UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica etsinf

Tema 6. Distribución Multinomial

Percepción (PER)

Curso 2017/2018

Departamento de Sistemas Informáticos y Computación

Índice

Índice

Distribución 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* $\Rightarrow$

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* $\Rightarrow$

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* $\Rightarrow$

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 $p(\boldsymbol{x}|c)$

Índice

Definición: distribución multinomial

Sea una población $\mathcal{Y} = \{y_1, \ldots, y_n\}$ 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 $x_+ = N$.

Índice

Clasificador multinomial

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

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

Por tanto:

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

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

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

Clasificador multinomial

Eliminando el término independiente de c:

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

Expresando el sumatorio en forma de producto escalar:

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

En forma de clasificador lineal:

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

Con:

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

Índice

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:

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

Índice

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)*