



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica



Tema 5. Distribuciones de probabilidad

Percepción (PER)

Curso 2020/2021

Departamento de Sistemas Informáticos y Computación

Índice

- 1 Introducción y motivación ▷ 3
- 2 Distribución de Bernoulli ▷ 11
- 3 Distribución multinomial ▷ 20
- 4 Distribución Gaussiana ▷ 28

Índice

- 1 *Introducción y motivación* ▷ 3
- 2 Distribución de Bernoulli ▷ 11
- 3 Distribución multinomial ▷ 20
- 4 Distribución Gaussiana ▷ 28

Introducción

Clasificador de Bayes (datos vectoriales):

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} P(c | \mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \frac{P(c) p(\mathbf{x} | c)}{p(\mathbf{x})} =$$

$$\operatorname{argmax}_{c=1,\dots,C} P(c) p(\mathbf{x} | c) = \operatorname{argmax}_{c=1,\dots,C} \log P(c) + \log p(\mathbf{x} | c)$$

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

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

$$c^*(\mathbf{x}) \approx \operatorname{argmax}_{c=1,\dots,C} \log \hat{P}(c) + \log \hat{p}(\mathbf{x} | c)$$

Introducción

$\hat{P}(c)$ y $\hat{p}(\mathbf{x} \mid c)$ se estiman desde N muestras etiquetadas, $(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)$, extraídas aleatoriamente de $p(\mathbf{x}, c)$

Estimación de la probabilidad *a priori*:

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

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

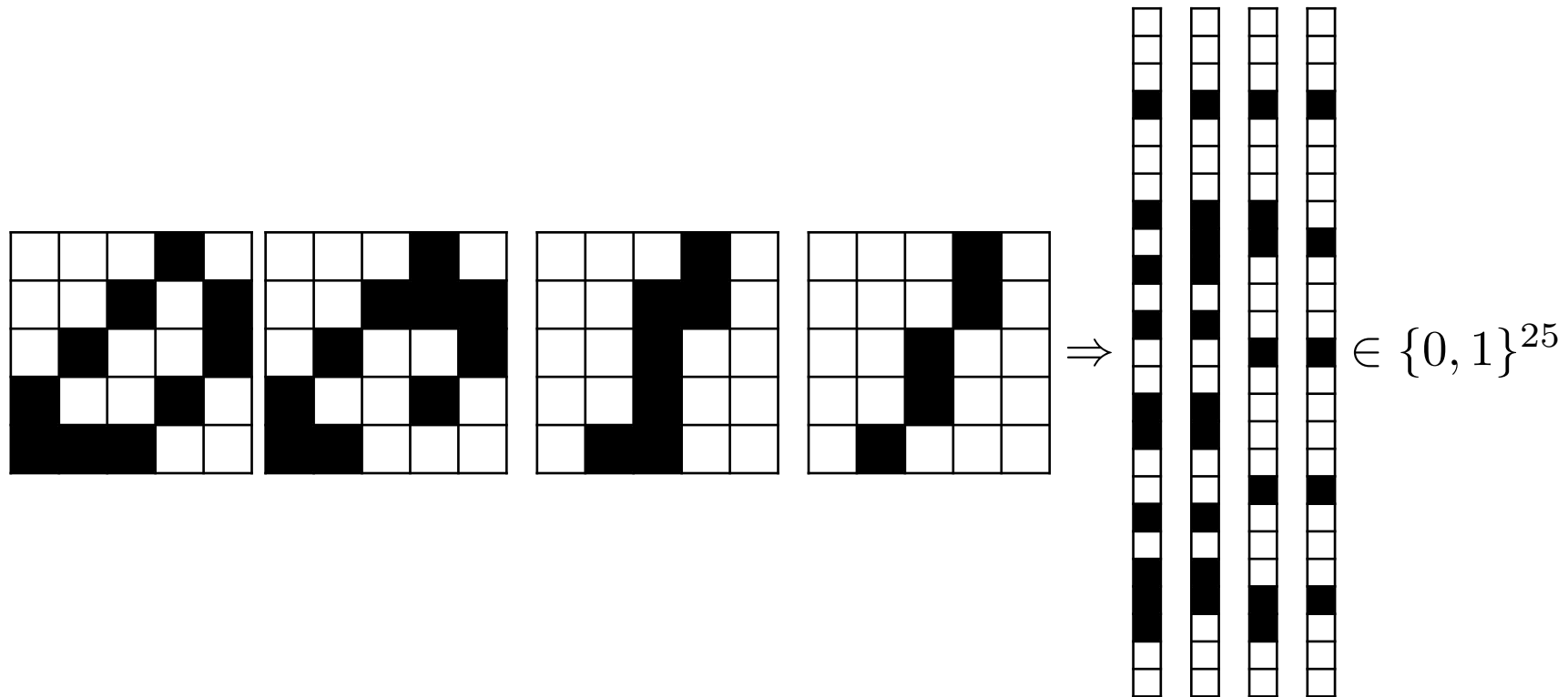
Estudiaremos tres tipos de distribución de probabilidad $p(\mathbf{x} \mid c)$ que son aplicables en función de los datos a modelizar:

- Bernoulli: datos binarios (vectores binarios)
- Multinomial: datos que son contadores (vectores de enteros positivos)
- Gaussiana: datos que son números reales (vectores 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 $p(\mathbf{x}|c)$

Multinomial: Motivación

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

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

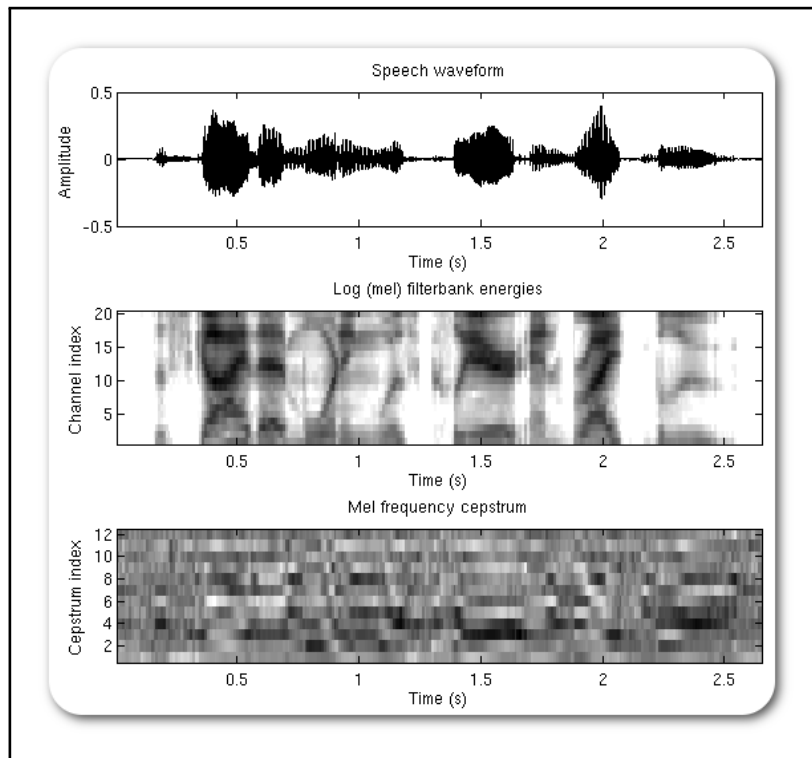
3 mensajes enviados a alt.atheism	⇒	0	0	0	windows
		4	11	3	god
		0	0	0	dod
		0	3	0	government
		2	0	1	writes
		14	7	15	people
		0	0	0	team
		0	0	0	bike
		0	0	0	game
		0	3	0	car
		0	1	0	article
		0	0	0	hockey
		0	0	0	rutgers
		0	0	0	encryption
		0	0	0	israel
		4	1	3	jesus
		0	0	0	clipper
		1	2	11	christians
		8	8	0	bible
		7	4	3	christian
3 mensajes enviados a comp.windows.x	⇒	17	17	9	windows
		0	0	0	god
		0	0	0	dod
		0	0	0	government
		0	1	0	writes
		4	3	5	people
		1	0	0	team
		0	0	0	bike
		0	1	0	game
		0	0	0	car
		3	0	8	article
		0	0	0	hockey
		0	0	0	rutgers
		0	0	0	encryption
		0	0	0	israel
		0	0	0	jesus
		0	0	0	clipper
		0	0	0	christians
		2	0	0	bible
		0	1	0	christian
3 mensajes enviados a rec.sport.hockey	⇒	0	0	0	windows
		0	0	0	god
		0	0	0	dod
		1	0	0	government
		0	0	2	writes
		8	0	7	people
		9	10	0	team
		0	0	0	bike
		3	13	10	game
		0	0	0	car
		0	0	0	article
		8	2	5	hockey
		0	0	0	rutgers
		0	0	0	encryption
		0	0	0	israel
		0	0	0	jesus
		0	0	0	clipper
		0	0	0	christians
		0	0	0	bible
		0	0	0	christian

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

Gaussiana: Motivación

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

Ejemplo: Señal acústica mediante vectores de coeficientes cepstrales



⇒

```
18637.9492 520.1318 -421.2418 1373.5566 53.3002
1169.5877 -212.9738 ... 266.1113 1038.8430
41.1315 284.1508 -51.6237 27.0556 177.5606
```

```
15014.2197 -261.0466 -803.4806 1420.2153
190.7374 1422.7503 -405.5645 ... -120.4282
108.4680 29.8098 69.2215 94.6452 20.1369 -27.4755
```

...

```
10495.8291 -551.9980 -471.0134 815.3850 325.7713
891.6923 -281.7178 ... -267.4365 -107.5793
85.9412 -184.7203 -33.2691 95.3320 -81.5344
```

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

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

- 1 Introducción y motivación ▷ 3
- 2 *Distribución de Bernoulli* ▷ 11
- 3 Distribución multinomial ▷ 20
- 4 Distribución Gaussiana ▷ 28

Definición: Bernoulli unidimensional

Sea $p \in [0, 1]$ y $q = 1 - p$. Pero no tiene que sumar 1.

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

x = número de éxitos

La f.d. de x es:

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

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

La distribución binomial está asociada a experimentos del siguiente tipo:

- Realizamos n veces cierto experimento en el que consideramos sólo la posibilidad de éxito o fracaso.

- La probabilidad de obtener éxito o fracaso siempre es la misma en cada ocasión. Siendo p la probabilidad de éxito y q la probabilidad de fracaso ($q=1-p$)

La distribución binomial se suele representar por $B(n, p)$.

Siendo n el número de intentos y p la probabilidad de éxito.

Definición: Bernoulli multidimensional

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

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

La f.d. de \mathbf{x} es:

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

f.d. = función de distribución

Clasificador Bernoulli

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

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

Por tanto:

$$\begin{aligned} c^*(\mathbf{x}) &= \operatorname{argmax}_{c=1, \dots, C} \log P(c) + \log p(\mathbf{x} | c) \\ &= \operatorname{argmax}_{c=1, \dots, C} \log P(c) + \log \prod_{d=1}^D p_{cd}^{x_d} (1 - p_{cd})^{(1-x_d)} \\ &= \operatorname{argmax}_{c=1, \dots, C} \log P(c) + \sum_{d=1}^D x_d \log p_{cd} + (1 - x_d) \log(1 - p_{cd}) \end{aligned}$$

Clasificador Bernoulli

Agrupando términos dependientes e independientes de x_d :

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

Reescribimos la expresión anterior como:

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

donde

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

Clasificador Bernoulli

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

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} g_c(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \sum_{d=1}^D w_{cd} x_d + w_{c0}$$

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

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \mathbf{w}_c^t \mathbf{x} + w_{c0}$$

donde

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

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:

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

Conjunto de parámetros a estimar Θ :

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

Por ***criterio de máxima verosimilitud*** (MV), se estima Θ como:

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

Para cada vector:

$$\hat{\mathbf{p}}_c = \frac{1}{N_c} \sum_{n: c_n=c} \mathbf{x}_n \quad c = 1, \dots, C$$

Suma de valores de
vector/nº componentes
de vector

Suavizado de la distribución Bernoulli

Problema: muchos criterios de entrenamiento (incluido 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 $(\mathbf{1}, c)$, $c = 1, \dots, C$.

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

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

Índice

- 1 Introducción y motivación ▷ 3
- 2 Distribución de Bernoulli ▷ 11
- 3 *Distribución multinomial* ▷ 20
- 4 Distribución Gaussiana ▷ 28

Definición: distribución multinomial

fronteras de decisión cuadráticas

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

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

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

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

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

Donde $w_n \in \mathcal{Y}, n = 1, \dots, N$

Número de secuencias distintas de longitud N :

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

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 de y_d en w_1^N
- $\mathbf{x} = (x_1, \dots, x_D)^t$: vector de ocurrencias (número de ocurrencias de cada y_d 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*:

$$\binom{N}{\mathbf{x}} = \binom{N}{x_1, \dots, x_D} = \frac{N!}{x_1! \cdots x_D!}$$

Definición: distribución multinomial

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

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

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

$p(\mathbf{x})$ es una f.d. (función de distribución) multinomial:

- D -dimensional
- Longitud $N = \sum_{d=1}^D x_d$
- Prototipo \mathbf{p}
(vector)

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

Clasificador multinomial

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

$$p(\mathbf{x} | c) \sim \text{Mult}_D(x_+, \mathbf{p}_c), \quad c = 1, \dots, C.$$

Por tanto:

$$\begin{aligned} c^*(\mathbf{x}) &= \operatorname{argmax}_{c=1, \dots, C} \log P(c) + \log p(x | c) \\ &= \operatorname{argmax}_{c=1, \dots, C} \log P(c) + \log \frac{x_+!}{x_1! \cdots x_D!} \prod_{d=1}^D p_{cd}^{x_d} \\ &= \operatorname{argmax}_{c=1, \dots, C} \log P(c) + \log \frac{x_+!}{x_1! \cdots x_D!} + \sum_{d=1}^D x_d \log p_{cd} \end{aligned}$$

Clasificador multinomial

Eliminando el término independiente de c :

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \log P(c) + \sum_{d=1}^D x_d \log p_{cd}$$

Expresando el sumatorio en forma de producto escalar:

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} (\log \mathbf{p}_c)^t \mathbf{x} + \log P(c)$$

En forma de clasificador lineal:

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} g_c(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \mathbf{w}_c^t \mathbf{x} + w_{c0}$$

Con:

$$\mathbf{w}_c = \log \mathbf{p}_c \quad w_{c0} = \log P(c)$$

Entrenamiento por máxima verosimilitud

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

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

Conjunto de parámetros a estimar Θ :

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

Por ***criterio de máxima verosimilitud*** (MV), se estima Θ como:

$$\hat{P}(c) = \frac{N_c}{N} \quad \hat{\mathbf{p}}_c = \frac{1}{\sum_{n: c_n=c} \sum_{d=1}^D x_{nd}} \sum_{n: c_n=c} \mathbf{x}_n \quad c = 1, \dots, C$$

Para cada vector:

Suma de valores de clase i de vector/suma de todos los valores de la clase i

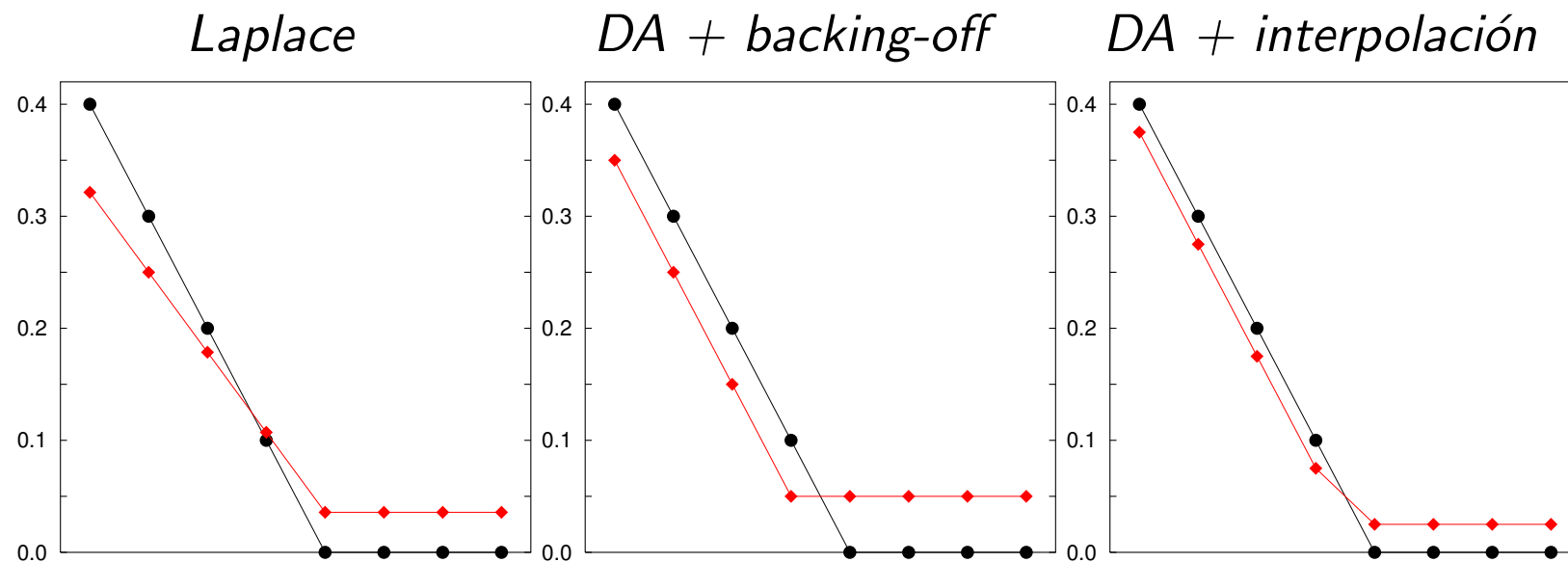
Suavizado de la distribución multinomial

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

Descuento Absoluto (DA): nunca puede ser menor que 0.

1. Descuenta una constante $\epsilon > 0$ (pequeña) a cada parámetro mayor que cero
2. Distribuir la probabilidad descontada según una *distribución generalizada*: Con $c = 0.2$
 - Entre todos los parámetros nulos (*backing-off*)
 - Entre todos los parámetros (*interpolación*)

$\hat{p} = (0.1 \ 0.7 \ 0.2)$ $\epsilon = 0.2$
 $\hat{p}' = (0 \ 0.5 \ 0)$ DESC = 0.5 \Rightarrow BACK OFF $\Rightarrow \hat{p} = (0.25 \ 0.5 \ 0.25)$
 \Rightarrow INTERP. $\Rightarrow \hat{p} = (\frac{0.5}{3} \ \frac{0.5+0.5}{3} \ \frac{0.5}{3})$



Índice

- 1 Introducción y motivación ▷ 3
- 2 Distribución de Bernoulli ▷ 11
- 3 Distribución multinomial ▷ 20
- 4 *Distribución Gaussiana* ▷ 28

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 $\mathbf{x} = (x_1, \dots, x_D)^t$ una variable aleatoria D -dimensional

Gaussiana estandarizada

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

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

Definición: distribución gaussiana multidimensional

Gaussiana general

Sean:

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

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

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

Clasificador gaussiano

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

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

Por tanto:

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

Clasificador gaussiano

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} -\frac{1}{2}\mathbf{x}^t \Sigma_c^{-1} \mathbf{x} + \boldsymbol{\mu}_c^t \Sigma_c^{-1} \mathbf{x} + \left(\log P(c) - \frac{1}{2} \log |\Sigma_c| - \frac{1}{2} \boldsymbol{\mu}_c^t \Sigma_c^{-1} \boldsymbol{\mu}_c \right)$$

con la dimensión

Clasificador *cuadrático* con \mathbf{x} :

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} g_c(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \mathbf{x}^t W_c \mathbf{x} + \mathbf{w}_c^t \mathbf{x} + w_{c0}$$

Con:

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

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

Clasificador gaussiano

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

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

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \mu_c^t \Sigma^{-1} \mathbf{x} + \left(\log P(c) - \frac{1}{2} \mu_c^t \Sigma^{-1} \mu_c \right)$$

El clasificador gaussiano es *lineal*: con el número de clases

$$c^*(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} g_c(\mathbf{x}) = \operatorname{argmax}_{c=1,\dots,C} \mathbf{w}_c^t \mathbf{x} + w_{c0}$$

Con:

$$\mathbf{w}_c = \Sigma^{-1} \mu_c \quad w_{c0} = \log P(c) - \frac{1}{2} \mu_c^t \Sigma^{-1} \mu_c$$

Entrenamiento por máxima verosimilitud

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

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

Conjunto de parámetros a estimar Θ :

- Probabilidades *a priori*: $P(1), \dots, P(C)$
 - Medias para cada clase: μ_1, \dots, μ_C
 - Matrices de covarianza para cada clase: $\Sigma_1, \dots, \Sigma_C$
- $N_c = \text{nº de muestras de clase}$
 $N = \text{nº de muestras}$

Por **criterio de máxima verosimilitud** (MV), se estima Θ como:

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

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

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

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*: $P(1), \dots, P(C)$
- Medias para cada clase: μ_1, \dots, μ_C
- Matriz de covarianza común: Σ

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

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

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:

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

Donde:

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

Flat smoothing

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

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

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