

En la figura se representan 4 muestras de aprendizaje de sendas clases:  $x_1 = (1, 2)^T$  de la clase  $c_1$ ,  $x_2 = (1, 2)^T$  de  $c_2$ ,  $x_3 = (2, 3)^T$  de  $c_3$ , y  $x_4 = (3, 2)^T$  de  $c_4$ . Supóngase que se ejecuta el algoritmo Perceptrón a partir de la parte más simple, con factor de aprendizaje  $\alpha = 1$ , margen  $\gamma = 0.1$  y vectores de pesos iniciales nulos (en notación homogénea). A continuación se muestra el inicio de una traza de ejecución del mismo:

**■ Iteración 1,  $x_1$ :**

$g_1(x_1) = 0$	$g_2(x_1) = \gamma > g_1(x_1)$ ? Si	$w_2 = w_2 - x_1 = (-1 - 1 - 2)^T$
$g_1(x_1) = 0$	$g_3(x_1) > g_1(x_1)$ ? Si	$w_3 = w_3 - x_1 = (-1 - 1 - 2)^T$
$g_1(x_1) = 0$	$g_4(x_1) > g_1(x_1)$ ? Si	$w_4 = w_4 - x_1 = (-1 - 1 - 2)^T$
$g_1(x_1) = 0$		$w_1 = w_1 + x_1 = (1, 2)^T$

**■ Iteración 1,  $x_2$ :**

$g_1(x_2) = -5$	$g_2(x_2) > g_1(x_2)$ ? Si	$w_1 = w_1 - x_2 = (0 - 1 - 1)^T$
$g_1(x_2) = 5$	$g_3(x_2) > g_1(x_2)$ ? Si	$w_3 = w_3 - x_2 = (-2 - 3 - 3)^T$
$g_1(x_2) = -5$	$g_4(x_2) > g_1(x_2)$ ? Si	$w_4 = w_4 - x_2 = (-2 - 3 - 3)^T$
$g_1(x_2) = -5$		$w_2 = w_2 + x_2 = (0 - 1 - 1)^T$

Se pide:

- Completa la traza hasta finalizar la primera iteración.
- ¿Cuáles datos de entrenamiento se clasifican erróneamente con los pesos obtenidos al finalizar la primera iteración?
- ¿Crees que el algoritmo Perceptrón convergería en este caso tras realizar un número finito de iteraciones? ¿Por qué?

**1. Iteración 1,  $x_3$ :**

$g_1(x_3) = -17$	$g_2(x_3) > g_1(x_3)$ ? Si	$w_1 = w_1 - x_3 = (-1 - 3 - 2)^T$
$g_1(x_3) = 1$	$g_3(x_3) > g_1(x_3)$ ? Si	$w_2 = w_2 - x_3 = (-1 - 1 - 4)^T$
$g_1(x_3) = -12$	$g_4(x_3) > g_1(x_3)$ ? Si	$w_4 = w_4 - x_3 = (-3 - 5 - 0)^T$
$g_1(x_3) = -17$		$w_3 = w_3 + x_3 = (-1 - 1 - 0)^T$

**■ Iteración 1,  $x_4$ :**

$g_1(x_4) = -30$	$g_2(x_4) > g_1(x_4)$ ? Si	$w_1 = w_1 - x_4 = (-2 - 0 - 4)^T$
$g_1(x_4) = -14$	$g_3(x_4) > g_1(x_4)$ ? Si	$w_3 = w_3 - x_4 = (-2 - 4 - 6)^T$
$g_1(x_4) = -12$	$g_4(x_4) > g_1(x_4)$ ? Si	$w_4 = w_4 - x_4 = (-2 - 4 - 2)^T$
$g_1(x_4) = -4$		$w_2 = w_2 + x_4 = (-2 - 2 - 4)^T$

**2. Dos, ya que se clasifican mal las muestras  $x_1$  y  $x_2$  (discriminante maximizada en negrita):**

$g_1(x_1) = -16$	$g_2(x_1) = -18$	$g_3(x_1) = -10$	$g_4(x_1) = -12$
$g_1(x_2) = -16$	$g_2(x_2) = -16$	$g_3(x_2) = -12$	$g_4(x_2) = -10$
$g_1(x_3) = -26$	$g_2(x_3) = -28$	$g_3(x_3) = -16$	$g_4(x_3) = -18$
$g_1(x_4) = -28$	$g_2(x_4) = -26$	$g_3(x_4) = -18$	$g_4(x_4) = -16$

3. El algoritmo Perceptrón sólo converge en este caso tras un número finito de iteraciones ya que las muestras son linealmente separables y el margen  $\gamma = 0.1$  es cercano a 0.

**Entrada:**  $\{(x_n, c_n)\}_{n=1}^N, \{\mathbf{w}_c\}_{c=0}^C, \alpha \in \mathbb{R}^{>0}$  y  $b \in \mathbb{R}$

**Salida:**  $\{\mathbf{w}_c\}^* = \arg \min_{\{\mathbf{w}_c\}} \sum_{\{P\}} \left[ \max_{c \neq c_n} \mathbf{w}_c^T \mathbf{x}_n + b > \mathbf{w}_{c_n}^T \mathbf{x}_n \right]$

**Método:**  $\{\mathbf{w}_c\} = \begin{cases} 1 & \text{si } P = \text{verdadero} \\ 0 & \text{si } P = \text{falso} \end{cases}$

**repetir**

**para todo dato  $\mathbf{x}_n$**

- err = falso**
- para toda clase  $c$  distinta de  $c_n$**
- si  $\mathbf{w}_c^T \mathbf{x}_n + b > \mathbf{w}_{c_n}^T \mathbf{x}_n$ :  $\mathbf{w}_c = \mathbf{w}_c - \alpha \cdot \mathbf{x}_n$ ,  $err = \text{verdadero}$**
- si  $err = 0$ :  $\mathbf{w}_{c_n} = \mathbf{w}_{c_n} + \alpha \cdot \mathbf{x}_n$**

**hasta que no quedan muestras mal clasificadas**

Todo clasificador puede representarse como:

$$c(\mathbf{x}) = \arg \max_c g_c(\mathbf{x})$$

donde cada clase  $c$  utiliza una **función discriminante**  $g_c(\mathbf{x})$  que mide el grado de pertenencia de un objeto  $\mathbf{x}$  a la clase  $c$ .

Las funciones discriminantes más utilizadas son **lineales** (con  $\mathbf{x}$ ):

$$g_c(\mathbf{x}) = \mathbf{w}_c^T \mathbf{x} + w_{c0} \quad \text{donde } \mathbf{x} = \begin{pmatrix} x_1 \\ i \\ x_D \end{pmatrix} \quad \text{y } \mathbf{w}_c = \begin{pmatrix} w_{c1} \\ i \\ w_{cD} \end{pmatrix}$$

Con notación **homogénea**:

$$g_c(\mathbf{x}) = \mathbf{w}_c^T \mathbf{x} \quad \text{donde } \mathbf{x} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \quad \text{y } \mathbf{w}_c = \begin{pmatrix} w_{c0} \\ \mathbf{w}_c \end{pmatrix}$$

Se tiene un problema de clasificación en dos clases, 0 y 1, para objetos representados en  $\{0, 1\}^2$ , esto es, vectores de bits de la forma  $\mathbf{x} = (x_1, x_2)$  con  $x_1, x_2 \in \{0, 1\}$ . Asimismo, disponemos de cuatro muestras de entrenamiento:

$x_{c1} = (0, 0)$	$x_{c2} = (1, 0)$	$x_{c3} = (0, 1)$	$x_{c4} = (1, 1)$
-------------------	-------------------	-------------------	-------------------

El clasificador Gaussiano se define como el clasificador de Bayes particularizando al caso en el que las funciones de densidad de probabilidad condicionales de las clases son de tipo Gaussiano:

$$p(\mathbf{x}|c) = (2\pi)^{-\frac{D}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \|\mathbf{x} - \mu_c\|^2 \Sigma_c^{-1} (\mathbf{x} - \mu_c)\right)$$

o, equivalentemente,

$$p(\mathbf{x}|c) = (2\pi)^{-\frac{D}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \|\mathbf{x} - \mu_c\|^2 \Sigma_c^{-1} (\mathbf{x} - \mu_c)\right)$$

para todo  $c = 1, \dots, C$ .

Supón que se tienen dos clases, A y B, de probabilidades a priori iguales y funciones de densidad de probabilidad condicionales de las clases tipo Gaussiano:

$p(\mathbf{x} A) \sim N_2(\mu_A, \Sigma_A)$ con $\mu_A = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ y $\Sigma_A = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$	$x = (x_1, x_2) \in \mathcal{R}_A : x_1 \in \mathbb{R} \wedge x_2 \in [-5, 1]$
$p(\mathbf{x} B) \sim N_2(\mu_B, \Sigma_B)$ con $\mu_B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ y $\Sigma_B = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}$	$x = (x_1, x_2) \in \mathcal{R}_B : x_1 \in \mathbb{R} \wedge x_2 \in [-\infty, 5] \setminus [5, 1], +\infty$

a)  $x_2 = -2 \pm 3 = \begin{cases} 1 \\ -5 \end{cases}$

El clasificador Gaussiano asociado, en términos de funciones discriminantes simplificadas es:

$g_A(\mathbf{x}) = -x_1^2 - x_2^2$	$x = (x_1, x_2) \in \mathcal{R}_A$
$g_B(\mathbf{x}) = -x_1^2 + 2x_2 + \frac{5}{2}$	$x = (x_1, x_2) \in \mathcal{R}_B$

Determina:

- La frontera de decisión inducida por el clasificador.
- La región de decisión asociada a cada clase.
- La clase a la cual pertenece el punto  $x = (-10, -10)$ .

Como se deduce del apartado c),  $x = (-10, -10) \in B$ .

**1. (1 punto)** Sean A y B dos clases de igual probabilidad a priori y distribuciones de probabilidad condicionales gaussianas:  $p(\mathbf{x}|A) \sim N_2(\mu_A, \Sigma_A)$  y  $p(\mathbf{x}|B) \sim N_2(\mu_B, \Sigma_B)$ , tal que:

$\begin{pmatrix} \mu_A \\ \Sigma_A \end{pmatrix} = \begin{pmatrix} \mu_B \\ \Sigma_B \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 2 & 6 \end{pmatrix}$	$\begin{pmatrix} \mu_A \\ \mu_B \\ \Sigma_A = \Sigma_B \\ \Sigma_B \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 & 6 \\ -1 & 1/2 \end{pmatrix}$
---	---

Se pide:

- Define el clasificador gaussiano de mínimo error (funciones discriminantes para A y B).
- Calcula la frontera de decisión entre ambos.
- Clasifica el punto  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ .

a) Como las matrizes de covarianzas son comunes, la función discriminante, para cada clase  $c$  se define como:  $g_c(\mathbf{x}) = \mathbf{w}_c^T \mathbf{x} + w_{c0}$ , donde:  $\mathbf{w}_c = \Sigma^{-1} \mu_c$ , y  $w_{c0} = \log(p(c)) - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c$ .

De este modo:

$$\mathbf{w}_A = \Sigma^{-1} \mu_A = \begin{pmatrix} 3 & -1 \\ -1 & 1/2 \end{pmatrix} \begin{pmatrix} 2 \\ 8 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$$

$$\mathbf{w}_B = \Sigma^{-1} \mu_B = \begin{pmatrix} 3 & -1 \\ -1 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$w_{AB} = \log(p(C)) - \frac{1}{2} \mu_A^T \Sigma^{-1} \mu_A = \log(0.5) - \frac{1}{2} (2, 8) \begin{pmatrix} 3 & -1 \\ -1 & 1/2 \end{pmatrix} \begin{pmatrix} 2 \\ 8 \end{pmatrix} = \log(0.5) - 6$$

$$w_{BB} = \log(p(C)) - \frac{1}{2} \mu_B^T \Sigma^{-1} \mu_B = \log(0.5) - \frac{1}{2} (1, 1) \begin{pmatrix} 3 & -1 \\ -1 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \log(0.5) - 1.5$$

$$g_A(\mathbf{x}) = \mathbf{w}_A^T \mathbf{x} + w_{A0} = \begin{pmatrix} 2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \log(0.5) - 6 = -2x_1 + 2x_2 + \log(0.5) - 6$$

$$g_B(\mathbf{x}) = \mathbf{w}_B^T \mathbf{x} + w_{B0} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \log(0.5) - 1.5$$

b) La frontera de decisión se definirá como:

$$g_A(\mathbf{x}) = g_B(\mathbf{x}) \quad \text{Clasificación del punto } \begin{pmatrix} 1 \\ 1 \end{pmatrix}:$$

$$-2x_1 + 2x_2 + \log(0.5) - 6 = -x_1 + x_2 + \log(0.5) - 1.5 \quad \text{en la clase } B$$

$$x_1 = -x_2 + 4.5 \quad \text{y } g_B(-1, -1) = 0.5 + \log(0.5) \Rightarrow \begin{pmatrix} -1 \\ -1 \end{pmatrix} \in B$$

Se tienen los 6 datos de entrenamiento bidimensionales mostrados abajo. Determina la proyección unidimensional óptima de los datos según PCA.

$x_1 = \begin{pmatrix} -3 \\ 1 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$	$x_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 1 \\ 1 \end{pmatrix}$
--	--

a)  $S = \begin{pmatrix} 6 & 0 \\ 0 & 1+2z^2 \end{pmatrix}$

b)  $E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$  y  $A = \begin{pmatrix} 6 & 0 \\ 0 & 1+2z^2 \end{pmatrix}$

Siendo  $z$  una constante positiva próxima a cero.

**Matriz covarianzas**

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Sea  $n$  un entero no negativo y  $p$  un real en  $[0, 1]$ . Decimos que una variable aleatoria  $x$  en  $\{0, 1, 2, \dots, n\}$  es Binomial( $n, p$ ) si su función de probabilidad es:

$$p_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ con } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

La distribución Binomial se puede ver como la distribución de la suma de  $n$  Bernoulli independientes con identica probabilidad de tomar el valor uno. Por ejemplo, dando un clasificador de probabilidad de error  $c$ , el número de errores cometidos  $n$  muestra de test sigue una distribución Binomial. La figura a la derecha muestra  $p_{n,p}(x)$ , con  $n = 100$  y  $p = 5\%$ , para todo  $x \in \{0, 1, \dots, 10\}$ .



Sea  $\lambda \in \mathbb{R}^+$ . Decimos que una variable aleatoria  $x \in \{0, 1, 2, \dots\}$  es Poisson( $\lambda$ ) si su función de masa de probabilidad es:

$$p_{\lambda}(x) = \frac{\exp(-\lambda) \lambda^x}{x!}$$

La distribución de Poisson se emplea para modelar la probabilidad de que un evento dado ocurra un cierto número de veces en un contexto prefijado. El parámetro  $\lambda$  es la media o esperanza de la distribución de probabilidad del evento. Por ejemplo,  $x$  podría ser el número de cibolas telefónicas que recibieron en un día o el número de extracciones de una muestra publicadas en un documento dado. La figura a la derecha muestra  $p_{\lambda}(x)$  para todo  $x \in \{0, 1, \dots, 11\}$ .



**(2 puntos)** Sea  $X = (x_1, \dots, x_N)^t$  una muestra aleatoria simple de una mixtura finita de  $C$  componentes unidimensionales con idéntica varianza:

$$p_{\Theta}(x) = \sum_c p_c p(x | c), \quad \text{con } p(x | c) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu_c)^2}{2\sigma^2}\right)$$

Supongamos que los coeficientes de la mixtura son conocidos, así como la varianza  $\sigma^2$ . Deriva una instancia del algoritmo EM (una iteración) para la estimación máximo-verosímil de las medias de las componentes; esto es, suponiendo que el vector de parámetros desconocidos es  $\Theta = (\mu_1, \dots, \mu_C)^t$ .

La log-verosimilitud de  $\Theta$  con respecto a  $X$  es:

$$L(\Theta) = \log p_{\Theta}(X) = \sum_n \log p_{\Theta}(x_n) = \sum_n \log \sum_c p_c \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n - \mu_c)^2}{2\sigma^2}\right)$$

Sea  $\Theta^{(k)} = (\mu_1^{(k)}, \dots, \mu_C^{(k)})^t$  una estimación de  $\Theta$  tras  $k$  iteraciones del algoritmo EM ( $k = 0, 1, \dots$ ).

El paso E es "idéntico" para toda mixtura finita: consiste en calcular, para todo  $\Theta$ , la función:

$$Q(\Theta | \Theta^{(k)}) = \sum_n \sum_c z_{nc}^{(k)} (\log p_c + \log p_{\Theta}(x_n | c))$$

donde  $z_{nc}^{(k)}$  es la probabilidad a posteriori de que  $x_n$  sea generado por la componente  $c$  según  $\Theta^{(k)}$ ; esto es,

$$z_{nc}^{(k)} = E_{p_{\Theta^{(k)}}(x_n | c)}(z_{nc}) = p_{\Theta^{(k)}}(z_{nc} = 1 | x_n) = \frac{p_c p_{\Theta^{(k)}}(x_n | c)}{\sum_c p_c p_{\Theta^{(k)}}(x_n | c)}.$$

El paso M consiste en optimizar  $Q$  con respecto a las medias:

$$\frac{\partial Q}{\partial \mu_c} = \sum_{n: c_n=c} z_{nc}^{(k)} \frac{\partial Q}{\partial \mu_c} = \frac{1}{\sigma^2} \sum_{n: c_n=c} z_{nc}^{(k)} (x_n - \mu_c) \Big|_{\mu^{(k+1)}} = 0 \rightarrow \boxed{\mu^{(k+1)} = \frac{1}{\sum_{n: c_n=c} z_{nc}^{(k)}} \sum_{n: c_n=c} z_{nc}^{(k)} x_n}$$

## Propiedades matrices

### (f) Propiedades de la transpuesta:

1.  $(A^T)^T = A$
2.  $(A+B)^T = A^T + B^T$
3.  $(AB)^T = B^T A^T$
4.  $(\alpha A)^T = \alpha A^T$

### (c) Propiedades de la inversa:

1.  $A^{-1}$  es única
2.  $(A^{-1})^{-1} = A$
3.  $(AB)^{-1} = B^{-1} A^{-1}$
4.  $(\alpha A)^{-1} = \frac{1}{\alpha} A^{-1} \quad \forall \alpha \neq 0$
5.  $(A^*)^{-1} = (A^{-1})^*$
6.  $(A^T)^{-1} = (A^{-1})^T$
7.  $A^{-1} = \frac{1}{\det(A)} (Adj(A)) \quad \text{donde } Adj(A) \text{ es la adjunta de } A$

### (b) Multiplicación de matrices:

1.  $A(B+C) = AB + AC$
2.  $(A+B)C = AC + BC$
3.  $A(BC) = (AB)C$
4.  $\alpha(AB) = (\alpha A)B = A(\alpha B)$
5.  $A_0 = 0_A = 0_B$
6.  $B_1 = I_B = B$
7. En general,  $AB \neq BA$  (la multiplicación no es commutativa)
8.  $AB = 0$  no implica necesariamente que  $A = 0$  ó  $B = 0$
9.  $AB = AC$  no implica necesariamente que  $B = C$

**Ejemplo:**  $c^*(x) = \frac{1}{C} \arg \max_c p(c | x)$  (para el problema ejemplo)

Su probabilidad de error para un  $x$  cualquiera es mínima:

$$\epsilon(c^*(x)) = 1 - P(c^*(x) | x) = 1 - \max_c P(c | x)$$

por lo cual también lo es su error, el **error de Bayes**:

$$\epsilon^* = E(\epsilon(c^*(x))) = \frac{1}{C} \sum_c P(c) \epsilon(c^*(x)) \text{ si } E \text{ es discreto}$$

**Ejemplo (cont.):**  $C = 2, D = 1$   $p_{11} = p_{22} = \frac{1}{2}, p_{12} = \frac{1}{2}, p_{21} = \frac{1}{4}$

$$\begin{aligned} \{ & (0,1), (1,2), (1,1), (0,1), (0,2), (0,1), (1,1) \\ & \hat{p}(1) = \frac{4}{7}, \hat{p}(2) = \frac{3}{7}, \hat{p}_{11} = \frac{1}{4}, \hat{p}_{21} = \frac{1}{3} \end{aligned}$$