Clase 10 - Clasificación (parte 3)

Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

14 de marzo de 2021

Regresión Logística

Analizaremos ahora los supuestos sobre el modelo generativo (i.e., las probabilidades de clase y condicionales) para encontrar un r que resulte en la bien conocida regresión logística. Consideraremos el caso binario donde las densidades condicionales de clase son Gaussianas multivariadas, dadas por

$$p(x|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(x-\mu_k)^\top\Sigma^{-1}(x-\mu_k)) \quad k \in \{1,2\}.$$

Donde $u_k \in \mathbb{R}^M$ corresponde al centroide de la clase $\mathcal{C}_k$ y $\Sigma \in \mathbb{R}^{M\times M}$ simétrica y definida positiva, corresponde a la matriz de covarianza de las clases (misma matriz para todas las clases). Para este caso, se tiene que para la ecuación:
$r = r(x) = \ln\left(\frac{\mathbb{P}(x|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(x|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right)$

Regresión Logística

Analizaremos ahora los supuestos sobre el modelo generativo (i.e., las probabilidades de clase y condicionales) para encontrar un r que resulte en la bien conocida regresión logística. Consideraremos el caso binario donde las densidades condicionales de clase son Gaussianas multivariadas, dadas por

$$p(x|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)) \quad k \in \{1, 2\}.$$

Donde $u_k \in \mathbb{R}^M$ corresponde al centroide de la clase $\mathcal{C}_k$ y $\Sigma \in \mathbb{R}^{M\times M}$ simétrica y definida positiva, corresponde a la matriz de covarianza de las clases (misma matriz para todas las clases). Para este caso, se tiene que para la ecuación:
$r = r(x) = \ln\left(\frac{\mathbb{P}(x|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(x|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right)$

Regresión Logística

Analizaremos ahora los supuestos sobre el modelo generativo (i.e., las probabilidades de clase y condicionales) para encontrar un r que resulte en la bien conocida regresión logística. Consideraremos el caso binario donde las densidades condicionales de clase son Gaussianas multivariadas, dadas por

$$p(x|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)) \quad k \in \{1,2\}.$$

Donde $u_k \in \mathbb{R}^M$ corresponde al centroide de la clase $\mathcal{C}_k$ y $\Sigma \in \mathbb{R}^{M\times M}$ simétrica y definida positiva, corresponde a la matriz de covarianza de las clases (misma matriz para todas las clases). Para este caso, se tiene que para la ecuación:
$r = r(x) = \ln\left(\frac{\mathbb{P}(x|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(x|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right)$

Regresión Logística

$$
\begin{aligned}
r &= \ln\left(\frac{\mathbb{P}(x|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(x|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right) = \ln\left(\frac{\exp(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1))\mathbb{P}(\mathcal{C}_1)}{\exp(-\frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2))\mathbb{P}(\mathcal{C}_2)}\right) \\
&= -\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&= \frac{1}{2}\left(x^\top\Sigma^{-1}(\mu_1-\mu_2) + (\mu_1-\mu_2)^\top\Sigma^{-1}x - \mu_1\Sigma^{-1}\mu_1 + \mu_2\Sigma^{-1}\mu_2\right) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&= (\mu_1-\mu_2)^\top\Sigma^{-1}x + \frac{1}{2}\left(\mu_2\Sigma^{-1}\mu_2 - \mu_1\Sigma^{-1}\mu_1\right) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&= a^\top x + b
\end{aligned}
$$

donde hemos usado la notación

$$
\begin{aligned}
a &= \Sigma^{-1}(\mu_1-\mu_2) \\
b &= \frac{1}{2}(\mu_2^\top\Sigma^{-1}\mu_2 - \mu_1^\top\Sigma^{-1}\mu_1) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right).
\end{aligned}
$$

Regresión Logística

$$
\begin{aligned}
r &= \ln\left(\frac{\mathbb{P}(x|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(x|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right) = \ln\left(\frac{\exp(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1))\mathbb{P}(\mathcal{C}_1)}{\exp(-\frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2))\mathbb{P}(\mathcal{C}_2)}\right) \\
&= -\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\top\Sigma^{-1}(x-\mu_2) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&= \frac{1}{2}\left(x^\top\Sigma^{-1}(\mu_1-\mu_2) + (\mu_1-\mu_2)^\top\Sigma^{-1}x - \mu_1\Sigma^{-1}\mu_1 + \mu_2\Sigma^{-1}\mu_2\right) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&= (\mu_1-\mu_2)^\top\Sigma^{-1}x + \frac{1}{2}\left(\mu_2\Sigma^{-1}\mu_2 - \mu_1\Sigma^{-1}\mu_1\right) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right) \\
&= a^\top x + b
\end{aligned}
$$

donde hemos usado la notación

$$
\begin{aligned}
a &= \Sigma^{-1}(\mu_1-\mu_2) \\
b &= \frac{1}{2}(\mu_2^\top\Sigma^{-1}\mu_2 - \mu_1^\top\Sigma^{-1}\mu_1) + \ln\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right).
\end{aligned}
$$

Regresión Logística

Lo anterior nos entrega la regresión logística (lineal) para el caso binario, donde al incorporar la expresión anterior en la ecuación logística obtenemos

$$p(\mathcal{C}_k|x) = \sigma(a^\top x + b) = \frac{1}{1 + \exp\left(-a^\top x - b\right)}.$$

Ahora que hemos definido el modelo para nuestro problema de clasificación, aflora naturalmente la siguiente pregunta: ¿Cómo ajustar los parámetros de las condicionales a la clase y priors respectivamente? Para esto, reiteremos que los parámetros del modelos serán los de la probabilidad de clase $p(\mathcal{C}_k)$ y de la probabilidades condicionales de clase $p(x|\mathcal{C}_k)$. Respectivamente:

- Probabilidad de clase:

  $$p(\mathcal{C}_1) = \pi, \quad p(\mathcal{C}_2) = 1 - \pi,$$

  es decir, un parámetro π (por determinar).
- Probabilidad condicional de clase:

  $$p(x|\mathcal{C}_k) = \mathcal{N}(\mu_k, \Sigma); k = 1, 2$$

  es decir, parámetros $\mu_1 \in \mathbb{R}^M, \mu_2 \in \mathbb{R}^M, \Sigma \in \mathbb{R}^M \times \mathbb{R}^M$ (por determinar) o, equivalentemente, $M + M + M(M+1)/2 = M(M+5)/2$ parámetros escalares (considerando que Σ es simétrica).

Denotaremos todos los parámetros mediante el parámetro agregado $\theta = \{\pi, \mu_1, \mu_2, \Sigma\}$.

Regresión Logística

Lo anterior nos entrega la regresión logística (lineal) para el caso binario, donde al incorporar la expresión anterior en la ecuación logística obtenemos

$$p(\mathcal{C}_k|x) = \sigma(a^\top x + b) = \frac{1}{1 + \exp(-a^\top x - b)}.$$

Ahora que hemos definido el modelo para nuestro problema de clasificación, aflora naturalmente la siguiente pregunta: ¿Cómo ajustar los parámetros de las condicionales a la clase y priors respectivamente? Para esto, reiteremos que los parámetros del modelos serán los de la probabilidad de clase $p(\mathcal{C}_k)$ y de la probabilidades condicionales de clase $p(x|\mathcal{C}_k)$. Respectivamente:

- Probabilidad de clase:

$$p(\mathcal{C}_1) = \pi, \quad p(\mathcal{C}_2) = 1 - \pi,$$

  es decir, un parámetro π (por determinar).
- Probabilidad condicional de clase:

$$p(x|\mathcal{C}_k) = \mathcal{N}(\mu_k, \Sigma); k = 1, 2$$

  es decir, parámetros $\mu_1 \in \mathbb{R}^M, \mu_2 \in \mathbb{R}^M, \Sigma \in \mathbb{R}^M \times \mathbb{R}^M$ (por determinar) o, equivalentemente, $M + M + M(M+1)/2 = M(M+5)/2$ parámetros escalares (considerando que Σ es simétrica).

Denotaremos todos los parámetros mediante el parámetro agregado $\theta = \{\pi, \mu_1, \mu_2, \Sigma\}$.

Regresión Logística

Lo anterior nos entrega la regresión logística (lineal) para el caso binario, donde al incorporar la expresión anterior en la ecuación logística obtenemos

$$p(\mathcal{C}_k|x) = \sigma(a^\top x + b) = \frac{1}{1 + \exp\left(-a^\top x - b\right)}.$$

Ahora que hemos definido el modelo para nuestro problema de clasificación, aflora naturalmente la siguiente pregunta: ¿Cómo ajustar los parámetros de las condicionales a la clase y priors respectivamente? Para esto, reiteremos que los parámetros del modelos serán los de la probabilidad de clase $p(\mathcal{C}_k)$ y de la probabilidades condicionales de clase $p(x|\mathcal{C}_k)$. Respectivamente:

- Probabilidad de clase:

  $$p(\mathcal{C}_1) = \pi, \quad p(\mathcal{C}_2) = 1 - \pi,$$

  es decir, un parámetro π (por determinar).
- Probabilidad condicional de clase:

  $$p(x|\mathcal{C}_k) = \mathcal{N}(\mu_k, \Sigma); k = 1, 2$$

  es decir, parámetros $\mu_1 \in \mathbb{R}^M, \mu_2 \in \mathbb{R}^M, \Sigma \in \mathbb{R}^M \times \mathbb{R}^M$ (por determinar) o, equivalentemente, $M + M + M(M+1)/2 = M(M+5)/2$ parámetros escalares (considerando que Σ es simétrica).

Denotaremos todos los parámetros mediante el parámetro agregado $\theta = \{\pi, \mu_1, \mu_2, \Sigma\}$.

Regresión Logística

Lo anterior nos entrega la regresión logística (lineal) para el caso binario, donde al incorporar la expresión anterior en la ecuación logística obtenemos

$$p(\mathcal{C}_k|x) = \sigma(a^\top x + b) = \frac{1}{1 + \exp(-a^\top x - b)}.$$

Ahora que hemos definido el modelo para nuestro problema de clasificación, aflora naturalmente la siguiente pregunta: ¿Cómo ajustar los parámetros de las condicionales a la clase y priors respectivamente? Para esto, reiteremos que los parámetros del modelos serán los de la probabilidad de clase $p(\mathcal{C}_k)$ y de la probabilidades condicionales de clase $p(x|\mathcal{C}_k)$. Respectivamente:

- Probabilidad de clase:

  $$p(\mathcal{C}_1) = \pi, \quad p(\mathcal{C}_2) = 1 - \pi,$$

  es decir, un parámetro π (por determinar).
- Probabilidad condicional de clase:

  $$p(x|\mathcal{C}_k) = \mathcal{N}(\mu_k, \Sigma); k = 1, 2$$

  es decir, parámetros $\mu_1 \in \mathbb{R}^M, \mu_2 \in \mathbb{R}^M, \Sigma \in \mathbb{R}^M \times \mathbb{R}^M$ (por determinar) o, equivalentemente, $M + M + M(M+1)/2 = M(M+5)/2$ parámetros escalares (considerando que Σ es simétrica).

Denotaremos todos los parámetros mediante el parámetro agregado $\theta = \{\pi, \mu_1, \mu_2, \Sigma\}$.

Regresión Logística

Lo anterior nos entrega la regresión logística (lineal) para el caso binario, donde al incorporar la expresión anterior en la ecuación logística obtenemos

$$p(\mathcal{C}_k|x) = \sigma(a^\top x + b) = \frac{1}{1 + \exp\left(-a^\top x - b\right)}.$$

Ahora que hemos definido el modelo para nuestro problema de clasificación, aflora naturalmente la siguiente pregunta: ¿Cómo ajustar los parámetros de las condicionales a la clase y priors respectivamente? Para esto, reiteremos que los parámetros del modelos serán los de la probabilidad de clase $p(\mathcal{C}_k)$ y de la probabilidades condicionales de clase $p(x|\mathcal{C}_k)$. Respectivamente:

- Probabilidad de clase:

  $$p(\mathcal{C}_1) = \pi, \quad p(\mathcal{C}_2) = 1 - \pi,$$

  es decir, un parámetro π (por determinar).
- Probabilidad condicional de clase:

  $$p(x|\mathcal{C}_k) = \mathcal{N}(\mu_k, \Sigma); k = 1, 2$$

  es decir, parámetros $\mu_1 \in \mathbb{R}^M, \mu_2 \in \mathbb{R}^M, \Sigma \in \mathbb{R}^M \times \mathbb{R}^M$ (por determinar) o, equivalentemente, $M + M + M(M+1)/2 = M(M+5)/2$ parámetros escalares (considerando que Σ es simétrica).

Denotaremos todos los parámetros mediante el parámetro agregado $\theta = \{\pi, \mu_1, \mu_2, \Sigma\}$.

Regresión Logística

Realizaremos el entrenamiento del modelo mediante el método de máxima verosimilitud. Consideremos la codificación donde la observación (x_i, t_i) corresponde a clase $\mathcal{C}_1$ con $t_i = 1$ y a clase $\mathcal{C}_2$ con $t = 0$, podemos expresar la verosimilitud con una observación mediante:

$$L_i(\theta) = p(x_i, t_i|\theta) = p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i}.$$

Para un conjunto de $\mathcal{D}$ de la forma

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} \in \mathbb{R}^{N\times M}, \qquad T = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \in \{0,1\}^N \text{ es decir, codificación } 0-1.$$

podemos escribir la verosimilitud mediante $L(\theta) = p(X, T|\theta)$, luego:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} p(x_i, t_i|\theta) = \prod_{i=1}^{N} p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i} \\ &= \prod_{i=1}^{N} \left(p(x_i|\mathcal{C}_1, \theta) p(\mathcal{C}_1|\theta)\right)^{t_i} \left(p(x_i|\mathcal{C}_0, \theta) p(\mathcal{C}_0|\theta)\right)^{1-t_i} \\ &= \prod_{i=1}^{N} (\pi \mathcal{N}(x_i|\mu_1, \Sigma))^{t_i} ((1-\pi)\mathcal{N}(x_i|\mu_2, \Sigma))^{1-t_i}. \end{aligned}$$

Regresión Logística

Realizaremos el entrenamiento del modelo mediante el método de máxima verosimilitud. Consideremos la codificación donde la observación (x_i, t_i) corresponde a clase $\mathcal{C}_1$ con $t_i = 1$ y a clase $\mathcal{C}_2$ con $t = 0$, podemos expresar la verosimilitud con una observación mediante:

$$L_i(\theta) = p(x_i, t_i|\theta) = p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i}.$$

Para un conjunto de $\mathcal{D}$ de la forma

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} \in \mathbb{R}^{N\times M}, \qquad T = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \in \{0,1\}^N \text{ es decir, codificación } 0-1.$$

podemos escribir la verosimilitud mediante $L(\theta) = p(X, T|\theta)$, luego:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^N p(x_i, t_i|\theta) = \prod_{i=1}^N p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i} \\ &= \prod_{i=1}^N (p(x_i|\mathcal{C}_1, \theta)p(\mathcal{C}_1|\theta))^{t_i} (p(x_i|\mathcal{C}_0, \theta)p(\mathcal{C}_0|\theta))^{1-t_i} \\ &= \prod_{i=1}^N (\pi\mathcal{N}(x_i|\mu_1, \Sigma))^{t_i} ((1-\pi)\mathcal{N}(x_i|\mu_2, \Sigma))^{1-t_i}. \end{aligned}$$

Regresión Logística

Realizaremos el entrenamiento del modelo mediante el método de máxima verosimilitud. Consideremos la codificación donde la observación (x_i, t_i) corresponde a clase $\mathcal{C}_1$ con $t_i = 1$ y a clase $\mathcal{C}_2$ con $t = 0$, podemos expresar la verosimilitud con una observación mediante:

$$L_i(\theta) = p(x_i, t_i|\theta) = p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i}.$$

Para un conjunto de $\mathcal{D}$ de la forma

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} \in \mathbb{R}^{N\times M}, \qquad T = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \in \{0,1\}^N \text{ es decir, codificación } 0-1.$$

podemos escribir la verosimilitud mediante $L(\theta) = p(X, T|\theta)$, luego:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^N p(x_i, t_i|\theta) = \prod_{i=1}^N p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i} \\ &= \prod_{i=1}^N \left(p(x_i|\mathcal{C}_1, \theta)p(\mathcal{C}_1|\theta)\right)^{t_i} \left(p(x_i|\mathcal{C}_0, \theta)p(\mathcal{C}_0|\theta)\right)^{1-t_i} \\ &= \prod_{i=1}^N (\pi\mathcal{N}(x_i|\mu_1, \Sigma))^{t_i} ((1-\pi)\mathcal{N}(x_i|\mu_2, \Sigma))^{1-t_i}. \end{aligned}$$

Regresión Logística

Realizaremos el entrenamiento del modelo mediante el método de máxima verosimilitud. Consideremos la codificación donde la observación (x_i, t_i) corresponde a clase $\mathcal{C}_1$ con $t_i = 1$ y a clase $\mathcal{C}_2$ con $t = 0$, podemos expresar la verosimilitud con una observación mediante:

$$L_i(\theta) = p(x_i, t_i|\theta) = p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i}.$$

Para un conjunto de $\mathcal{D}$ de la forma

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} \in \mathbb{R}^{N\times M}, \qquad T = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \in \{0,1\}^N \text{ es decir, codificación } 0-1.$$

podemos escribir la verosimilitud mediante $L(\theta) = p(X, T|\theta)$, luego:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} p(x_i, t_i|\theta) = \prod_{i=1}^{N} p(x_i, \mathcal{C}_1|\theta)^{t_i} p(x_i, \mathcal{C}_0|\theta)^{1-t_i} \\ &= \prod_{i=1}^{N} \left(p(x_i|\mathcal{C}_1, \theta)p(\mathcal{C}_1|\theta)\right)^{t_i} \left(p(x_i|\mathcal{C}_0, \theta)p(\mathcal{C}_0|\theta)\right)^{1-t_i} \\ &= \prod_{i=1}^{N} (\pi \mathcal{N}(x_i|\mu_1, \Sigma))^{t_i} ((1-\pi)\mathcal{N}(x_i|\mu_2, \Sigma))^{1-t_i}. \end{aligned}$$

Regresión Logística

Nuestro interés se encuentra en la log-verosimilitud

$$l(\theta) := \log L(\theta)$$

$$= \sum_{i=1}^{N} (t_i(\log(\pi) + \log(\mathcal{N}(x_i|\mu_1, \Sigma))) + (1 - t_i)(\log(1 - \pi) + \log(\mathcal{N}(x_i|\mu_2, \Sigma))))$$

Aplicando condiciones de primer orden

- **1)** Con respecto a π:

$$\frac{\partial \log(L)}{\partial \pi} = \sum_{i=1}^{N} \frac{t_i}{\pi} - \frac{1 - t_i}{1 - \pi} = 0$$

$$\Rightarrow \quad (1 - \pi) \sum_{i=1}^{N} t_i = \pi \sum_{i=1}^{N} (1 - t_i)$$

$$\Rightarrow \quad \sum_{i=1}^{N} t_i = \pi N \quad \Rightarrow \quad \pi = \frac{\sum_{i=1}^{N} t_i}{N} = \frac{N_1}{N_1 + N_2} \tag{1.1}$$

Donde $N_i := \text{Card}(x : x \in \mathcal{C}_i)$. Por lo tanto, el EMV de *pi* colapsa a la regla de Laplace.

Regresión Logística

Nuestro interés se encuentra en la log-verosimilitud

$$l(\theta) := \log L(\theta)$$

$$= \sum_{i=1}^{N} (t_i(\log(\pi) + \log(\mathcal{N}(x_i|\mu_1, \Sigma))) + (1-t_i)(\log(1-\pi) + \log(\mathcal{N}(x_i|\mu_2, \Sigma))))$$

Aplicando condiciones de primer orden

- **1)** Con respecto a π:

$$\frac{\partial \log(L)}{\partial \pi} = \sum_{i=1}^{N} \frac{t_i}{\pi} - \frac{1-t_i}{1-\pi} = 0$$

$$\Rightarrow \quad (1-\pi)\sum_{i=1}^{N} t_i = \pi \sum_{i=1}^{N}(1-t_i)$$

$$\Rightarrow \quad \sum_{i=1}^{N} t_i = \pi N \quad \Rightarrow \quad \pi = \frac{\sum_{i=1}^{N} t_i}{N} = \frac{N_1}{N_1 + N_2} \tag{1.1}$$

Donde $N_i := \text{Card}(x : x \in \mathcal{C}_i)$. Por lo tanto, el EMV de *pi* colapsa a la regla de Laplace.

Regresión Logística

- **2)** Con respecto a μ_1:

$$\frac{\partial \log(L)}{\partial \mu_1} = \sum_{i=1}^{N} t_i \frac{\partial}{\partial \mu_1}(-\frac{1}{2}(x_i - \mu_1)^\top \Sigma^{-1}(x_i - \mu_1))$$

$$= \sum_{i=1}^{N} t_i(\Sigma^{-1}(x_i - \mu_1)) = \Sigma^{-1} \sum_{i=1}^{N} t_i(x_i - \mu_1) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{N} t_i x_i = \mu_1 \sum_{i=1}^{N} t_i \quad \Rightarrow \quad \mu_1 = \frac{1}{N_1} \sum_{i=1}^{N} t_i x_i = \frac{1}{N_1} \sum_{x_i \in \mathcal{C}_1} x_i$$

De forma análoga:

$$\mu_2 = \frac{1}{N_2} \sum_{x_i \in \mathcal{C}_2} x_i$$

Queda la siguiente pregunta, **¿Cuál es el estimador MV de Σ?**

Regresión Logística

- **2)** Con respecto a μ_1:

$$\frac{\partial \log(L)}{\partial \mu_1} = \sum_{i=1}^{N} t_i \frac{\partial}{\partial \mu_1}(-\frac{1}{2}(x_i - \mu_1)^\top \Sigma^{-1}(x_i - \mu_1))$$

$$= \sum_{i=1}^{N} t_i(\Sigma^{-1}(x_i - \mu_1)) = \Sigma^{-1} \sum_{i=1}^{N} t_i(x_i - \mu_1) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{N} t_i x_i = \mu_1 \sum_{i=1}^{N} t_i \quad \Rightarrow \quad \mu_1 = \frac{1}{N_1} \sum_{i=1}^{N} t_i x_i = \frac{1}{N_1} \sum_{x_i \in \mathcal{C}_1} x_i$$

De forma análoga:

$$\mu_2 = \frac{1}{N_2} \sum_{x_i \in \mathcal{C}_2} x_i$$

Queda la siguiente pregunta, **¿Cuál es el estimador MV de Σ?**

Regresión Logística v/s modelo generativo

Recordemos que los supuestos tomados sobre el modelo generativo para el problema de clasificación resultaron en:

$$p(\mathcal{C}_1|x) = \sigma(w^\top x) = \frac{1}{1 + e^{-w^\top x}}$$

donde por claridad de notación hemos elegido la representación lineal $(w^\top x)$ y no afín $(a^\top x + b)$. En el caso anterior se ha entrenado el modelo generativo completo, es decir, $\pi, \mu_1, \mu_2, \Sigma$, lo cual tiene la ventaja de tener solución en forma cerrada, sin embargo, puede ser innecesario cuando solo necesitamos conocer el peso w en la ecuación anterior. Calculemos la verosimilitud de la regresión logística con datos $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^N$, para hacer la notación más compacta denotamos $\sigma_i = \sigma(w^\top x_i)$. Entonces:

$$p((t_i)_{i=1}^N|(x_i)_{i=1}^N, w) = \prod_{i=1}^N p(t_i|x_i, w) = \prod_{i=1}^N p(\mathcal{C}_1|x_i)^{t_i} p(\mathcal{C}_2|x_i)^{1-t_i} = \prod_{i=1}^N \sigma_i^{t_i}(1-\sigma_i)^{1-t_i}$$

Regresión Logística v/s modelo generativo

Recordemos que los supuestos tomados sobre el modelo generativo para el problema de clasificación resultaron en:

$$p(\mathcal{C}_1|x) = \sigma(w^\top x) = \frac{1}{1+e^{-w^\top x}}$$

donde por claridad de notación hemos elegido la representación lineal $(w^\top x)$ y no afín $(a^\top x + b)$. En el caso anterior se ha entrenado el modelo generativo completo, es decir, $\pi, \mu_1, \mu_2, \Sigma$, lo cual tiene la ventaja de tener solución en forma cerrada, sin embargo, puede ser innecesario cuando solo necesitamos conocer el peso w en la ecuación anterior. Calculemos la verosimilitud de la regresión logística con datos $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^N$, para hacer la notación más compacta denotamos $\sigma_i = \sigma(w^\top x_i)$. Entonces:

$$p((t_i)_{i=1}^N|(x_i)_{i=1}^N, w) = \prod_{i=1}^N p(t_i|x_i, w) = \prod_{i=1}^N p(\mathcal{C}_1|x_i)^{t_i} p(\mathcal{C}_2|x_i)^{1-t_i} = \prod_{i=1}^N \sigma_i^{t_i}(1-\sigma_i)^{1-t_i}$$

Regresión Logística v/s modelo generativo

Recordemos que los supuestos tomados sobre el modelo generativo para el problema de clasificación resultaron en:

$$p(\mathcal{C}_1|x) = \sigma(w^\top x) = \frac{1}{1 + e^{-w^\top x}}$$

donde por claridad de notación hemos elegido la representación lineal $(w^\top x)$ y no afín $(a^\top x + b)$. En el caso anterior se ha entrenado el modelo generativo completo, es decir, $\pi, \mu_1, \mu_2, \Sigma$, lo cual tiene la ventaja de tener solución en forma cerrada, sin embargo, puede ser innecesario cuando solo necesitamos conocer el peso w en la ecuación anterior. Calculemos la verosimilitud de la regresión logística con datos $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^N$, para hacer la notación más compacta denotamos $\sigma_i = \sigma(w^\top x_i)$. Entonces:

$$p((t_i)_{i=1}^N|(x_i)_{i=1}^N, w) = \prod_{i=1}^N p(t_i|x_i, w) = \prod_{i=1}^N p(\mathcal{C}_1|x_i)^{t_i} p(\mathcal{C}_2|x_i)^{1-t_i} = \prod_{i=1}^N \sigma_i^{t_i}(1-\sigma_i)^{1-t_i}$$

Regresión Logística v/s modelo generativo

Con lo que la log-verosimilitud está dada por

$$l(w) = \sum_{i=1}^{N} t_i \log(\sigma_i) + (1 - t_i) \log(1 - \sigma_i).$$

Notemos que este problema de optimización no exhibe una solución en forma cerrada, por lo que podemos resolverlo mediante gradiente, para lo cual es necesario calcular el gradiente de $l(w)$ respecto a w:

$$\nabla_w l(w) = \sum_{i=1}^{N} (t_i - \sigma_i) x_i,$$

lo cual nos da una regla de ajuste $\theta \mapsto \theta - \eta \sum_{i=1}^{N} (\sigma_i - t_i) x_i$, o bien

$$\theta \mapsto \theta + \eta (t_i - \sigma_i) x_i,$$

si tomamos los datos de "a uno" (método del gradiente estocástico).

Regresión Logística v/s modelo generativo

Con lo que la log-verosimilitud está dada por

$$l(w) = \sum_{i=1}^{N} t_i \log(\sigma_i) + (1 - t_i) \log(1 - \sigma_i).$$

Notemos que este problema de optimización no exhibe una solución en forma cerrada, por lo que podemos resolverlo mediante gradiente, para lo cual es necesario calcular el gradiente de $l(w)$ respecto a w:

$$\nabla_w l(w) = \sum_{i=1}^{N} (t_i - \sigma_i) x_i,$$

lo cual nos da una regla de ajuste $\theta \mapsto \theta - \eta \sum_{i=1}^{N} (\sigma_i - t_i) x_i$, o bien

$$\theta \mapsto \theta + \eta (t_i - \sigma_i) x_i,$$

si tomamos los datos de “a uno” (método del gradiente estocástico).

Regresión Logística v/s modelo generativo

Con lo que la log-verosimilitud está dada por

$$l(w) = \sum_{i=1}^{N} t_i \log(\sigma_i) + (1 - t_i) \log(1 - \sigma_i).$$

Notemos que este problema de optimización no exhibe una solución en forma cerrada, por lo que podemos resolverlo mediante gradiente, para lo cual es necesario calcular el gradiente de $l(w)$ respecto a w:

$$\nabla_w l(w) = \sum_{i=1}^{N} (t_i - \sigma_i) x_i,$$

lo cual nos da una regla de ajuste $\theta \mapsto \theta - \eta \sum_{i=1}^{N} (\sigma_i - t_i) x_i$, o bien

$$\theta \mapsto \theta + \eta (t_i - \sigma_i) x_i,$$

si tomamos los datos de "a uno" (método del gradiente estocástico).

Regresión Logística v/s modelo generativo

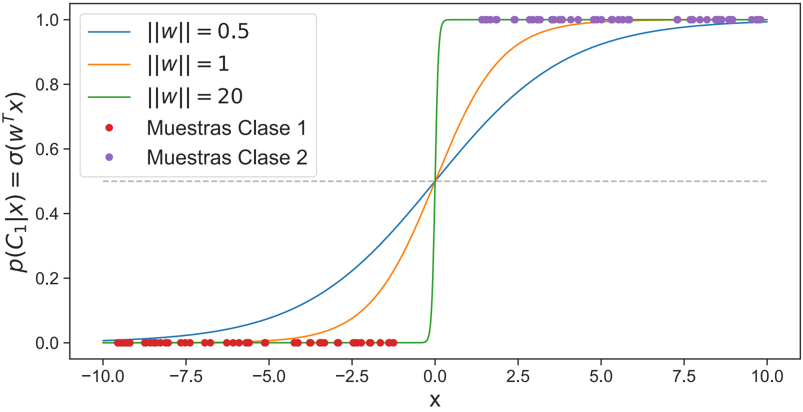

Fig.. En gris la frontera de decisión: una nueva entrada $x_\star$ será asignada a la clase $\mathcal{C}_1$ si $p(\mathcal{C}_1|x_\star) > \frac{1}{2}$, en caso contrario, será asignada a $\mathcal{C}_2$. Se observa que al entrenar con más muestras, $\|w\|$ crece por lo que el parámetro se sobreajusta a los datos y el clasificador converge a una función indicatriz.

Clase 10 - Clasificación (parte 3)

Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

14 de marzo de 2021

UNIVERSIDAD
DE CHILE