UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria **Informàtica**

etsinf

Tema 5. Distribuciones de probabilidad

Percepción (PER)

Curso 2019/2020

Departamento de Sistemas Informáticos y Computación

Índice

Índice

Introducción

Clasificador de Bayes:

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

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

- $P(c)$: 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 $p(x|c)$:

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

Introducción

$\hat{P}(c)$ y $\hat{p}(x \mid c)$ se estiman desde 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***

Introducción

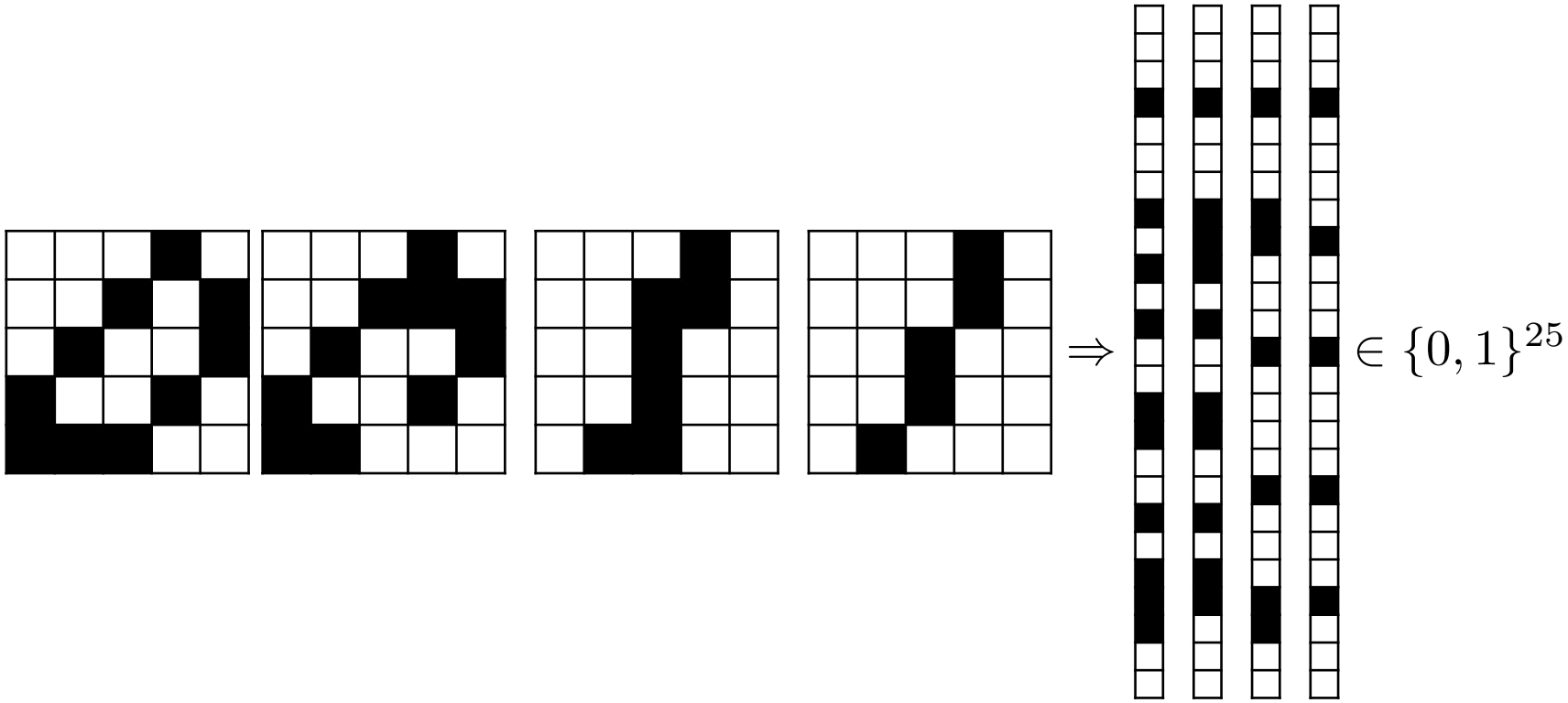
Estudiaremos tres tipos de distribución de probabilidad $p(x \mid c)$ que son aplicables en función de los datos a modelizar:

- Distribución de Bernoulli: datos binarios
- Distribución multinomial: datos que son contadores (enteros positivos)
- Distribución Gaussiana: datos que son números reales

Bernoulli: 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

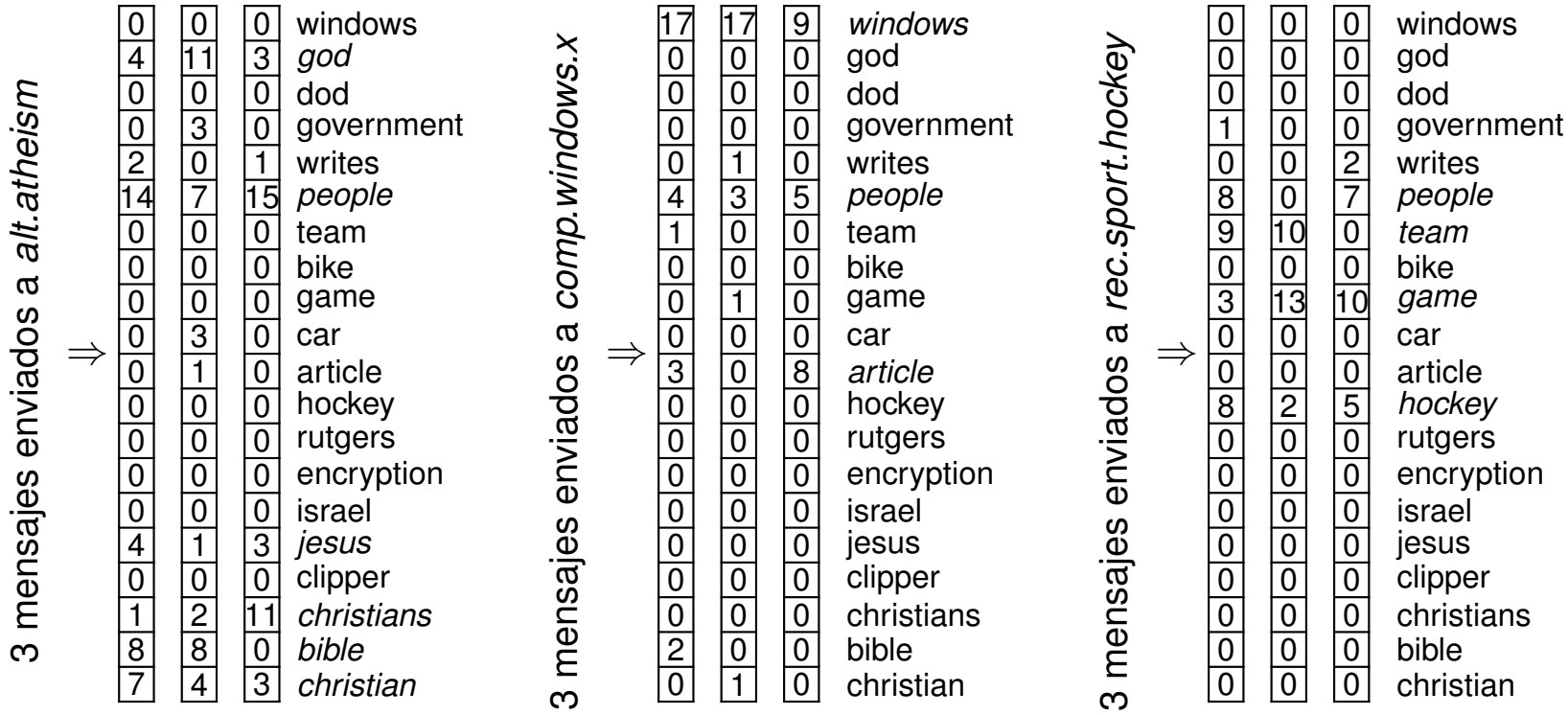


Idea: distribuciones de Bernoulli para modelar la condicional $p(x|c)$

Multinomial: Motivación

Algunas tareas de RF representan objetos como *vectores de cuentas*

Ejemplo: Texto representado como *bag-of-words*

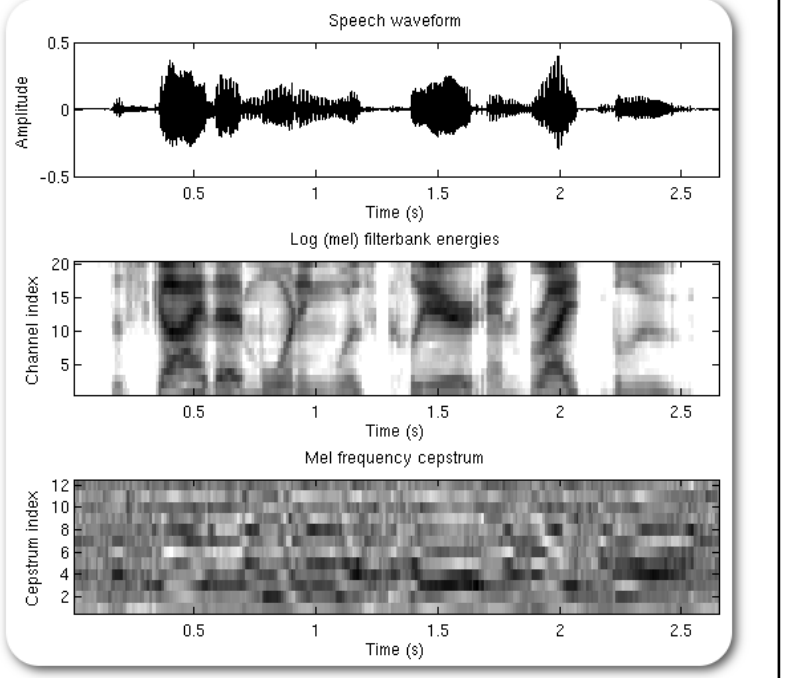


Idea: usar la *distribución multinomial* para modelizar la condicional $p(\boldsymbol{x}|c)$

Gaussiana: Motivación

Algunas tareas representan objetos por *vectores de características reales* ($\mathbb{R}^D$)

Ejemplo: Señal acústica mediante vectores de coeficientes cepstrales



$\Rightarrow$

18637.949219 520.131897 -421.241852 1373.556641 53.300270 1169.587769 -212.973862 . . . 266.111328 1038.843018 -41.131569 284.150818 -51.623795 27.055664 177.560684

15014.219727 -261.046661 -803.480652 1420.215332 -190.737427 1422.750366 -405.564545 . . . -120.428261 -108.468079 29.809811 69.221519 94.645294 20.136948 -27.475578

. . .

10495.829102 -551.998047 -471.013458 815.385010 -325.771301 891.692322 -281.717865 . . . -267.436554 -107.579346 -85.941284 -184.720337 -33.269119 95.332092 -81.534462

Idea: usar la *distribución gaussiana* para modelizar la condicional $p(\boldsymbol{x}|c)$

Introducción

Para cada distribución de probabilidad veremos:

- Su definición formal
- El clasificador asociado
- Su estimación por máxima verosimilitud (MV)
- Las técnicas de suavizado asociadas

Índice

Definición: Bernoulli unidimensional

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

Sea x una variable aleatoria binaria 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: $\quad 0^0 = 1 \quad$ y $\quad 0 \log 0 = 0$

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:

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

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:

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

Reescribimos la expresión anterior como:

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

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 $\boldsymbol{x}$:

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

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

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

donde

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

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:

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

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

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:

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

Índice

Definición: distribución multinomial

Sea una población $\mathcal{Y} = \{y_1, \ldots, y_l\}$ con $y_i \in \{1, \ldots, D\}$

Sean las proporciones p_d de los tipos de elemento $\{1, \ldots, D\}$ dadas por:

$$\boldsymbol{p} = (p_1, \ldots, p_D)^t \in [0,1]^D \qquad \text{con} \qquad \sum_{d=1}^{D} p_d = 1$$

Sea una secuencia de N elementos formada por extracción aleatoria con reemplazo desde $\mathcal{Y}$

$$w_1^N = w_1\, w_2 \cdots w_N$$

Número de secuencias distintas de longitud N:

$$\mathrm{VR}_{D,N} = D^N$$

Definición: distribución multinomial

Asumiendo independencia entre elementos:

$$p(w_1^N) = p_{w_1} p_{w_2} \cdots p_{w_N}$$

No depende del orden de los elementos, sino de su número de ocurrencias:

- x_d: el número de ocurrencias del elemento d en w_1^N
- $\boldsymbol{x} = (x_1, \ldots, x_D)^t$: vector de ocurrencias (número de ocurrencias de cada elemento en w_1^N)

$$p(w_1^N) = p_1^{x_1} \cdots p_D^{x_D} = \prod_{d=1}^{D} p_d^{x_d}$$

El número de secuencias diferentes con el mismo vector de ocurrencias es un *coeficiente multinomial:*

$$\binom{N}{\boldsymbol{x}} = \binom{N}{x_1, \ldots, x_D} = \frac{N!}{x_1! \cdots x_D!}$$

Definición: distribución multinomial

Distribución multinomial: se define sobre el espacio de vectores de ocurrencias

La probabilidad de $\boldsymbol{x}$ es la suma de probabilidades de todas las secuencias con vector de ocurrencias $\boldsymbol{x}$:

$$p(\boldsymbol{x}) = \binom{N}{\boldsymbol{x}} \prod_{d=1}^{D} p_d^{x_d}$$

$p(\boldsymbol{x})$ es una f.d. multinomial:

- D-dimensional
- Longitud $N = \sum_{d=1}^{D} x_d$
- Prototipo $\boldsymbol{p}$

De ahora en adelante, usaremos $x_+ = N$.

Clasificador multinomial

Clasificador multinomial: clasificador de Bayes donde la f.d. condicional $p(\boldsymbol{x}|c)$ es una multinomial

$$p(\boldsymbol{x} \mid c) \sim Mult_D(x_+, \boldsymbol{p}_c), \quad c = 1, \ldots, C.$$

Por tanto:

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

$$= \underset{c=1,\ldots,C}{\text{argmax}} \ \log P(c) + \log \frac{x_+!}{x_1! \cdots x_D!} \prod_{d=1}^{D} p_{cd}^{x_d}$$

$$= \underset{c=1,\ldots,C}{\text{argmax}} \ \log P(c) + \log \frac{x_+!}{x_1! \cdots x_D!} + \sum_{d=1}^{D} x_d \log p_{cd}$$

Clasificador multinomial

Eliminando el término independiente de c:

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

Expresando el sumatorio en forma de producto escalar:

$$c^*(\boldsymbol{x}) = \underset{c=1,\ldots,C}{\text{argmax}} \ (\log \boldsymbol{p}_c)^t \, \boldsymbol{x} + \log \ P(c)$$

En forma de clasificador lineal:

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

Con:

$$\boldsymbol{w}_c = \log \boldsymbol{p}_c \qquad w_{c0} = \log P(c)$$

Entrenamiento por máxima verosimilitud

Sean N muestras de entrenamiento aleatoriamente extraídas de C distribuciones multinomiales independientes:

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

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

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

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

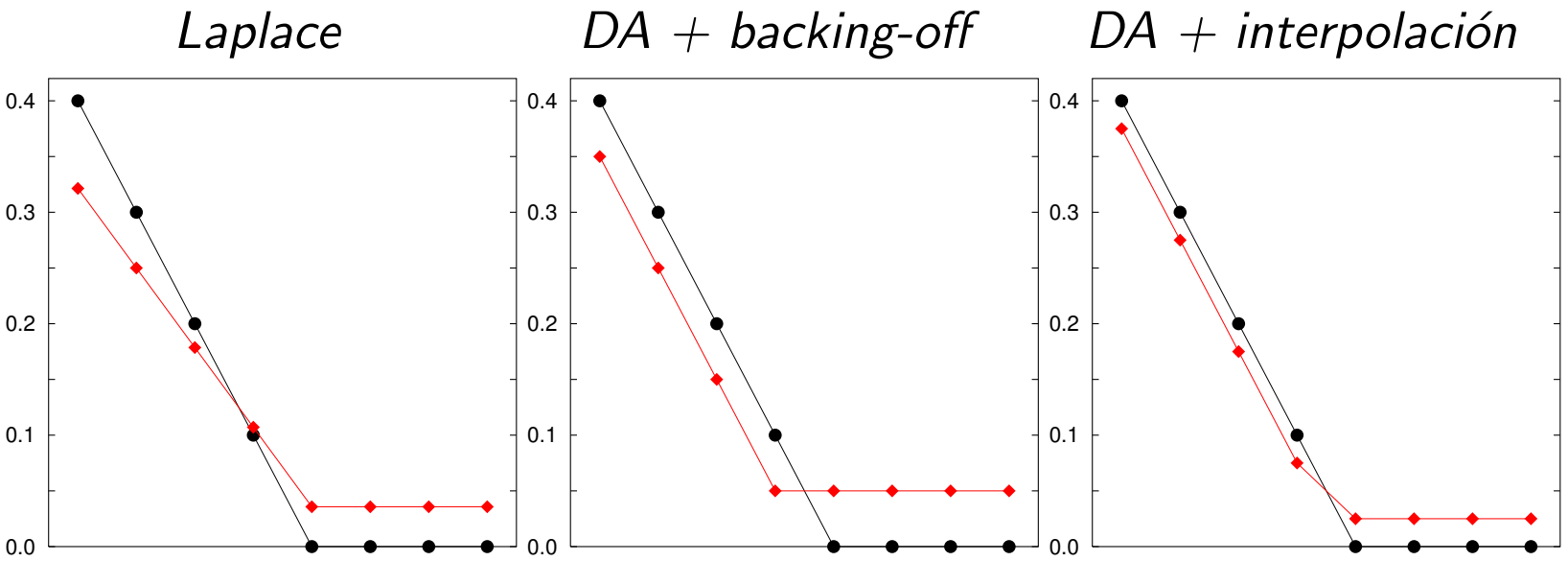
$$\hat{P}(c) = \frac{N_c}{N} \qquad \hat{\boldsymbol{p}}_c = \frac{1}{\sum\limits_{n\,:\,c_n=c} \sum\limits_{d=1}^{D} x_{nd}} \sum_{n\,:\,c_n=c} \boldsymbol{x}_n \qquad c = 1, \ldots, C$$

Suavizado de la distribución multinomial

Laplace: suma una constante $\epsilon > 0$ a cada parámetro y renormaliza

Descuento Absoluto (DA):

1. Descuenta una constante $\epsilon > 0$ (pequeña) a cada parámetro mayor que cero
2. Distribuir la probabilidad descontada según una *distribución generalizada*:
   - Entre todos los parámetros nulos *(backing-off)*
   - Entre todos los parámetros *(interpolación)*



Índice

Definición: distribución gaussiana unidimensional

Sea x una variable aleatoria unidimensional

Gaussiana unidimensional estandarizada

$x \sim \mathcal{N}(0, 1)$ presenta una distribución de probabilidad

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2\right)$$

Gaussiana unidimensional general

$x \sim \mathcal{N}(\mu, \sigma)$, con media $\mu \in \mathbb{R}$ y varianza $\sigma^2 \in \mathbb{R}^+$, presenta una distribución de probabilidad

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Definición: distribución gaussiana multidimensional

Sea $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ una variable aleatoria D-dimensional

Gaussiana estandarizada

$\boldsymbol{x} \sim \mathcal{N}_D(\mathbf{0}, I_D)$, donde $x_1, \ldots, x_D \sim \mathcal{N}(0,1)$ independientes, presenta una distribución de probabilidad:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{x}\right)$$

Definición: distribución gaussiana multidimensional

Gaussiana general

Sean:

- $\boldsymbol{z} \sim \mathcal{N}_D(\mathbf{0}, I_D)$
- $\boldsymbol{\mu} \in \mathbb{R}^D$
- $A \in \mathbb{R}^{D \times D} : |A| \neq 0$
- $\Sigma = AA^t$ (simétrica y definida positiva) con:
  - $A = W\Delta^{\frac{1}{2}}$
  - W vectores propios de Σ
  - Δ valores propios de Σ
- $\boldsymbol{x} = A\boldsymbol{z} + \boldsymbol{\mu}$

$\boldsymbol{x} \sim \mathcal{N}_D(\boldsymbol{\mu}, \Sigma)$, con media $\boldsymbol{\mu}$ y matriz de covarianzas Σ, presenta una distribución de probabilidad:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

Clasificador gaussiano

Clasificador gaussiano: clasificador de Bayes donde la f.d. condicional $p(\boldsymbol{x}|c)$ es una gaussiana

$$p(\boldsymbol{x} \,|\, c) \sim \mathcal{N}_D(\boldsymbol{\mu}_c, \Sigma_c), \quad c = 1, \ldots, C$$

Por tanto:

$$\begin{aligned}
c^*(\boldsymbol{x}) &= \underset{c=1,\ldots,C}{\text{argmax}} \; \log P(c) + \log p(\boldsymbol{x} \mid c) \\
&= \underset{c=1,\ldots,C}{\text{argmax}} \; \log P(c) - \frac{D}{2}\log 2\pi - \frac{1}{2}\log |\Sigma_c| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^t\Sigma_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) \\
&= \underset{c=1,\ldots,C}{\text{argmax}} \; \log P(c) - \frac{1}{2}\log |\Sigma_c| - \frac{1}{2}\boldsymbol{x}^t\Sigma_c^{-1}\boldsymbol{x} + \boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{\mu}_c \\
&= \underset{c=1,\ldots,C}{\text{argmax}} \; -\frac{1}{2}\boldsymbol{x}^t\Sigma_c^{-1}\boldsymbol{x} + \boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{x} + \left(\log P(c) - \frac{1}{2}\log |\Sigma_c| - \frac{1}{2}\boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{\mu}_c\right)
\end{aligned}$$

Clasificador gaussiano

$$c^*(\boldsymbol{x}) = \underset{c=1,\ldots,C}{\text{argmax}} \; -\frac{1}{2}\boldsymbol{x}^t\Sigma_c^{-1}\boldsymbol{x} + \boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{x} + \left(\log P(c) - \frac{1}{2}\log\,|\Sigma_c| - \frac{1}{2}\boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{\mu}_c\right)$$

Clasificador ***cuadrático*** con $\boldsymbol{x}$:

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

Con:

$$W_c = -\frac{1}{2}\Sigma_c^{-1} \qquad \boldsymbol{w}_c = \Sigma_c^{-1}\boldsymbol{\mu}_c$$

$$w_{c0} = \log P(c) - \frac{1}{2}\log\,|\Sigma_c| - \frac{1}{2}\boldsymbol{\mu}_c^t\Sigma_c^{-1}\boldsymbol{\mu}_c$$

Clasificador gaussiano

Caso particular: ***matriz de covarianzas común***, $\Sigma_c = \Sigma$

En ese caso, tanto $-\frac{1}{2}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x}$ como $-\frac{1}{2}\log|\Sigma|$ son independientes de c

$$c^*(\boldsymbol{x}) = \underset{c=1,\ldots,C}{\text{argmax}}\ \boldsymbol{\mu}_c^t\Sigma^{-1}\boldsymbol{x} + \left(\log P(c) - \frac{1}{2}\boldsymbol{\mu}_c^t\Sigma^{-1}\boldsymbol{\mu}_c\right)$$

El clasificador gaussiano es ***lineal***:

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

Con:

$$\boldsymbol{w}_c = \Sigma^{-1}\boldsymbol{\mu}_c \qquad w_{c0} = \log P(c) - \frac{1}{2}\boldsymbol{\mu}_c^t\Sigma^{-1}\boldsymbol{\mu}_c$$

Entrenamiento por máxima verosimilitud

Sean N muestras de entrenamiento aleatoriamente extraídas de C distribuciones gaussianas independientes

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

Conjunto de parámetros a estimar Θ:

- Probabilidades *a priori*: $P(1), \ldots, P(C)$
- Medias para cada clase: $\mu_1, \ldots, \mu_C$
- Matrices de covarianza para cada clase: $\Sigma_1, \ldots, \Sigma_C$

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

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

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{n:c_n=c} \boldsymbol{x}_n \qquad (2)$$

$$\hat{\Sigma}_c = \frac{1}{N_c} \sum_{n\,:\,c_n=c} (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_c)^t = \left(\frac{1}{N_c} \sum_{n\,:\,c_n=c} \boldsymbol{x}_n \boldsymbol{x}_n^t \right) - \hat{\boldsymbol{\mu}}_c \hat{\boldsymbol{\mu}}_c^t \qquad (3)$$

Entrenamiento por máxima verosimilitud

En el caso de Σ común para todas las clases ($\Sigma_c = \Sigma$), el conjunto de parámetros a estimar Θ es:

- Probabilidades *a priori*: $P(1), \ldots, P(C)$
- Medias para cada clase: $\mu_1, \ldots, \mu_C$
- Matriz de covarianza común: Σ

Por ***criterio de máxima verosimilitud***, la estimación de $\boldsymbol{\Theta}$ se calcula como en el caso general (Ecuaciones (1) y (2) para $\hat{P}(c)$ y $\hat{\boldsymbol{\mu}}_c$ respectivamente) y:

$$\hat{\Sigma} = \sum_c \hat{P}(c)\,\hat{\Sigma}_c = \frac{1}{N}\sum_n \boldsymbol{x}_n \boldsymbol{x}_n^t - \sum_c \hat{P}(c)\,\hat{\boldsymbol{\mu}}_c \hat{\boldsymbol{\mu}}_c^t \tag{4}$$

Con $\hat{\Sigma}_c$ calculada según Ecuación (3)

Suavizado

Umbralizado de covarianza [Duda01, pág. 113]

Covarianzas con magnitud de la correlación no cercana a uno valen cero:

$$\tilde{\sigma}^2_{cdd'} = \begin{cases} \hat{\sigma}^2_{cdd'} & \text{si } |\hat{\rho}_{cdd'}| > 1 - \epsilon \\ 0 & \text{otro caso} \end{cases} \qquad \forall c, d, d' = 1, \ldots, D;\ d \neq d'$$

Donde:

- ϵ es una constante pequeña no negativa ($\epsilon = 0 \rightarrow \Sigma$ diagonal)
- Coeficiente de correlación: $\hat{\rho}_{cdd'} = \frac{\hat{\sigma}^2_{cdd'}}{\hat{\sigma}_{cdd}\,\hat{\sigma}_{cd'd'}}$

Flat smoothing

Combinación lineal de cada $\hat{\Sigma}_c$ y $\tilde{\Sigma}$ (matriz de covarianza global suavizada):

$$\tilde{\Sigma}_c = \alpha\,\hat{\Sigma}_c + (1 - \alpha)\,\tilde{\Sigma} \qquad \forall c \quad \alpha \in [0, 1]$$

Donde: $\tilde{\Sigma} = \beta\hat{\Sigma} + (1 - \beta)I$, $\beta \in [0, 1]$