad (10.28)$$

donde

$$p(\mathbf{x})=\sum_{j=1}^{c} p(\mathbf{x}/y=c_j)P(y=c_j) \qquad (10.29)$$

A partir de (10.28) dado $\mathbf{x}$ la regla de decisión viene establecida por,

$$\mathbf{x}\in c_i \;\; sii \;\; p(y=c_i/\mathbf{x})>p(y=c_j/\mathbf{x}) \;\; \forall i\neq j, \;\; i,j=1,2,..c \qquad (10.30)$$

Fijándose en el segundo término de la expresión (10.21) del teorema de Bayes y eliminado el término no discriminante $p(\mathbf{x})$ (no aporta nada en la decisión), se tiene una forma alternativa de clasificar el vector de atributos $\mathbf{x}$:

$$\mathbf{x} \in c_i \ sii \ p(\mathbf{x}/y=c_i)P(y=c_i) > p(\mathbf{x}/y=c_j)P(y=c_j) \ \forall i \neq j, \ i,j=1,2,..,c \quad (10.31)$$

Generalmente las distribuciones de densidad de probabilidad se eligen Normales o Gaussianas.

Un caso especial surge cuando las probabilidades a priori son iguales para todas las clases, ya que en esta situación la distancia de Mahalanobis se puede utilizar como función discriminante mediante la siguiente regla de decisión a partir de (10.15) y teniendo en cuenta el signo negativo en el término exponencial de la función de densidad de probabilidad Normal, así

$$\mathbf{x} \in c_i \ sii \ d_M^2(\mathbf{x},\mathbf{m}_i) < d_M^2(\mathbf{x},\mathbf{m}_j) \ \forall i \neq j, \ i,j=1,2,..,c \quad (10.32)$$

donde $\mathbf{m}_i$, $\mathbf{m}_j$ son los vectores media de las clases c_i y c_j respectivamente. La distancia de Mahalanobis es una distancia implícita en la ecuación (10.15). Sin pérdida de generalidad, la distancia de un vector $\mathbf{x}_k$ a la clase c_i resulta ser: $d_M^2(\mathbf{x}_k,\mathbf{m}_i) = (\mathbf{x}_k - \mathbf{m}_i)^t C_i^{-1}(\mathbf{x}_k - \mathbf{m}_i)$.

En el supuesto de que las matrices de covarianza sean la identidad, la distancia de Mahalanobis al cuadrado resulta ser la distancia Euclídea al cuadrado, en cuyo caso tendríamos,

$$d_E^2(\mathbf{x},\mathbf{m}_i) = (\mathbf{x}-\mathbf{m}_i)^t(\mathbf{x}-\mathbf{m}_i) = \mathbf{x}^t\mathbf{x} - 2\mathbf{x}^t\mathbf{m}_i + \mathbf{m}_i^t\mathbf{m}_i \quad (10.33)$$

En la expresión anterior el término $\mathbf{x}^t\mathbf{x}$ no discrimina, ya que se repite en todas las clases, de forma que puede despreciarse. Ahora, si se cambia de signo y se divide por 2 en la ecuación (10.33) se obtiene la siguiente función discriminante,

$$fd_i(\mathbf{x}) = \mathbf{x}^t\mathbf{m}_i - \frac{1}{2}\mathbf{m}_i^t\mathbf{m}_i \quad (10.34)$$

Como el resultado de (10.33) es una cantidad positiva, al haber eliminado el término $\mathbf{x}^t\mathbf{x}$ y cambiado de signo los restantes, se deduce que la distancia Euclídea al cuadrado mínima hace la expresión (10.34) máxima.

10.3.3 Medidas estadísticas

A veces puede resultar de interés conocer el grado de separabilidad entre dos clases c_i y c_j. Algunas de estas medidas son las siguientes:

Divergencia

$$Diverg_{ij} = \frac{1}{2}Tr\left[(C_i - C_j)(C_j^{-1} - C_i^{-1})\right] + \frac{1}{2}Tr\left[(\mathbf{m}_i - \mathbf{m}_j)^t (C_i^{-1} + C_j^{-1})(\mathbf{m}_i - \mathbf{m}_j)\right] \quad (10.35)$$

donde $Tr[\cdot]$ es la traza de una matriz (suma de los elementos de la diagonal principal), C_i y C_j son las matrices de covarianza de las clases c_i y c_j y $\mathbf{m}_i$ y $\mathbf{m}_j$ son los vectores media de las clases c_i y c_j, finalmente el superíndice t indica vector transpuesto. A mayor divergencia, mayor separabilidad de las clases.

Coseno del ángulo formado por los vectores media

$$\cos\alpha_{ij} = \frac{\mathbf{m}_i^t \mathbf{m}_j}{|\mathbf{m}_i||\mathbf{m}_j|} \quad (10.36)$$

Es evidente que a mayor paralelismo entre ambos vectores, las correspondientes clases serán más similares. Por tanto, un valor alto de *cos* α expresa mayor similitud entre las clases.

Distancia de Jeffries-Matusita

$$J_{ij} = 2\left(1 - e^{-B_{ij}}\right) \quad (10.37)$$

donde B es la distancia de *Bhattacharya* dada por,

$$B_{ij} = \frac{1}{8}(\mathbf{m}_i - \mathbf{m}_j)^t \left(\frac{C_i + C_j}{2}\right)^{-1} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2}\ln\frac{2\left|(C_i + C_j)\right|}{\sqrt{|C_i|}\sqrt{|C_j|}} \quad (10.38)$$

donde $|C|$ indica el determinante de C. Cuanto mayor sea su valor mayor separabilidad.

10.4 CLASIFICADOR NO PARAMÉTRICO: VENTANA DE PARZEN

A diferencia de la estimación paramétrica, donde la función de densidad de probabilidad *fdp* se obtiene estimando los parámetros desconocidos de un modelo conocido, en la estimación no paramétrica no se conoce el modelo. Las técnicas no paramétricas son básicamente variaciones de la *aproximación del histograma* de una *fdp* desconocida. Consideremos el caso unidimensional. La figura 13.1 muestra dos ejemplos

de una *fpd* y su aproximación por el método del histograma, esto es, el eje x (espacio unidimensional) se divide primero en intervalos de longitud h.

La probabilidad de una muestra x localizada en un intervalo se puede deducir fácilmente. Si N es el número total de muestras y una fracción de ellas k_N se sitúan dentro de uno de los intervalos, la correspondiente probabilidad se aproxima por la *razón de frecuencia*

$$P \approx k_N / N \quad (10.39)$$

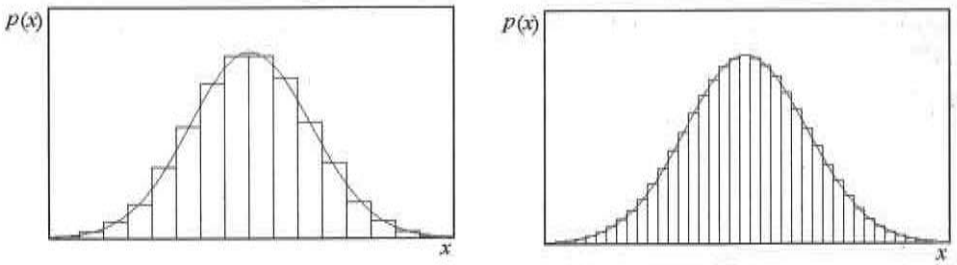


Figura 10.1 Aproximación de la función de densidad de probabilidad por el método del histograma con (a) intervalos de longitud grande y (b) pequeña

Esta aproximación tiende a la verdadera probabilidad P a medida que $N \to \infty$. El correspondiente valor de la *pdf* se considera constante a lo largo del intervalo y se aproxima por,

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{1}{h}\frac{k_N}{N}, \qquad |x - \hat{x}| \leq \frac{h}{2} \quad (10.40)$$

donde $\hat{x}$ es el punto medio del intervalo. Esto determina la amplitud de la curva del histograma sobre el intervalo. Ésta es una aproximación razonable para $p(x)$ continua y h suficientemente pequeña. Se puede demostrar que $\hat{p}(x)$ converge hacia el verdadero valor $p(x)$ a medida que $N \to \infty$ dado que,

$$h_N \to 0 \quad k_N \to \infty \quad \frac{k_N}{N} \to 0 \qquad (10.41)$$

se introduce h_N para mostrar la dependencia de N. Estas condiciones pueden comprenderse sin grandes detalles matemáticos. La primera ya ha sido discutida. Las otras dos muestran la forma en la que debe crecer k_N para garantizar la convergencia. En efecto, en todos los puntos donde $p(x) \neq 0$, una vez fijada la dimensión h_N

suficientemente pequeña, la probabilidad P de los puntos que caen en este intervalo es finita. Además, $k_N \approx PN$ y k_N tiende a infinito a medida que N crece a infinito. Por otra parte, a medida que la dimensión h_N del intervalo tiende a cero, la correspondiente probabilidad también tiende a cero, justificando la última condición. En la práctica el número N de datos es finito. Las condiciones precedentes dan idea de cómo deben elegirse los diferentes parámetros. N debe ser suficientemente grande, h_N suficientemente pequeño y el número de puntos dentro de cada intervalo suficientemente grande también. Cómo de grandes o pequeños depende del tipo de *fdp* y del grado de aproximación que se desea. A continuación describimos el método de la *Ventana de Parzen*.

En el caso multidimensional, en lugar de intervalos de dimensión h, el espacio p-dimensional se divide en hipercubos con la longitud de los lados h y volumen h^p. Sean $\mathbf{x}_i$, $i = 1,2,\ldots,N$ los vectores de atributos disponibles. Definimos la función $\phi(\mathbf{x})$ de modo que,

$$\phi(\mathbf{x}_i) = \begin{cases} 1 & para \quad |x_{ij}| \le \frac{1}{2} \\ 0 & en\ otros\ casos \end{cases} \qquad (10.42)$$

donde x_{ij}, $j = 1,..,d$ son las componentes de $\mathbf{x}_i$. En otras palabras, la función es igual a 1 para todos los puntos dentro del hipercubo de lado la unidad centrado en el origen y 0 fuera de él. Con esto la ecuación (13.40) se puede reescribir como,

$$\hat{p}(\mathbf{x}) = \frac{1}{h^p}\left(\frac{1}{N}\sum_{i=1}^{N}\phi\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)\right) \qquad (10.43)$$

La interpretación de la ecuación (10.43) es sencilla, consideramos un hipercubo con longitud de lado h centrado en $\mathbf{x}$, que es el punto donde queremos estimar la *fdp*. La suma es k_N, esto es, el número de puntos que caen dentro del hipercubo. Entonces la estima *fdp* se obtiene dividiendo k_N por N y el respectivo volumen del hipercubo h^p. No obstante, volviendo a la ecuación (10.43) y observándola desde una perspectiva diferente, vemos que estamos intentando aproximar una función continua $p(\mathbf{x})$ mediante una expansión de términos de funciones discontinuas $\phi(\cdot)$. Esto condujo a la generalización de (10.43) propuesta por Parzen (1962) utilizando funciones continuas en lugar de $\phi(\cdot)$, dichas funciones son conocidas como *núcleos* o *funciones potenciales* o *ventanas de Parzen*. Ejemplos típicos de funciones de este tipo son los núcleos Gaussianos,

$$p(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{(2\pi)^{p/2}|C|^{1/2}h^p}\exp\left[-\frac{1}{2h^2}(\mathbf{x}-\mathbf{x}_j)^t C^{-1}(\mathbf{x}-\mathbf{x}_j)\right]\right\} \qquad (10.44)$$

Una vez estimadas las funciones de densidad de probabilidad para cada una de las $c_1, c_2, \ldots, c_c$ clases a través de (10.44), la clasificación de una muestra x se lleva a cabo mediante la siguiente decisión,

$$\mathrm{x} \in c_i \quad \text{sii} \quad p_i(\mathrm{x}) > p_j(\mathrm{x}) \quad \forall j \neq i \qquad (10.45)$$

10.5 EJERCICIOS RESUELTOS

10.5.1 Agrupamiento borroso

***Ejercicio 10.1** Dado el conjunto de datos X de la tabla adjunta en el espacio bidimensional $\Re^2$

x_1	1	1	1	2	2	6	6	7	7	7
x_2	1	3	5	2	3	3	4	1	3	5

Clasificar dichos datos en dos clústeres con un valor del peso exponencial $m = 2$ y tomando como valor de $\varepsilon = 0.01$. Calcular los diferentes coeficientes de validación.

Dado el siguiente punto $x_k = (3,4)$ determinar los grados de pertenencia a cada uno de los clústeres.

Solución:

Antes de comenzar a resolver el ejercicio normalizamos los datos a clasificar en el rango [0,1] mediante la siguiente expresión, la normalización resulta conveniente para controlar el hecho de que todos los datos contribuyan por igual en el cómputo.

$$\boldsymbol{Y} = \frac{X - \boldsymbol{MIN}}{\boldsymbol{MAX} - \boldsymbol{MIN}} \qquad \boldsymbol{MIN} = \min\{X\}; \quad \boldsymbol{MAX} = \max\{X\}$$

Los valores $\boldsymbol{Y}$ normalizados a partir de X se muestran en la siguiente tabla, teniendo en cuenta que $\boldsymbol{MIN} = (1,1)$ y $\boldsymbol{MAX} = (7,5)$:

y_1	0.00	0.00	0.00	0.17	0.17	0.83	0.83	1.00	1.00	1.00
y_2	0.00	0.50	1.00	0.25	0.50	0.50	0.75	0.00	0.50	1.00

Como no nos han dado los valores iniciales de los centros de los clústeres, se utiliza el procedimiento pseudoaleatorio descrito en Balasko y col. (2006), dado por la siguiente expresión y teniendo en cuenta que el número de clústeres es dos.

$$\boldsymbol{v} = 2D\overline{\boldsymbol{M}} \circ \boldsymbol{R} + D\overline{\boldsymbol{m}}$$

donde $\overline{m}$ es la media de los valores de Y con dimensión $1xp$, $\overline{M} = \max(abs(Y - \overline{m}))$ es una matriz de dimensión $1xp$, $D = \left[\overbrace{1 \cdots 1}^{c}\right]^t$ cuya dimensión es $cx1$; R es una matriz de números aleatorios de dimensión cxp donde cada elemento de la matriz se obtiene utilizando la función $(rand(.) - 0.5)$, los valores aleatorios están restringidos al rango [0,1]. El operador $\circ$ expresa multiplicación de matrices elemento a elemento. Utilizando las expresiones anteriores obtenemos los dos clústeres dados a continuación,

$$v_1 = (0.95,\ 0.61); \qquad v_2 = (0.23,\ 0.49)$$

Utilizamos la distancia Euclídea e inicializamos la matriz $U^{(0)}$

$$U^{(0)} = 10^{-1}\begin{bmatrix} 0.49 & 0.03 & 0.83 & 0.06 & 0.00 & 9.95 & 9.94 & 8.33 & 9.99 & 9.67 \\ 9.51 & 9.97 & 9.17 & 9.93 & 10.0 & 0.05 & 0.06 & 1.67 & 0.00 & 0.33 \end{bmatrix}^t$$

Iteración 1:

$$v_1 = (0.93,\ 0.58); \qquad v_2 = (0.08,\ 0.44)$$

$$U^{(1)} = 10^{-1}\begin{bmatrix} 0.27 & 0.00 & 0.88 & 0.04 & 0.00 & 9.99 & 9.97 & 9.03 & 9.99 & 9.76 \\ 9.73 & 9.99 & 9.12 & 9.96 & 9.99 & 0.00 & 0.03 & 0.96 & 0.00 & 0.24 \end{bmatrix}^t$$

$\Delta = 0.0707 > \varepsilon$ (luego, seguir iterando)

Iteración 2:

$$v_1 = (0.93,\ 0.57); \qquad v_2 = (0.07,\ 0.44)$$

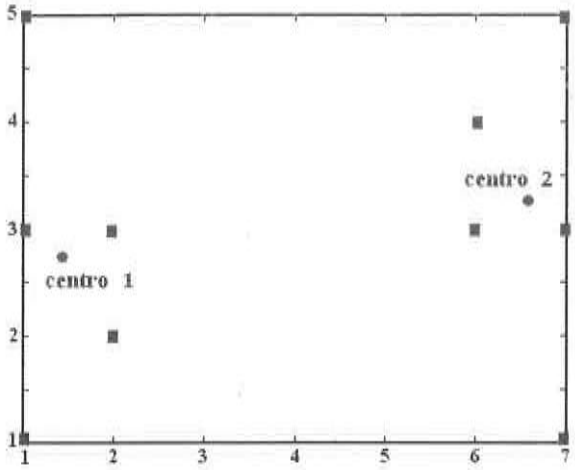
$$U^{(2)} = 10^{-1}\begin{bmatrix} 0.26 & 0.00 & 0.87 & 0.04 & 0.00 & 9.99 & 9.96 & 9.12 & 9.99 & 9.74 \\ 9.74 & 9.99 & 9.13 & 9.96 & 9.99 & 0.00 & 0.04 & 0.88 & 0.00 & 0.26 \end{bmatrix}^t$$

$\Delta = 0.0084 < \varepsilon$ (luego, terminar proceso iterativo)

Finalmente, se deshace la normalización para los vectores correspondientes a los centros de los clústeres aplicando la transformada inversa a la realizada previamente obteniendo los valores dados a continuación.

$$v_1 = (6.57,\ 3.27); \qquad v_2 = (1.43,\ 2.74)$$

En la figura siguiente aparecen los datos en forma de cuadrados y los centros en forma de círculos, correspondientes a los dos clústeres.



Los coeficientes de validación se muestran en la siguiente tabla, habiendo sido obtenidos a partir de las ecuaciones (10.8) a (10.14).

	CP	CE	SC	S	XB	DI	ADI
Validación	0.96	0.09	0.79	0.08	2.71	0.67	0.16

Mediante la ecuación (10.6) calculamos las distancias del punto $x_k = (2,3)$ dado a cada uno de los centros de los clústeres como sigue

$$\|x_k - v_1\|_G^2 = (2-6.57)^2 + (3-3.27)^2 = 20.96$$

$$\|x_k - v_2\|_G^2 = (2-1.43)^2 + (3-2.74)^2 = 0.39$$

Sustituyendo estos valores en la ecuación (10.6) se obtienen finalmente los siguientes valores para los grados de pertenencia: $\mu_{1k} = 0.1575$ y $\mu_{2k} = 0.8425$. Esto significa que el nuevo punto pertenece principalmente al clúster 2 debido a su mayor grado de pertenencia.

***Ejercicio 10.2** Un sistema basado en visión por computador trata de clasificar tres tipos de texturas naturales basadas en las tres componentes de color (R,G,B). Los tres tipos de texturas corresponden a paisajes naturales en los que se pretende distinguir áreas de cielo azul, zonas boscosas con predominio de verdes y ocres correspondientes a zonas sin cultivar. Se han extraído los siguientes datos de la imagen.