UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Regresión Logística

Jorge Civera
Alfons Juan
Albert Sanchis

Departamento de Sistemas Informáticos y Computación

Objetivos formativos

- Explicar la distribución categórica
- Representar la codificación one-hot
- Describir el modelo probabilístico de clasificación con la función softmax
- Describir el modelo de regresión logística
- Describir el aprendizaje por máxima verosimilitud
- Aplicar descenso por gradiente en regresión logística

Índice

1. Distribución categórica y codificación one-hot

- ***Variable categórica:*** variable aleatoria que toma un valor de un conjunto finito de categorías (no ordenadas)
- ***Ejemplos de variables categóricas: etiqueta de clase***, palabra de un vocabulario, etc.
- ***Codificación one-hot:*** de una variable categórica y que toma un valor entre C posibles, $\{1, \dots, C\}$

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

- ***Ejemplo***: Codificación one-hot de la muestra $\boldsymbol{x_1} = (0, 0)^t$ de la clase $c_1 = 1$ y $\boldsymbol{x_2} = (1, 1)^t$ de la clase $c_2 = 2$:

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

Codificación one-hot y distribución categórica

- ***Distribución categórica:*** distribución de probabilidades entre las C posibles categorías de una variable categórica, que viene dada por un vector de parámetros $\boldsymbol{\theta} \in [0,1]^C$ tal que $\sum_c \theta_c = 1$

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

- ***Convención:*** $0^0 = 1$

- ***Ejemplo:***

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

2. Modelo probabilístico de clasificación softmax

- ***Normalización probabilística de clasificadores:*** sea G cualquier clasificador definido con funciones discriminantes generales $[g_1, \cdots, g_C]$, se puede definir uno equivalente G' con funciones discriminantes normalizadas probabilísticamente $[g'_1, \cdots, g'_C]$

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

  Por tanto, $g'_c(\boldsymbol{x}) = \frac{e^{g_c(\boldsymbol{x})}}{\sum_{c'} e^{g_{c'}(\boldsymbol{x})}}$ define un clasificador equivalente

  Esta transformación es conocida como ***función softmax***

- En este modelo probabilístico se asume que los valores $g_c(\boldsymbol{x})$ son log-probabilidades no normalizadas denominados ***logits***

Modelo probabilístico de clasificación softmax

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

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

donde se cumple

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

3. Regresión logística

- ***Regresión logística:*** modelo con softmax y funciones discriminantes lineales (en notación homogénea)

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

donde

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

- No hay diferencia con los clasificadores basados en funciones discriminantes lineales, a excepción de que ahora predecimos las probabilidades de todas las clases

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Ejemplo

Sea un modelo de regresión logística en notación homogénea para un problema de clasificación en $C = 2$ clases y datos representados mediante vectores de dimensión $D = 2$

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

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

Calcula la probabilidad de que $\boldsymbol{x} = (1, 0, 0)^t$ y $\boldsymbol{x} = (1, 1, 1)^t$ pertenezcan a cada clase:

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

4. Aprendizaje por máxima verosimilitud

- ***Objetivo:*** establecer un criterio para aprender $\mathbf{W}$ a partir de un conjunto de datos de entrenamiento, $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$
- ***Log-verosimilitud (condicional):*** log-probabilidad de $\mathcal{D}$ interpretada como función de $\mathbf{W}$

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

- ***Aprendizaje por máxima verosimilitud:*** elegimos una $\mathbf{W}$ que otorgue máxima probabilidad a $\mathcal{D}$

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

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Ejemplo

Calcula la log-verosimilitud de $\mathbf{W}^t = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \end{pmatrix}$ con dos datos $\mathcal{D} = \{((1,0,0)^t, (1,0)^t), ((1,1,1)^t, (0,1)^t)\}$

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

Planteamiento como problema de minimización

- ***Neg-log-verosimilitud:*** log-verosimilitud con el signo cambiado y normalizada por el número de datos

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

- ***Ejemplo:*** neg-log-verosimilitud del ejemplo anterior

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

- ***Aprendizaje por mínima NLL:*** aprendizaje por máxima verosimilitud planteado como un problema de minimización

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

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Ejemplo

Supongamos que tenemos que elegir por mínima NLL entre:

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

con los datos $\mathcal{D} = \{((1,0,0)^t, (1,0)^t), ((1,1,1)^t, (0,1)^t)\}$

Según el ejemplo anterior, la NLL de $\mathbf{W}$ es 0.1269 y la de $\tilde{\mathbf{W}}$:

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

Por tanto, elegiríamos $\mathbf{W}$ ya que su NLL es menor que la de $\tilde{\mathbf{W}}$

5. Aprendizaje con descenso por gradiente

- ***Descenso por gradiente:*** algoritmo iterativo para minimizar una función $\mathcal{L}(\boldsymbol{\theta})$ a partir de un valor inicial de los parámetros $\boldsymbol{\theta}_0$ dado

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

- ***Factor de aprendizaje:*** $\eta_i > 0$ juega el mismo papel que en Perceptrón; podemos elegir un valor pequeño constante, $\eta_i = \eta$

- ***Dirección de descenso más pronunciada:*** $-\boldsymbol{\nabla} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_i}$ es el neg-gradiente de la función evaluada en $\boldsymbol{\theta}_i$

- ***Convergencia:*** si η no es muy grande y la función es convexa (con forma de bol), converge a un mínimo (global)

Descenso por gradiente en regresión logística

- La ***NLL*** es una función convexa
- ***Gradiente de la NLL:*** haremos uso del siguiente resultado

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

- ***Descenso por gradiente aplicado a regresión 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

Ejemplo

Sea un modelo de regresión logística en notación homogénea para un problema de clasificación en $C = 2$ clases y datos de dimensión $D = 2$, actualiza el valor de $\mathbf{W}$ aplicando descenso por gradiente con $\eta = 0.1$, matriz de pesos iniciales nulos y conjunto de entrenamiento $\mathcal{D} = \{\boldsymbol{x_1} = (1, 0, 0)^t, \boldsymbol{y_1} = (1, 0)^t)\}$

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

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

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

$$
\begin{aligned}
\mathbf{W} &= \mathbf{W} - \eta\, \boldsymbol{x}(\boldsymbol{\mu} - \boldsymbol{y})^t \\
&= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - 0.1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} ((0.5, 0.5) - (1, 0)) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - 0.1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (-0.5, 0.5) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} 0.1 \\ 0 \\ 0 \end{pmatrix} (-0.5, 0.5) \\
&= \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} -0.05 & 0.05 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.05 & -0.05 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}
\end{aligned}
$$

6. Conclusiones

Hemos visto:

- La distribución categórica y la codificación one-hot
- El modelo probabilístico de clasificación con la función softmax y, en particular, el modelo de regresión logística
- El método de aprendizaje por máxima verosimilitud en regresión logística aplicando descenso por gradiente

UNIVERSITAT POLITÈCNICA DE VALÈNCIA