UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica etsinf

Tema 5. Distribución de Bernoulli

Percepción (PER)

Curso 2017/2018

Departamento de Sistemas Informáticos y Computación

Índice

Índice

Introducción

Clasificador de Bayes:

$$c^*(x) = \underset{c=1,\dots,C}{\text{argmax}}\ P(c \mid x) = \underset{c=1,\dots,C}{\text{argmax}}\ \frac{P(c)\, p(x \mid c)}{p(x)}$$

$$= \underset{c=1,\dots,C}{\text{argmax}}\ P(c)\, p(x \mid c) = \underset{c=1,\dots,C}{\text{argmax}}\ \log P(c) + \log p(x \mid c)$$

- $P(c)$: probabilidad *a priori*
- $p(x|c)$: función de densidad (f.d.) de probabilidad condicional para clase c

En la práctica, se usan ***estimaciones*** de $P(c)$ y $p(x|c)$:

$$c^*(x) \approx \underset{c=1,\dots,C}{\text{argmax}}\ \log \hat{P}(c) + \log \hat{p}(x \mid c)$$

Introducción

$\hat{P}(c)$ y $\hat{p}(x \mid c)$ se estiman a partir de N muestras etiquetadas, $(x_1, c_1), \ldots, (x_N, c_N)$, extraídas aleatoriamente de $p(x, c)$

Estimación de la probabilidad *a priori*:

$$\hat{P}(c) = \frac{N_c}{N} \qquad\qquad N_c = \sum_{n\,:\,c_n = c} 1$$

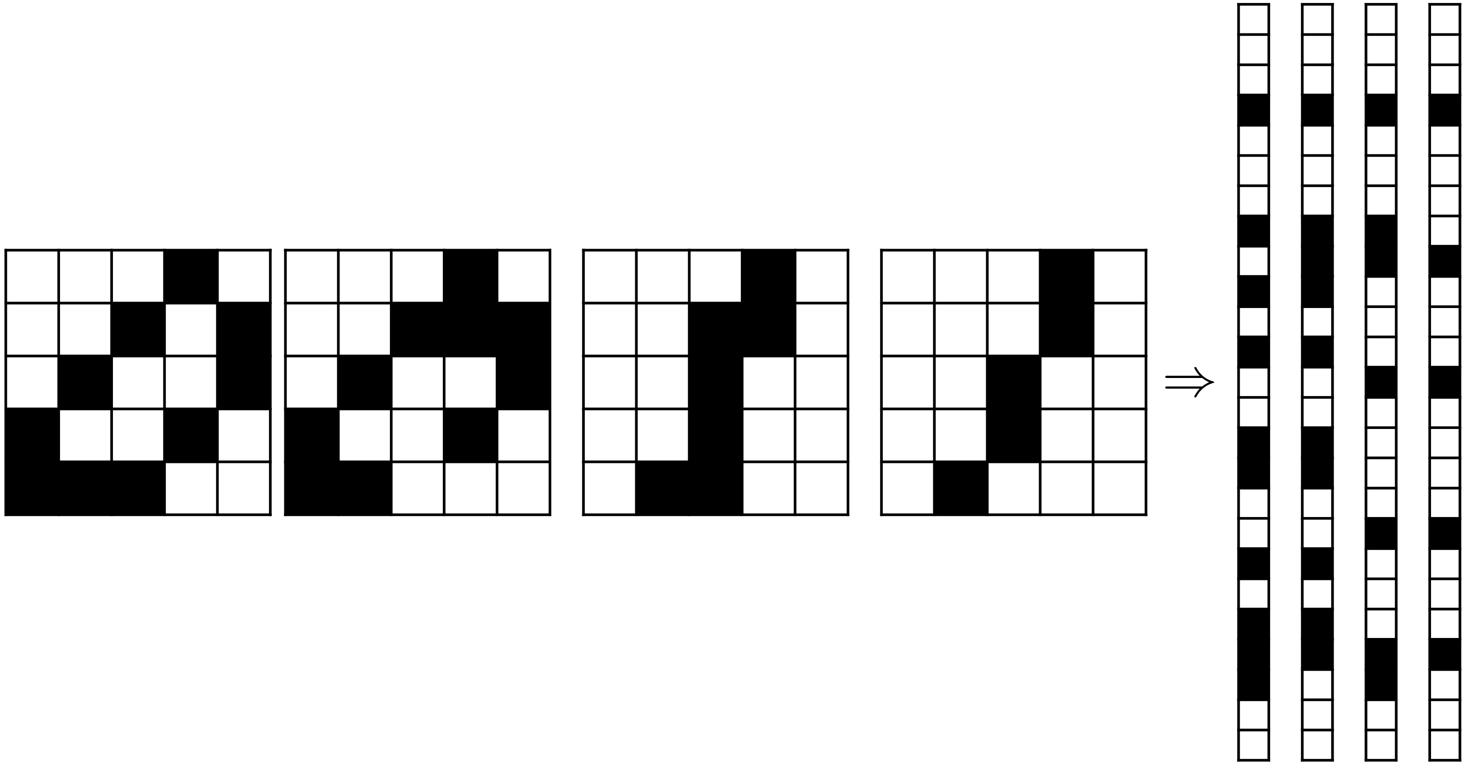
Estimación de la condicional $\hat{p}(x|c)$: a partir de las muestras x_n con $c_n = c$

Forma habitual: se asume un ***tipo de distribución*** sobre los datos de la clase y ***se estiman sus parámetros***

Motivación

Algunas tareas de RF conllevan objetos representados como un *vector de bits.*

Ejemplo: imágenes binarias de $5 \times 5 \rightarrow$ vectores de bits de 25 dimensiones



$\in \{0,1\}^{25}$

Idea: distribución de Bernoulli para modelar la condicional $p(x|c)$

Índice

Definición: Bernoulli unidimensional

Sea $p \in [0,1]$ y $q = 1 - p$.

Sea x una variable aleatoria que sigue una distribución de Bernoulli de parámetro p ($x \sim Be(p)$)

La f.d. de x es:

$$p(x) = \begin{cases} p & \text{si } x = 1 \\ q & \text{si } x = 0 \end{cases} \quad = \quad p\,x + q\,(1-x) \quad = \quad p^x\,q^{1-x}$$

Nota: $0^0 = 1$ y $0 \log 0 = 0$

Definición: Bernoulli multidimensional

Sean $x_1 \sim Be(p_1), \ldots, x_D \sim Be(p_D)$ independientes

En ese caso, $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ sigue una Bernoulli D-dimensional de parámetro $\boldsymbol{p} = (p_1, \ldots, p_D)^t$

La f.d. de $\boldsymbol{x}$ es:

$$p(\boldsymbol{x}) = \prod_{d=1}^{D} p(x_d) = \prod_{d=1}^{D} p_d\, x_d + q_d\, (1 - x_d) = \prod_{d=1}^{D} p_d^{x_d}\, q_d^{(1-x_d)}$$

Índice

Clasificador Bernoulli

Clasificador Bernoulli: clasificador de Bayes en el caso particular en que la f.d. condicional $p(\boldsymbol{x}|c)$ es una Bernoulli:

$$p(\boldsymbol{x} \,|\, c) \sim Be_D(\boldsymbol{p}_c), \quad c = 1, \ldots, C.$$

Por tanto:

$$\begin{aligned}
c^*(\boldsymbol{x}) &= \underset{c=1,\ldots,C}{\text{argmax}} \; \log\, P(c) + \log\, p(\boldsymbol{x} \mid c) \\
&= \underset{c=1,\ldots,C}{\text{argmax}} \; \log\, P(c) + \log \prod_{d=1}^{D} p_{cd}^{x_d} \, (1 - p_{cd})^{(1-x_d)} \\
&= \underset{c=1,\ldots,C}{\text{argmax}} \; \log\, P(c) + \sum_{d=1}^{D} x_d \log p_{cd} + (1 - x_d) \log(1 - p_{cd})
\end{aligned}$$

Clasificador Bernoulli

Agrupando términos dependientes e independientes de x_d:

$$c^*(\boldsymbol{x}) = \underset{c=1,\dots,C}{\text{argmax}} \left(\sum_{d=1}^{D} x_d (\log p_{cd} - \log(1-p_{cd})) \right) + \left(\log\ P(c) + \sum_{d=1}^{D} \log(1-p_{cd}) \right)$$

Reescribimos la expresión anterior como:

$$c^*(\boldsymbol{x}) = \underset{c=1,\dots,C}{\text{argmax}} \sum_{d=1}^{D} w_{cd}\, x_d + w_{c0}$$

donde

$$w_{cd} = \log p_{cd} - \log(1-p_{cd}) \qquad w_{c0} = \log\ P(c) + \sum_{d=1}^{D} \log(1-p_{cd})$$

Clasificador Bernoulli

Por tanto, es un ***clasificador lineal*** sobre $\boldsymbol{x}$:

$$c^*(\boldsymbol{x}) = \underset{c=1,\dots,C}{\text{argmax}}\; g_c(\boldsymbol{x}) = \underset{c=1,\dots,C}{\text{argmax}} \sum_{d=1}^{D} w_{cd}\, x_d + w_{c0}$$

Reescribiendo la expresión anterior como un producto escalar de dos vectores:

$$c^*(\boldsymbol{x}) = \underset{c=1,\dots,C}{\text{argmax}}\, \boldsymbol{w}_c^t\, \boldsymbol{x} + w_{c0}$$

donde

$$\boldsymbol{w}_c = \log \boldsymbol{p}_c - \log(\boldsymbol{1} - \boldsymbol{p}_c)$$

Índice

Entrenamiento por máxima verosimilitud

Sean un conjunto de entrenamiento de N muestras independientes e idénticamente distribuidas (i.i.d.) extraídas aleatoriamente de C distribuciones Bernoulli:

$$\{(\boldsymbol{x}_n, c_n)\}_{n=1}^{N} \quad \text{i.i.d.} \quad p(\boldsymbol{x}, c) = P(c)\, p(\boldsymbol{x}|c), \quad p(\boldsymbol{x}|c) \sim Be_D(\boldsymbol{p}_c)$$

Conjunto de parámetros a estimar $\boldsymbol{\Theta}$:

- Probabilidades *a priori*: $P(1) \ldots, P(C)$
- Parámetros de las Bernoulli para cada clase c: $\boldsymbol{p}_c$, $c = 1, \ldots, C$

Por ***criterio de máxima verosimilitud*** (MV), se estima $\boldsymbol{\Theta}$ como:

$$\hat{P}(c) = \frac{N_c}{N} \qquad c = 1, \ldots, C$$

$$\hat{\boldsymbol{p}}_c = \frac{1}{N_c} \sum_{n\,:\,c_n = c} \boldsymbol{x}_n \qquad c = 1, \ldots, C$$

Índice

Suavizado de la distribución Bernoulli

Problema: muchos criterios de entrenamiento (incluído MV) pueden generar clasificadores sobreentrenados

Soluciones:

- Cambiar el criterio de aprendizaje
- ***Suavizar*** los parámetros estimados

Opciones de suavizado en Bernoulli:

- Truncamiento simple
- Muestra ficticia

Suavizado de la distribución Bernoulli

Truncamiento simple

Dado ϵ, $0 \leq \epsilon \leq 0.5$, redefinir $\hat{p}_{cd}$ como:

$$\tilde{p}_{cd} = \begin{cases} \epsilon & \text{si } \hat{p}_{cd} < \epsilon \\ 1-\epsilon & \text{si } \hat{p}_{cd} > 1-\epsilon \\ \hat{p}_{cd} & \text{en otro caso} \end{cases}$$

Muestra ficticia

Añadir al conjunto de aprendizaje $(\mathbf{0}, c)$ y $(\mathbf{1}, c)$, $c = 1, \ldots, C$.

Equivale a redefinir la estimación de $\hat{\boldsymbol{p}}_c$ como:

$$\tilde{\boldsymbol{p}}_c = \frac{1}{N_c + 2}\left(\mathbf{1} + \sum_{n\,:\,c_n = c} \boldsymbol{x}_n\right)$$