# Ejercicios: clasificación supervisada

**1**. Considera un problema de clasificación en el que las dos clases son igualmente probables a priori. La distribución de $X$ condicionada a $Y = 1$ es normal con vector de medias $\mu_1$ y matriz de covarianzas $\sigma^2 I$, donde $I$ es la identidad. La distribución de $X$ condicionada a $Y = 0$ es normal con vector de medias $\mu_0$ y la misma matriz de covarianzas $\sigma^2 I$.

(a) Demuestra que en esta situación, la regla Bayes consiste en clasificar el punto $x$ en la clase $Y = 0$ si $x$ está más cerca de $\mu_0$ que de $\mu_1$ en el sentido de la distancia euclídea.
(b) Explica cómo se modificaría la regla anterior si la probabilidad a priori de $Y = 1$ es el doble que la de $Y = 0$.

**2**. Supongamos que la distribución de $X$ condicionada a $Y = 1$ es normal con vector de medias $\mu_1$ y matriz de covarianzas $\Sigma$, mientras que la distribución de $X$ condicionada a $Y = 0$ es normal con vector de medias $\mu_0$ y la misma matriz de covarianzas $\Sigma$ (caso homocedástico). Demuestra que la frontera que separa ambas poblaciones de acuerdo con la regla de clasificación lineal de Fisher es ortogonal a $\mu_0 - \mu_1$ si y sólo si $\mu_0 - \mu_1$ es un autovector de $\Sigma$.

**3**. Supongamos que la distribución de $X$ condicionada a $Y = 1$ tiene vector de medias $\mu_1$ y matriz de covarianzas $\Sigma$, mientras que la distribución de $X$ condicionada a $Y = 0$ tiene vector de medias $\mu_0$ y la misma matriz de covarianzas $\Sigma$ (caso homocedástico). Sea $w = \Sigma^{-1}(\mu_1 - \mu_0)$ y sea $\tilde{w}$ la función lineal discriminante canónica de Fisher. Demuestra que $w = [\tilde{w}'(\mu_1 - \mu_0)]\tilde{w}$.

**4**. Sea $(X, Y) \in \mathbb{R}^2$ un vector aleatorio tal que la distribución de $Y$ condicionada a $X$ es de Bernoulli de parámetro $(1 + e^{-\beta X})^{-1}$, donde $\beta \in \mathbb{R}$, y $\mathrm{P}(Y = 0) = \mathrm{P}(Y = 1) = 1/2$. Supongamos que queremos predecir $Y$ a partir de $X$. Responde a las siguientes preguntas, dejando el resultado en función del parámetro $\beta$.

(a) Determina la regla de clasificación óptima en este modelo.
(b) Si $X$ tiene distribución uniforme en $(0, 1)$, calcula el error de la regla de clasificación del apartado anterior (error Bayes).
(c) Si $X$ tiene distribución uniforme en $(0, 1)$, calcula $\lim_{n\to\infty} \mathrm{E}L_n$, donde $L_n$ es la probabilidad de error correspondiente al clasificador del vecino más próximo. Compara el resultado con el del apartado anterior.

**5**. En un grupo de 435 pacientes que habían sufrido quemaduras de tercer grado se midió el área de la zona afectada por las quemaduras [la variable `x` corresponde a $\log(\text{área} + 1)$]. Algunos de los pacientes sobrevivieron (`y=1`) y otros fallecieron (`y=0`). Con el fin de estudiar cómo influye el área de las quemaduras en la probabilidad de supervivencia se ajustó un modelo de regresión logística a los datos con los resultados siguientes:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8518  -0.6998   0.1860   0.5239   2.2089

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   22.708      2.266  10.020   <2e-16
x            -10.662      1.083  -9.848   <2e-16
---

    Null deviance: 524.69  on 433  degrees of freedom
Residual deviance: 335.23  on 432  degrees of freedom
AIC: 339.23
```

(a) ¿Aportan estos datos evidencia (a nivel $\alpha = 0{,}01$) de que cuanto mayor es el área afectada menor es la probabilidad de que el individuo sobreviva?

(b) Calcula un intervalo de confianza con nivel aproximado del 95 % para el parámetro $\beta_1$, donde $\beta_1$ es el coeficiente correspondiente a la variable `x`.

(c) Determina una regla de clasificación para predecir si un individuo sobrevivirá o no en función del valor de la variable `x`. ¿Cuál sería la predicción si `x` vale 2,5?

**6**. En un experimento descrito en Prentice (1976) se expuso una muestra de escarabajos a cierto pesticida. Tras cinco horas de exposición a distintos niveles de concentración del pesticida algunos de los escarabajos murieron y otros sobrevivieron. Los resultados para cada dosis aparecen en la tabla siguiente:

| Dosis ($\log_{10} CS_2 mgl^{-1}$) | N. insectos | N. muertos |
|---|---|---|
| 1.6907 | 59 | 6 |
| 1.7242 | 60 | 13 |
| 1.7552 | 62 | 18 |
| 1.7842 | 56 | 28 |
| 1.8113 | 63 | 52 |
| 1.8369 | 59 | 53 |
| 1.8610 | 62 | 61 |
| 1.8839 | 60 | 60 |

Formula un modelo de regresión logística para analizar estos datos y estima la probabilidad de que muera un escarabajo expuesto durante cinco horas a una dosis de concentración 1.8.

**7**. Supongamos que la distribución de $X$ condicionada a $Y = 1$ es normal con vector de medias $\mu_1$ y matriz de covarianzas $\Sigma$, mientras que la distribución de $X$ condicionada a $Y = 0$ es normal con vector de medias $\mu_0$ y la misma matriz de covarianzas $\Sigma$ (caso homocedástico). Demuestra que el error Bayes del correspondiente problema de clasificación es:

$$L^* = 1 - \Phi(\Delta/2),$$

donde $\Delta^2 = (\mu_0 - \mu_1)'\Sigma^{-1}(\mu_0 - \mu_1)]$ es el cuadrado de la distancia de Mahalanobis entre los dos vectores de medias y $\Phi$ es la función de distribución de una v.a. normal estándar. (Se supone $\pi_0 = \pi_1 = 1/2$).

**8**. Los datos del fichero `lirios.RData` corresponden a la longitud y anchura del pétalo y del sépalo de 100 lirios, 50 de ellos correspondientes a la especies *versicolor* y otros 50 de la especie *virginica*.

(a) Considera primero únicamente las dos variables correspondientes al sépalo. Calcula los coeficientes de la función discriminante lineal de Fisher y estima la probabilidad de error de esta regla mediante el riesgo empírico $\hat{L}_n$ y la tasa de error por validación cruzada $\hat{L}_n^{\text{vc}}$. Compara los valores de estos estimadores con el estimador *paramétrico* basado en el resultado del problema anterior: $1 - \Phi(\hat{\Delta}^2/2)$, donde $\hat{\Delta}^2 = (\hat{\mu}_0 - \hat{\mu}_1)'\hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1)]$.

(b) Repite el apartado anterior pero considerando las cuatro variables.

**9**. Sea $Y$ una variable aleatoria tal que $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$, y sea $X$ un vector aleatorio de dimensión $k$ tal que la distribución de $X$ condicionada a $Y = 1$ es $\mathrm{N}_k(0, 2I)$, donde $I$ es la matriz identidad de orden $k$, mientras que la distribución de $X$ condicionada a $Y = 0$ es $\mathrm{N}_k(0, I)$.

(a) Calcula la regla Bayes para clasificar un vector $x$ como correspondiente a $Y = 0$ ó $Y = 1$.

(b) Utiliza la regla Bayes para clasificar el punto $x = (2, \ldots, 2)' \in \mathbb{R}^k$.

(c) Da una expresión del error Bayes (es decir, el error de clasificación de la regla obtenida en el primer apartado) tan explícita como sea posible.