CLASIFICACIÓN DE OBJETOS

Regresión Logística – Aprendizaje (III)

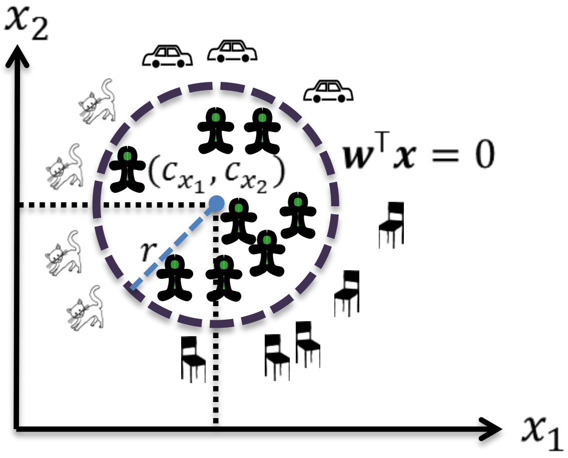
Antonio M. López

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

Universitat Autònoma de Barcelona

- Fronteras no lineales:

$$(x_1 - c_{x_1})^2 + (x_2 - c_{x_2})^2 = r^2$$



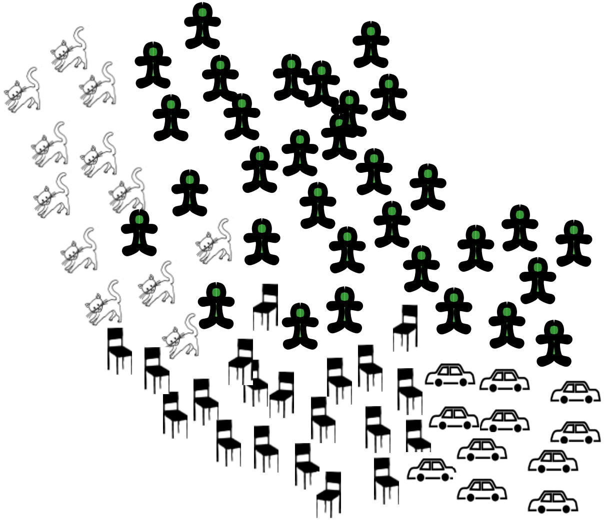
$$\boldsymbol{w} = (w_0, w_1, w_2, w_3, w_4, w_5)^\top$$

$$\boldsymbol{x} = (1, x_1, x_2, {x_1}^2, x_1 x_2, {x_2}^2)^\top$$

$$\boldsymbol{w} = \left((c_{x_1}{}^2 + c_{x_2}{}^2 - r^2), -2c_{x_1}, -2c_{x_2}, 1, 0, 1\right)^\top$$

- <u>Sobreajuste (*overfitting*)</u>:

  - Las muestras tienen ruido o son incompletas.
  - El modelo de frontera es una aproximación.



- Aunque se consiga un clasificador sin error en la clasificación de las muestras de entrenamiento, esto no garantiza la correcta clasificación de muestras nuevas (capacidad de Generalización).

- El problema tiende a manifestarse especialmente cuando el número de parámetros del modelo es muy alto (dimensión alta del descriptor) respecto al número de muestras de entrenamiento.

- Aumentar el número de muestras de entrenamiento.
- Bajar dimensión del descriptor (selección de características).
- Selección del modelo.
- **Regularización del modelo**.

- Regularización de la Regresión Logística:

$$\mathcal{J}_R(\boldsymbol{w}) \leftarrow \mathcal{J}(\boldsymbol{w}) + \frac{\lambda}{2M}\sum_{i=1}^{n} {w_i}^2$$

Repetir (en paralelo)

$$ {w_0}^{(k)} \longleftarrow {w_0}^{(k-1)} - \frac{\alpha}{M}\sum_{j=1}^{M}\left(h_{\boldsymbol{w}}(\boldsymbol{x}^j) - y^j\right){x_0}^j$$

$\forall i \in \{1, \dots, n\}$

$$ {w_i}^{(k)} \longleftarrow {w_i}^{(k-1)} - \frac{\alpha}{M}\left[\sum_{j=1}^{M}\left(h_{\boldsymbol{w}}(\boldsymbol{x}^j) - y^j\right){x_i}^j + \lambda\, w_i\right]$$

Hasta converger /* p.e. monitorizar $\mathcal{J}_R(\boldsymbol{w})$ */

$\lambda \in \mathbb{R}^+$, determina el grado de penalización.

UAB Universitat Autònoma de Barcelona

- Punto de vista de la Probabilidad:

$$P(Y=1|X=\boldsymbol{x};\boldsymbol{w}) = h_{\boldsymbol{w}}(\boldsymbol{x}) = Logistic\,(\boldsymbol{w}^{\top}\boldsymbol{x})$$

$$P(Y=0|X=\boldsymbol{x};\boldsymbol{w}) = 1 - h_{\boldsymbol{w}}(\boldsymbol{x})$$

  - Podemos usar un algoritmo tipo máxima verosimilitud (*likelihood;* ML) para aprender $\boldsymbol{w}$.
  - En particular, queremos obtener el $\boldsymbol{w}^*$ que maximice la verosimilitud de que el conjunto de muestras $\mathcal{S} = \{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^M, y^M)\}$ fuese generado según $P(Y|X;\boldsymbol{w})$
    ➔ Máxima Verosimilitud Condicional (*Maximum Conditional Likelihood*).
  - Esto se reduce a resolver el siguiente problema:

$$\mathfrak{L}(\boldsymbol{w}) = \prod_{j=1}^{M} P(Y=y^j|X=\boldsymbol{x}^j;\boldsymbol{w})$$

$$\boldsymbol{w}^* \longleftarrow \underset{\boldsymbol{w}\,\in\,\mathbb{R}^{n+1}}{argmax}\,\mathfrak{L}(\boldsymbol{w})$$

Muestras *iid*:
independientes
idénticamente distribuidas

Universitat Autònoma de Barcelona

- Alternativamente:

$$\mathfrak{L}(\boldsymbol{w}) = \ln \prod_{j=1}^{M} \mathrm{P}(\mathrm{Y}{=}y^j|\mathrm{X}{=}\boldsymbol{x}^j;\boldsymbol{w}) = \sum_{j=1}^{M} \ln \mathrm{P}(\mathrm{Y}{=}y^j|\mathrm{X}{=}\boldsymbol{x}^j;\boldsymbol{w})$$

$$\boldsymbol{w}^* \longleftarrow \underset{\boldsymbol{w}\,\in\,\mathbb{R}^{n+1}}{argmax}\,\mathfrak{L}(\boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = -\sum_{j=1}^{M} \ln \mathrm{P}(\mathrm{Y}{=}y^j|\mathrm{X}{=}\boldsymbol{x}^j;\boldsymbol{w}) = -\sum_{j=1}^{M} y^j\, ln\big(h_{\boldsymbol{w}}(\boldsymbol{x}^j)\big) \;+\; (1-y^j)\, ln\Big(1-h_{\boldsymbol{w}}(\boldsymbol{x}^j)\Big)$$

$$\boldsymbol{w}^* \longleftarrow \underset{\boldsymbol{w}\,\in\,\mathbb{R}^{n+1}}{argmin}\,\mathcal{J}(\boldsymbol{w})$$

- Con regularización:

$$\mathfrak{L}(\boldsymbol{w}) = \sum_{j=1}^{M} \ln \mathrm{P}(\mathrm{Y}{=}y^j | \mathrm{X}{=}\boldsymbol{x}^j; \boldsymbol{w}) - \frac{\lambda}{2} ||\boldsymbol{w}||^2$$

$$\boldsymbol{w}^* \longleftarrow \underset{\boldsymbol{w} \in \mathbb{R}^{n+1}}{argmax}\, \mathfrak{L}(\boldsymbol{w})$$

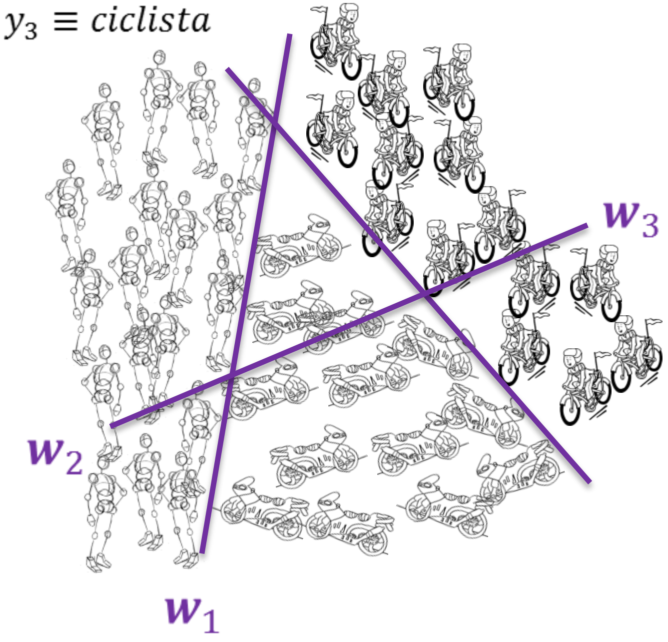
- Es interesante recordar que en una estimación de $\boldsymbol{w}$ tipo MAP (máximo a posteriori), la expresión a optimizar en este caso tendría la forma:

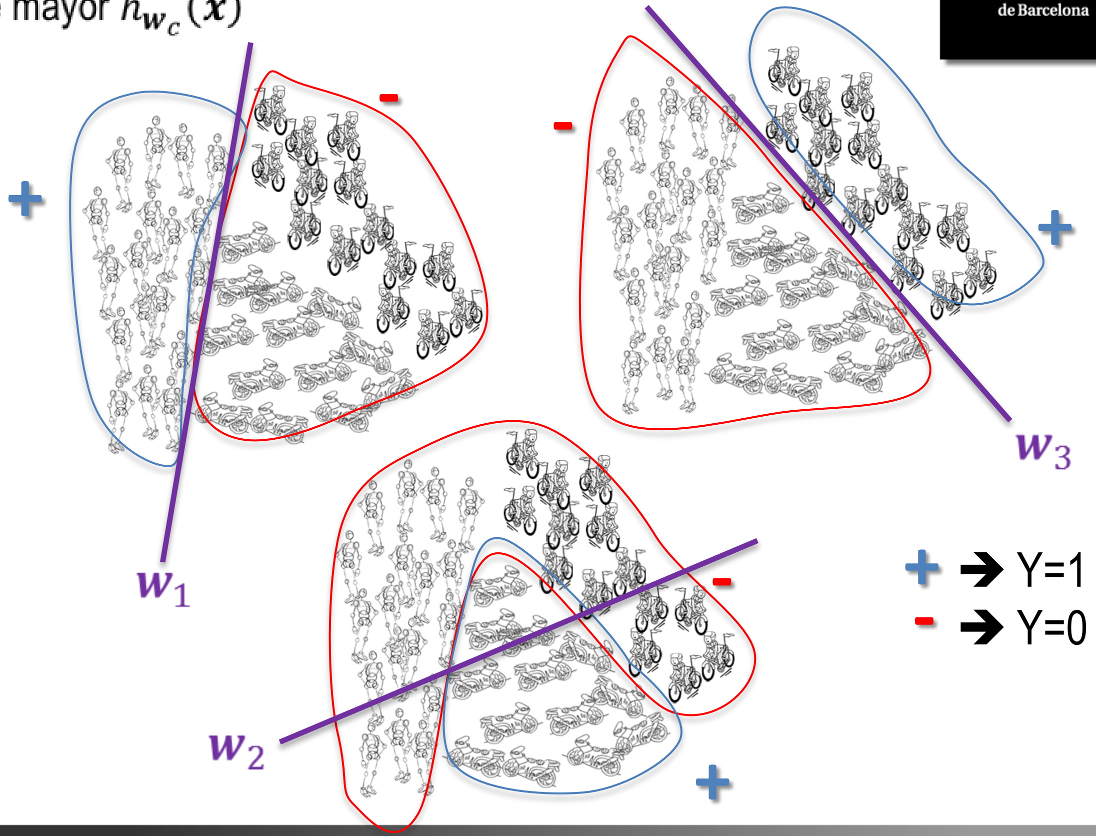
$$\sum_{j=1}^{M} \ln \mathrm{P}(\mathrm{Y}{=}y^j | \mathrm{X}{=}\boldsymbol{x}^j; \boldsymbol{w}) + \ln \mathrm{P}(\boldsymbol{w})$$

y si P($\boldsymbol{w}$) es una Gaussiana centrada en cero con varianza proporcional a 1/λ, entonces $\ln$ P($\boldsymbol{w}$) es proporcional a $\lambda||\boldsymbol{w}||^2$ ➔ la versión regularizada es tipo MAP.

- <u>Múltiples clases (K)</u>: Uno-contra-todos ("one-vs-all")
  - ➔ una regresión logística por clase, los $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_K\}$ se aprenden por separado
  - ➔ coger la clase y_c, de mayor $h_{\boldsymbol{w}_c}(\boldsymbol{x})$

$K = 3$
$y_1 \equiv persona,$
$y_2 \equiv moto,$
$y_3 \equiv ciclista$

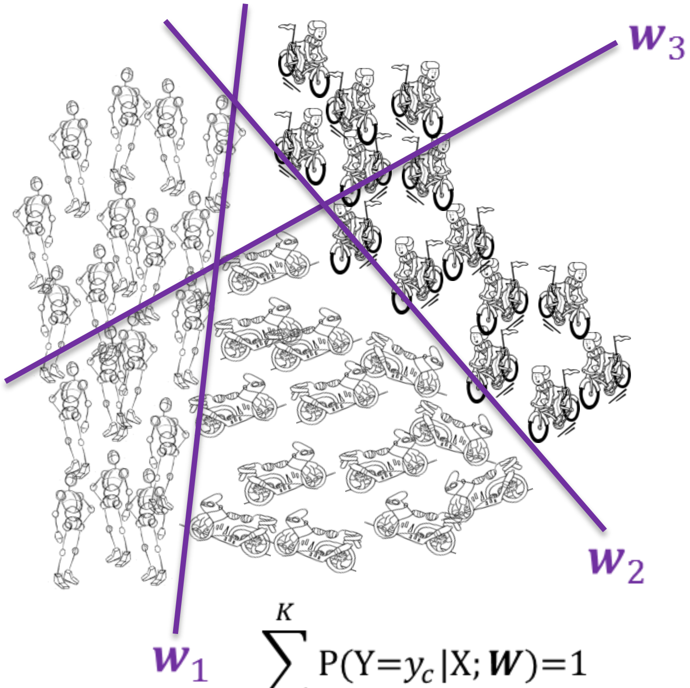




- Múltiples clases (K): Regresión Logística Multinómica (*multinomial logit*)

➔ Los $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_K\}$ se aprenden simultáneamente.

$K=3$
$y_1 \equiv persona,$
$y_2 \equiv moto,$
$y_3 \equiv ciclista$



$$\sum_{c=1}^{K} \mathrm{P}(\mathrm{Y}=y_c|\mathrm{X}; \boldsymbol{W})=1$$

$$\boldsymbol{W} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_K\}$$

$$\mathrm{P}(\mathrm{Y}=y_K|\mathrm{X}=\boldsymbol{x}; \boldsymbol{W}) = \frac{1}{1+\sum_{j=1}^{K-1} e^{\boldsymbol{w}_j^\top \boldsymbol{x}}}$$

$$\forall\, c \in \{1, \dots, K-1\}$$

$$\mathrm{P}(\mathrm{Y}=y_c|\mathrm{X}=\boldsymbol{x}; \boldsymbol{W}) = \frac{e^{\boldsymbol{w}_c^\top \boldsymbol{x}}}{1+\sum_{j=1}^{K-1} e^{\boldsymbol{w}_j^\top \boldsymbol{x}}}$$

- Conceptos clave de este vídeo:

  - Fronteras no lineales.
  - Problema del sobreajuste.
  - Regularización en regresión logística.
  - ML, MAP en regresión logística.
  - Multiclase en regresión logística.