Ejercicios. Clasificador Estadístico

1. Dadas las siguientes clases representadas por sus centroides:

$$\bar{z}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix};\ \bar{z}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix};\ \bar{z}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix};\ \bar{z}_4 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

   a) Construya un reconocedor basado en la distancia Euclídea que resuelva este problema de clasificación.

   b) Dibuje las fronteras de indecisión y las regiones del espacio asignadas a cada clase por el reconocedor.

2. Se pretende discriminar entre dos tipos de objetos empleando las características x_1 y x_2. Para ello se han realizado medidas en cuatro muestras de cada clase, con los siguientes resultados:

$$\alpha_1 : \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \end{pmatrix} \right\}; \qquad \alpha_2 : \left\{ \begin{pmatrix} 5 \\ 0 \end{pmatrix} \begin{pmatrix} 5 \\ -3 \end{pmatrix} \begin{pmatrix} 3 \\ -2 \end{pmatrix} \begin{pmatrix} 7 \\ -2 \end{pmatrix} \right\}$$

   Se pide:

   a) Obtener las funciones discriminantes del clasificador euclídeo "modificado" (es decir, aquél cuyas funciones discriminantes son lineales: $\vec{W}^t \vec{X}$).

   b) Considerando ambas clases equiprobables a priori, obtener el clasificador estadístico gaussiano.

   c) Lo mismo que en el apartado anterior, pero ahora con la probabilidad a priori de α_1 igual a 0.7 y la de α_2 igual a 0.3.

   d) Realizar un estudio geométrico comparado de los resultados obtenidos en los apartados anteriores.

3. Para la clasificación de dos tipos de objetos se va a emplear una sola característica discriminante. Se han realizado 10 medidas de objetos de ambas clases, obteniéndose los siguientes resultados:

$$\alpha_1 : 7, 8, 7, 6, 4, 7, 5, 6, 6, 7$$
$$\alpha_2 : 13, 12, 14, 15, 14, 13, 12, 13, 12, 14$$

   Se pide:

   a) Suponiendo que las clases son equiprobables a priori, calcular el valor óptimo del clasificador estadístico, considerando que ambas clases son gaussianas.

   b) Lo mismo que en el apartando anterior, pero ahora con las probabilidades a priori $p(\alpha_1) = 0{,}7$ y $p(\alpha_2) = 0{,}3$.

4. Considérense las siguientes clases:

$$\alpha_1 : \left\{ \begin{pmatrix} -5 \\ -8 \end{pmatrix} \begin{pmatrix} -3 \\ -7 \end{pmatrix} \begin{pmatrix} -4 \\ -6 \end{pmatrix} \begin{pmatrix} -2 \\ -5 \end{pmatrix} \begin{pmatrix} -3 \\ -6 \end{pmatrix} \begin{pmatrix} -4 \\ -7 \end{pmatrix} \right\}$$

$$\alpha_2 : \left\{ \begin{pmatrix} -2 \\ -2 \end{pmatrix} \begin{pmatrix} -3 \\ -3 \end{pmatrix} \begin{pmatrix} -4 \\ -4 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \begin{pmatrix} -3 \\ -2 \end{pmatrix} \begin{pmatrix} -2 \\ -3 \end{pmatrix} \right\}$$

   con $p(\alpha_1) = 0{,}4$ y $p(\alpha_2) = 0{,}6$.

   Se pide:

   a) Teniendo en cuenta exclusivamente la característica x_1, obtener el umbral óptimo del clasificador estadístico, suponiendo una distribución normal.

b) Lo mismo, para la característica x_2.

c) Considerando ahora simultáneamente las dos características x_1 y x_2, obtener el clasificador basado en la distancia Euclídea.

d) Lo mismo que en (c), obtener el clasificador estadístico gaussiano.

e) Comparar geométricamene las soluciones obtenidas en (c) y (d).

5. Se trata de clasificar dos tipos de piezas producidas en una factoría a partir de su perímetro. A tal fin se han tomado al azar doce muestras, obteniéndose los siguientes resultados numéricos:

$$\text{Clase 1:}\{74, 66, 73, 65, 68, 73\};\ \text{Clase 2:}\{151, 149, 149, 151, 149, 151\}$$

Sabiendo que la factoría produce igual número de piezas de cada clase, se pide:

a) Deduzca la fórmula general del clasificador estadístico a priori con hipótesis de distribución gaussiana para el caso en que la dimensión del vector de características sea 1.

b) Calcule el valor del umbral de decisión del problema planteado para los siguientes clasificadores:

- Basado en la distancia Euclídea.
- Basado en la distancia de Mahalanobis.
- Estadístico a priori con hipótesis de distribución gaussiana.

c)
- Compare y justifique los resultados obtenidos en el apartado anterior.
- ¿En qué caso coincidirá el resultado del clasificador estadístico y el del basado en la distancia Euclídea?
- Idem para el clasificador estadístico y el basado en la distancia de Mahalanobis.
- ¿Qué valores deben tomar las probabilidades a priori en el apartado (b) para que los umbrales de decisión del clasificador estadístico a priori y del clasificador basado en la distancia de Mahalanobis coincidan?

6. Se desea construir un programa que ayude al personal de una guardería infantil a evaluar la presencia de una deficiencia en su alumnado. La característica t, que se empleará para construir el clasificador, se obtiene a partir de un test que realizan todos los alumnos. Dado que es la primera vez que se utiliza dicha prueba, no se dispone de información sobre los valores del índice para el alumnado. Por ello, con el fin de obtener algunas muestras de entrenamiento, se ha decidido seleccionar al azar a 10 alumnos del centro a los que se les aplica el test y evaluar a dichos alumnos con un psicólogo. Los resultados de este proceso se muestran en la siguiente tabla:

	Sin deficiencia							Con deficiencia		
t	8	7	9	0	-9	-7	-8	1	-1	0

Se pide:

a) Deduzca la fórmula general del clasificador estadístico paramétrico con hipótesis de distribución gaussiana, para el caso en que el vector de características sea unidimensional.

b) Compare y justifique las fronteras de indecisión del problema planteado para los siguientes clasificadores:

- Basado en la distancia Euclídea ($f_{de}(t) = w_1 t + w_0$).
- Basado en la distancia de Mahalanobis.
- Estadístico a priori con hipótesis de distribución gaussiana.

c) ¿En qué caso coincidirá el resultado del clasificador estadístico y el del basado en la distancia Euclídea?

d) ¿Qué valores deben tomar las probabilidades a priori en el apartado (b) para que los umbrales de decisión del clasificador estadístico a priori y del clasificador basado en la distancia de Mahalanobis coincidan?

e) Demuestre que el clasificador del vecino más próximo es un clasificador estadístico no paramétrico. ¿Cuáles son las limitaciones de este clasficador?

f) Represente gráficamente las fronteras de indecisión que obtendría el clasificador del vecino más próximo para el problema anterior. Compare los resultados con los obtenidos anteriormente.

7. Sea un problema de reconocimiento de formas biclase en el que se dispone de cinco muestras de entrenamiento de cada una de las clases:

$$\alpha_1 : \{(0,0);(1,0);(2,0);(2,1);(2,-1)\};\quad \alpha_2 : \{(0,1);(0,-1);(-1,0);(-2,1);(-2,-1)\}$$

a) Calcule la frontera de indecisión para el algoritmo de la distancia euclídea y el del vecino más próximo.

b) ¿Qué algoritmo separa mejor las muestras de ambas clases? ¿Cuáles son las limitaciones de dicho algoritmo? ¿Qué soluciones existen para mitigar estas limitaciones?

c) Aplique al problema anterior el algoritmo de edición de datos de Hart (Webb, p.100, sec. 3.3.4). Estudie si dicho algoritmo propone una solución óptima.

8. Demuestra que la frontera de indecisión del clasificador de la distancia euclídea es la mediatriz del segmento que une los centroides de las clases.

9. Dada la siguiente regla de clasificación para un problema biclase en un espacio de una dimensión: $x \in \alpha_1 \Leftrightarrow x > \theta$; en otro caso $x \in \alpha_2$.

a) Justifica que la probabilidad de error viene dada por

$$P(\mathcal{E}) = P(\alpha_1)\int_{-\infty}^{\theta} p(x|\alpha_1)dx + P(\alpha_2)\int_{\theta}^{\infty} p(x|\alpha_2)dx$$

b) Derivando la ecuación anterior, demuestra que una condición necesaria para minimizar $P(\mathcal{E})$ es que θ satisfaga

$$p(\theta|\alpha_1)P(\alpha_1) = p(\theta|\alpha_2)P(\alpha_2)$$

10. Dado un problema de clasificación biclase, demuestra que $f(\mathbf{x})$ es una función discriminante basada en la regionalización del espacio de trabajo

$$f(\mathbf{x}) = \ln\frac{p(\mathbf{x}|\alpha_1)}{p(\mathbf{x}|\alpha_2)} + \ln\frac{P(\alpha_1)}{P(\alpha_2)}.$$

11. Sea $\alpha_m(\mathbf{x})$ una función que para cualquier punto del espacio de trabajo $\mathbf{x}$ determina la clase que maximiza la probabilidad a posteriori: $\alpha_m(\mathbf{x}) = \arg\max_{i=1\ldots c}\{P(\alpha_i|\mathbf{x})\}$.

a) Demuestra que $P(\alpha_m(\mathbf{x})|\mathbf{x}) \geq \frac{1}{c}$.

b) Demuestra que para la regla de clasificación de mínimo error, la probabilidad de error vendrá dada por

$$P(\mathcal{E}) = 1 - \int P(\alpha_m(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

c) Emplea los resultados anteriores para demostrar que $P(\mathcal{E}) \leq \frac{c-1}{c}$.

d) ¿Cuándo ocurrirá que $P(\mathcal{E}) = \frac{c-1}{c}$?

12. ¿Cuál es la regla de clasificación con rechazo para un problema de coste o riesgo mínimo?

13. Demuestra que el clasificador de mínimo riesgo es equivalente al de mínimo error cuando $\lambda_{ij} = 1 \forall i \neq j$.

14. Demuestra que los estimadores de máxima verosimilitud de la media y la varianza de una gaussiana son respectivamente la media y la varianza muestral.

15. Demuestra que la media muestral es un estimador insesgado de la media y que, en cambio, la varianza muestral es un estimador sesgado de la varianza. ¿Cuál sería el estimador insesgado de la varizanza?

16. En general, si dos v.a. son incorreladas no puede afirmarse que sean independientes. Demuestra que si dichas variables son gaussianas entonces si son incorreladas también puede afirmare que son independientes.
¿Cuál sería en este caso la expresión de la distancia de Mahalanobis? **Sugerencia:** Demuestra que sin son incorreladas, la f.d.p conjunta $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \prod_i \mathcal{N}(x_i|\mu_i, \sigma_i^2)$.

17. Demuestra que el clasificador de la distancia euclídea es óptimo (obtiene la frontera de mínimo error) cuando las clases son equiprobables y siguen una distribución gaussiana tal que $\Sigma_i = \sigma^2 \mathtt{I}$.
¿Qué ocurre cuando las clases no son equiprobables?

18. Justifica la certeza o falsedad de la siguiente afirmación: *En un problema de clasificación biclase con medias y matrices de covarianzas arbitrarias y $P(\alpha_1) = P(\alpha_2)$, la frontera de indecisión está formada por los puntos del espacio de trabajo que tienen igual distancia de mahalanobis a cada media.*