UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Regressió Logística

Jorge Civera
Alfons Juan
Albert Sanchis

DSIC
Departament de Sistemes
Informàtics i Computació

Objectius formatius

- Explicar la distribució categòrica
- Representar la codificació one-hot
- Descriure el model probabilístic de classificació amb la funció softmax
- Descriure el model de regressió logística
- Descriure l'aprenentatge per màxima versemblança
- Aplicar descens per gradient en regressió logística

Índex

1 Distribució categòrica i codificació one-hot

- ***Variable categòrica:*** variable aleatòria que pren un valor d'un conjunt finit de categories (no ordenades)
- ***Exemples de variables categòriques: etiqueta de classe***, paraula d'un vocabulari, etc.
- ***Codificació one-hot:*** d'una variable categòrica i que pren un valor entre C possibles, $\{1,\dots,C\}$

$$\text{one-hot}(y) = \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_C \end{pmatrix} \in \{0,1\}^C \quad \text{amb} \quad \sum_c y_c = 1$$

- ***Exemple***: Codificació one-hot de la mostra $\boldsymbol{x_1} = (0,0)^t$ de la classe $c_1 = 1$ i $\boldsymbol{x_2} = (1,1)^t$ de la classe $c_2 = 2$:

$$\boldsymbol{y_1} = (1,0)^t$$
$$\boldsymbol{y_2} = (0,1)^t$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Distribució categòrica i codificació one-hot

- ***Distribució categòrica:*** distribució de probabilitats entre les C possibles categories d'una variable categòrica, que ve donada per un vector de paràmetres $\boldsymbol{\theta} \in [0,1]^C$ tal que $\sum_c \theta_c = 1$

$$p(y \mid \boldsymbol{\theta}) = \operatorname{Cat}(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{y_c}$$

- ***Convenció:*** $0^0 = 1$

- ***Exemple:***

$$\boldsymbol{\theta} = (0.5, 0.5, 0)^t,\ \operatorname{Cat}(\boldsymbol{y} = (1,0,0)^t \mid \boldsymbol{\theta}) = 0.5^1 0.5^0 0^0 = 0.5$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

2 Model probabilístic de classificació softmax

- ***Normalització probabilística de classificadors:*** siga G qualsevol classificador definit amb funcions discriminants generals $[g_1, \cdots, g_C]$, es pot definir un equivalent G' amb funcions discriminants normalitzades probabilísticament $[g'_1, \cdots, g'_C]$

$$\begin{aligned} c(\boldsymbol{x}) &= \operatorname*{argmax}_c \; g_c(\boldsymbol{x}) \\ &= \operatorname*{argmax}_c \; e^{g_c(\boldsymbol{x})} \text{ amb } h(z) = e^z \in \mathbb{R}^{\geq 0} \text{ estrictament creixent} \\ &= \operatorname*{argmax}_c \; \frac{e^{g_c(\boldsymbol{x})}}{\sum_{c'} e^{g_{c'}(\boldsymbol{x})}} \text{ amb } h(z) = kz,\; k \text{ constant positiva} \end{aligned}$$

  Per tant, $g'_c(\boldsymbol{x}) = \frac{e^{g_c(\boldsymbol{x})}}{\sum_{c'} e^{g_{c'}(\boldsymbol{x})}}$ defineix un classificador equivalent

  Aquesta transformació és coneguda com ***funció softmax***

- En aquest model probabilístic s'assumeix que els valores $g_c(\boldsymbol{x})$ són log-probabilitats no normalitzades denominats ***logits***

Model probabilístic de classificació softmax

- ***La funció softmax:*** transforma un vector de ***logits*** (log-probabilitats no normalitzades) $G \in \mathbb{R}^C$ en un de probabilitats $G' \in [0,1]^C$

$$G' = \mathcal{S}(G) = \left[\frac{e^{g_1}}{\sum_{c'} e^{g_{c'}}}, \dots, \frac{e^{g_C}}{\sum_{c'} e^{g_{c'}}}\right]$$

on es compleix

$$0 \leq \mathcal{S}(G)_c \leq 1 \quad \text{i} \quad \sum_c \mathcal{S}(G)_c = 1$$

3 Regressió logística

- ***Regressió logística:*** model amb softmax i funcions discriminants lineals (en notació homogènia)

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbf{W}) = \mathrm{Cat}(\boldsymbol{y} \mid \boldsymbol{\mu})$$

  on

$$\boldsymbol{\mu} = \mathcal{S}(\boldsymbol{a}), \quad \boldsymbol{a} = f(\boldsymbol{x}; \mathbf{W}) = \mathbf{W}^t \boldsymbol{x}, \quad \mathbf{W} \in \mathbb{R}^{D \times C} \quad \text{i} \quad \boldsymbol{x} \in \mathbb{R}^D$$

- No hi ha diferència amb els classificadors basats en funcions discriminants lineals, a excepció que ara predim les probabilitats de totes les classes

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Exemple

Siga un model de regressió logística en notació homogènia per a un problema de classificació en $C = 2$ classes i dades representades mitjançant vectors de dimensió $D = 2$

$$\boldsymbol{\mu} = \mathcal{S}(\boldsymbol{a}), \quad \boldsymbol{a} = f(\boldsymbol{x}; \mathbf{W}) = \mathbf{W}^t \boldsymbol{x} \quad \text{amb}$$

$$\mathbf{W}^t = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \end{pmatrix} \quad \text{i} \quad \boldsymbol{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$$

Calcula amb quina probabilitat $\boldsymbol{x} = (1, 0, 0)^t$ i $\boldsymbol{x} = (1, 1, 1)^t$ pertanyen a cada classe:

$\boldsymbol{x}^t$	$\boldsymbol{a}^t$	$\mu_1 = \mathcal{S}(\boldsymbol{a})_1$	$\mu_2 = \mathcal{S}(\boldsymbol{a})_2$
$(1, 0, 0)$	$(1, -1)$	$\frac{e^1}{e^1+e^{-1}} = 0.8808$	$\frac{e^{-1}}{e^1+e^{-1}} = 0.1192$
$(1, 1, 1)$	$(-1, 1)$	$\frac{e^{-1}}{e^{-1}+e^1} = 0.1192$	$\frac{e^1}{e^{-1}+e^1} = 0.8808$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

4 Aprenentatge per màxima versemblança

- ***Objectiu:*** establir un criteri per a aprendre $\mathbf{W}$ a partir d'un conjunt de dades d'entrenament, $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$
- ***Log-versemblança (condicional):*** log-probabilitat de $\mathcal{D}$ interpretada com a funció de $\mathbf{W}$

$$\begin{aligned}
\mathrm{LL}(\mathbf{W}) &= \log p(\mathcal{D} \mid \mathbf{W}) = \log \prod_{n=1}^{N} p(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \mathbf{W}) \\
&= \sum_{n=1}^{N} \log \mathrm{Cat}(\boldsymbol{y}_n \mid \boldsymbol{\mu}_n) \quad \text{amb} \quad \boldsymbol{\mu}_n = \mathcal{S}(\boldsymbol{a}_n) \quad \text{y} \quad \boldsymbol{a}_n = \mathbf{W}^t \boldsymbol{x}_n \\
&= \sum_{n=1}^{N} \log \prod_{c=1}^{C} \mu_{nc}^{y_{nc}} = \sum_{n=1}^{N} \sum_{c=1}^{C} y_{nc} \log \mu_{nc}
\end{aligned}$$

- ***Aprenentatge per màxima versemblança:*** triem una $\mathbf{W}$ que atorgue màxima probabilitat a $\mathcal{D}$

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmax}} \ \mathrm{LL}(\mathbf{W})$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Exemple

Calcula la log-versemblança de $\mathbf{W}^t = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \end{pmatrix}$ amb dues dades $\mathcal{D} = \{((1,0,0)^t,(1,0)^t),((1,1,1)^t,(0,1)^t)\}$

$$\begin{aligned}
\mathrm{LL}(\mathbf{W}) = \log p(\mathcal{D} \mid \mathbf{W}) &= \sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc}\log\mu_{nc} \\
&= y_{11}\log\mu_{11} + y_{12}\log\mu_{12} + y_{21}\log\mu_{21} + y_{22}\log\mu_{22} \\
&= \log\mu_{11} + \log\mu_{22} \\
&= \log 0.8808 + \log 0.8808 = -0.1269 - 0.1269 = -0.2538
\end{aligned}$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Plantejament com a problema de minimització

- ***Neg-log-versemblança:*** log-versemblança amb el signe canviat i normalitzada pel nombre de dades

$$\mathrm{NLL}(\mathbf{W}) = -\frac{1}{N}\mathrm{LL}(\mathbf{W}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc}\log\mu_{nc}$$

- ***Exemple:*** neg-log-versemblança de l'exemple anterior

$$\mathrm{NLL}(\mathbf{W}) = -\frac{1}{2}\mathrm{LL}(\mathbf{W}) = 0.1269$$

- ***Aprenentatge per mínima NLL:*** aprenentatge per màxima versemblança plantejat com un problema de minimització

$$\mathbf{W}^* = \operatorname*{argmin}_{\mathbf{W}}\ \mathrm{NLL}(\mathbf{W})$$

Exemple

Suposem que hem de triar per mínima NLL entre:

$$\mathbf{W}^t = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \end{pmatrix} \qquad \text{y} \qquad \tilde{\mathbf{W}}^t = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix}$$

amb les dades $\mathcal{D} = \{((1,0,0)^t, (1,0)^t), ((1,1,1)^t, (0,1)^t)\}$

Segons l'exemple anterior, la NLL de $\mathbf{W}$ és 0.1269 i la de $\tilde{\mathbf{W}}$:

$$\text{NLL}(\tilde{\mathbf{W}}) = -\frac{1}{2}(\log \tilde{\mu}_{11} + \log \tilde{\mu}_{22}) = -\frac{1}{2}(\log \frac{e^{-1}}{e^{-1} + e^{1}} + \log \frac{e^{-1}}{e^{-1} + e^{1}}) = 2.1269$$

Per tant, triaríem $\mathbf{W}$ ja que la seua NLL és menor que la de $\tilde{\mathbf{W}}$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

5 Aprenentatge amb descens per gradient

- ***Descens per gradient:*** algorisme iteratiu per a minimitzar una funció $\mathcal{L}(\boldsymbol{\theta})$ a partir d'un valor inicial dels paràmetres $\boldsymbol{\theta}_0$ donat

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \eta_i \boldsymbol{\nabla} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_i}$$

- ***Factor d'aprenentatge:*** $\eta_i > 0$ juga el mateix paper que en Perceptró; podem triar un valor xicotet constant, $\eta_i = \eta$

- ***Direcció de descens més pronunciada:*** $-\boldsymbol{\nabla} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_i}$ és el neg-gradient de la funció avaluada en $\boldsymbol{\theta}_i$

- ***Convergència:*** si η no és molt gran i la funció és convexa (amb forma de bol), convergeix a un mínim (global)

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Descens per gradient en regressió logística

- La ***NLL*** és una funció convexa

- ***Gradient de la NLL:*** farem ús del següent resultat

$$\begin{pmatrix} \frac{\partial\,\mathrm{NLL}}{\partial W_{11}} & \cdots & \frac{\partial\,\mathrm{NLL}}{\partial W_{1C}} \\ \vdots & \ddots & \vdots \\ \frac{\partial\,\mathrm{NLL}}{\partial W_{D1}} & \cdots & \frac{\partial\,\mathrm{NLL}}{\partial W_{DC}} \end{pmatrix} = \frac{\partial\,\mathrm{NLL}}{\partial \mathbf{W}^t} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n(\boldsymbol{\mu}_n - \boldsymbol{y}_n)^t$$

- ***Descens per gradient aplicat a regressió logística:***

$$\mathbf{W}_0 = \mathbf{0}; \quad \mathbf{W}_{i+1} = \mathbf{W}_i - \eta_i \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n(\boldsymbol{\mu}_n - \boldsymbol{y}_n)^t$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Exemple

Siga un model de regressió logística en notació homogènia per a un problema de classificació en $C = 2$ classes i dades de dimensió $D = 2$, actualitza el valor de $\mathbf{W}$ aplicant descens per gradient amb $\eta = 0.1$, matriu de pesos inicials nuls i conjunt d'entrenament $\mathcal{D} = \{\boldsymbol{x_1} = (1, 0, 0)^t, \boldsymbol{y_1} = (1, 0)^t)\}$

$$\boldsymbol{a} = \mathbf{W}^t \boldsymbol{x} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\boldsymbol{\mu} = S(\boldsymbol{a}) = \left(\frac{e^0}{e^0 + e^0}, \frac{e^0}{e^0 + e^0} \right)^t = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

$$\mathbf{W} = \mathbf{W} - \eta\, \boldsymbol{x}(\boldsymbol{\mu} - \boldsymbol{y})^t$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - 0.1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} ((0.5, 0.5) - (1, 0))$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - 0.1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (-0.5, 0.5)$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0.1 \\ 0 \\ 0 \end{pmatrix} (-0.5, 0.5)$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} -0.05 & 0.05 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.05 & -0.05 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

6 Conclusions

Hem vist:

- La distribució categòrica i la codificació one-hot
- El model probabilístic de classificació amb la funció softmax i, en particular, el model de regressió logística
- El mètode d'aprenentatge per màxima versemblança en regressió logística aplicant descens per gradient