Clase 07: Modelos Lineales

Responsable: Jordi Legorreta Montaño

EST-25134, Primavera 2021
Dr. Alfredo Garbuno Iñigo
Febrero 04, 2021

1. Modelos Lineales

Antes : modelo de clasificación por semiespacios

(También conocido como discriminación lineal).
Algoritmo para entrenarlo: perceptrones.
NOTA: No confundir modelo con algoritmo.
$\hat{y} \in \{0, 1\}$
$y\langle w, x\rangle \geq 1 \quad \leftarrow \quad$ margen de clasificación
$y\langle w, x\rangle > 0$
$w^{(t+1)} = w^{(t)} + y^{(i)}x^{(i)}$

El algoritmo, lo que busca, es que después de cada iteración, los casos mal clasificados sean bien clasificados.

$$y^{(i)}\langle w^{(t+1)}, x^{(i)}\rangle \geq 0$$

Dado el criterio anterior, ¿cuál será la función de pérdida?

1.1. Modelo de Regresión

Sirven para describir una relación entre un conjunto de características $x \in \mathcal{X} \subseteq \mathbb{R}^p$, y la respuesta (en este caso) será $y \in \mathbb{R}$.

x : variables independientes, explicativas o exploratorias.

y : variable dependiente o variable objeto.

$$y = f(x) \quad \leftarrow \quad \text{relación entre } \mathcal{X} \text{ e } \mathcal{Y}$$

Queremos encontrar $h : \mathbb{R}^d \to \mathbb{R}$, la mejor aproximación a dicha relación.

$$\mathcal{H}_{reg} = L_p = \{h_{w,b} : w \in \mathbb{R}^p, \;\; b \in \mathbb{R}, \;\; h_{w,b}(x) = \langle x, w\rangle + b\}$$

La "mejor" $h \in \mathcal{H}_{reg}$ es la que minimice el error:

$l(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
$l(w, x, y) = \frac{1}{2}(\langle w, x\rangle - y)^2, \quad w \in \mathbb{R}^{p+1}$

$L_s(h) = L_s(w) = \frac{1}{2m}\sum_{i=1}^{m}(\langle w, x^{(i)}\rangle - y^{(i)})^2$

Con la función de perdida anterior, el principio de minimización de riesgo empírico nos dice que:

$$ERM \rightarrow h^* = h^*_w \mid w^* = \text{argmin } L_s(w)$$

$$L_s(w) = \tfrac{1}{2m}\|\mathbb{X}w - y\|^2$$

$$\mathbb{X} = \begin{bmatrix} -\mathrm{X}^{(1)T}- \\ -\mathrm{X}^{(2)T}- \\ \vdots \\ -\mathrm{X}^{(m)T}- \end{bmatrix} \in \mathbb{R}^{mxp}, \quad y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{bmatrix} \in \mathbb{R}^m$$

Ventajas de la función $L_s(w)$:

- Función diferenciable, por lo que podemos calcular la función gradiente parametrizada por w.

$$\nabla_w L_s(w) = \nabla_w(\frac{1}{2m}\|\mathbb{X}w - y\|^2) = \frac{2}{2m}(\mathbb{X}w - y)^T\mathbb{X}$$

$$\nabla_w L_s(w) = 0 \iff \frac{1}{m}(\mathbb{X}w - y)^T\mathbb{X} = 0 \iff w^* = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y$$

$$w^* = A^{-1}b \iff Aw^* = b$$

¿Qué condiciones debe satisfacer la matriz A?

- Que sea invertible
- Que sea positiva definida

  (i) $m > p$

  (ii) No haya atributos linealmente dependientes

- $A^\dagger$ es la pseudoinversa

$$A = vDv^T, \text{ donde } vv^T = v_T v = \mathbb{I}_p$$
$$A^\dagger = vD^\dagger v^T$$
$$D = diag(d_1, d_2, ..., d_p)$$
$$D^\dagger = diag(1/d_1, ..., 1/d_k, 0, ..., 0)$$

$ERM \rightarrow$ solución: $w^* = A^\dagger x^T y$

Concluyendo: este problema de regresión tiene una solución y se puede encontrar de manera analítica.

1.2. Conectar ERM con estimadores de máxima verosimilitud

$y = h_w(x) + \varepsilon$: consideramos cierta desviación

$\varepsilon \sim N(0, \sigma^2)$
ε: error de estimación, discrepancia que existe entre el modelo lineal y los datos.
$y = \hat{y} + \varepsilon$

Definición : $\pi(y \mid x, w) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{1}{2\sigma^2}(y - \langle w,x \rangle)^2}$, donde $y \sim N(\langle w, x \rangle, \sigma^2)$

$log\pi(y \mid x, w) = -\frac{1}{2\sigma^2}(y - \langle w, x \rangle)^2 + c$
$l(\hat{y}, y) = -log\pi(y \mid x, w)$
$L_s(w) = \frac{1}{m}\sum_{i=1}^{m} l(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{m} log\pi(y \mid x, w)$
$-\frac{1}{m}\sum_{i=1}^{m} log\pi(y \mid x, w) = -l_m(w)$
$\mathcal{L}_m(w) = \Pi_{i=1}^{m}\pi(y^{(i)} \mid x^{(i)}, w) \leftarrow verosimilitud$
$l_m(w) = log(\mathcal{L}_m(w)) = \frac{1}{m}\sum_{i=1}^{m} log\pi(y^{(i)} \mid x^{(i)}, w)$
$\frac{1}{m}\sum_{i=1}^{m} log\pi(y^{(i)} \mid x^{(i)}, w) \leftarrow log - verosimilitud$
$w^*_{ERM} = \text{argmin } L_s(w) = \text{argmax } l_m(w) = \hat{w}^{MLE}$

En el contexto de regresión lineal con errores gaussianos, el principio de minimización de riesgo empírico es equivalente a estimar un parámetro como si fuera un problema estadístico de máxima verosimilitud.

1.3. Conexión de ERM con Teoría de la Información

Teoría de la información : Codificar, decodificar, transmitir y manipular paquetes de información en canales de comunicación[1].

Lo que más nos interesa en Teoría de la Información es:

(I) Entropía: $H(P) = \sum_j -p_j log(p_j)$, la información con la que podemos codificar la realización de un evento aleatorio a través de una distribución P.
"Sorpresas": los eventos de probabilidad muy bajos, cuando ocurren, nos sorprenden, porque no los esperamos.

$$log(\frac{1}{p_j}) = -log(p_j)$$

De alguna manera la entropía es el valor esperado de la sorpresa.

(II) Entropía cruzada: $H(P, Q) = \sum_j -p_j log(q_j)$, los eventos aleatorios se están generando con respecto a una distribución P y los estamos codificando utilizando una distribución Q.
Por la desigualdad de Jensen, tenemos que: $H(P) = H(P, P) \leq H(P, Q)$

(III) Entropía cruzada relativa: $D_{kl}(P\|Q) = H(P, Q) - H(P, P) \geq 0$, donde D_{kl} se conoce como "Divergencia"de Kullback-Leibler

Entonces: $L_s(h) = \sum_{i=1}^{m} -log\pi(y^{(i)} \mid x^{(i)}, w)\frac{1}{m} = H(D^m, D_w^m)$, entropía cruzada entre la distribución empírica que generan los datos D^m y una distribución de m datos utilizando una codificación basada en w.

Por lo tanto, $w^* = \text{argmin } L_s(h) = \text{argmin } H(D^m, D_w^m)$

1.4. Familias Polinomiales

$P(x) = a_0 + a_1x + ... + a_px^p = \langle a, \Phi(x)\rangle$, donde $a = (a_0, a_1, ..., a_p)^T \in \mathbb{R}^{p+1}$ y $\Phi(x) = (1, x, ..., x^p) \in \mathbb{R}^{p+1}$.

$a^* \leftarrow$ utilizamos ERM con pérdida cuadrática.

Ya no son sólo rectas, sino que ahora trabajamos con polinomios.

2. Modelo de Regresión en clasificación (binaria)

$h : \mathbb{R}^p \to [0, 1]$, donde $h(x)$ es la probabilidad de obtener $y = 1$.
$\sigma(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$, donde $\sigma(x)$ es la función sigmoide, la cual nos ayuda a comprimir la recta real, de tal forma que este problema de usar modelos de regresión para clasificación le llamamos:

Regresión logística: $\mathcal{H}_{log} = \sigma_w L_p\{\sigma_w : w \in \mathbb{R}^p,\ \ \sigma_w(x) = \sigma(\langle x, w\rangle)\}$

- Si $\langle w, x\rangle$ es muy grande, entonces $\sigma_w(x) \approx 1$
- Si $\langle w, x\rangle$ es muy pequeño (en términos negativos), entonces $\sigma_w(x) \approx 0$
- Si $\langle w, x\rangle = 0$, entonces $\sigma_w(x) = \frac{1}{2}$

¿Qué función de pérdida utilizar?

$\sigma_w(x) = \mathbb{P}(y = 1 \mid x, w) \implies y \sim Bernoulli(\sigma_w(x))$
Verosimilitud de una Bernoulli: $(\sigma_w(x))^y(1 - \sigma_w(x))^{1-y}$
log-verosimilitud: $ylog(\sigma_w) + (1 - y)log(1 - \sigma_w)$

Si $\sigma_w(x) = \frac{e^{w^T x}}{1+e^{w^T x}}$, entonces:

log-verosimilitud: $-log(1 + e^{-y\langle w,x\rangle})$, donde $y \in \{1, -1\}$
$l(\hat{y}, y) = log(1 + e^{-\hat{y}\langle w,x\rangle})$, donde $\hat{y} = \sigma_w(x)$
$L_s(w) = \frac{1}{m}\sum_{i=1}^{m} log(1 + e^{-y^{(i)}\langle w,x^{(i)}\rangle})$

Agradecimientos

Este *t*emplate se ha adaptado y traducido del provisto en la clase ACM 204 (Otoño 2017) por el profesor Joel Tropp.

Referencias

[1] Mackay 2003: Information Theory, Inference and Learning.