Tema 7: Regresión Logística

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

http://www.sc.ehu.es/isg/

Introducción

- Regresión logística se basa en la *función logística*: $f(z) = \frac{1}{1+e^{-z}}$
  - $0 < f(z) < 1$ puede ser interpreta en términos de probabilidad
  - $\lim_{z \to -\infty} \frac{1}{1+e^{-z}} = 0$
  - $\lim_{z \to +\infty} \frac{1}{1+e^{-z}} = 1$
  - $f(0) = \frac{1}{1+e^{-0}} = \frac{1}{2}$
- Muy usada en Ciencias de la Salud: los parámetros tienen una interpretación en términos de riesgo

El modelo logístico

- Paradigma de regresión logística:

$$P(C=1|\mathbf{x}) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

  - $\beta_0, \beta_1, \ldots, \beta_n$ son los parámetros, que deben ser estimados a partir de los datos

- Si C es binaria:

$$P(C=0|\mathbf{x}) = 1 - \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

$$= \frac{e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

El modelo logístico

risk ratio ($RR(\mathbf{x}, \mathbf{x}')$)

- Variables: C `Enfermedad Coronaria` (1 si, 0 no); X_1 `Nivel de Colesterol` (1 alto, 0 bajo), X_2 `Edad`, y X_3 `Resultado del Electrocardiograma` (1 anormal, 0 normal)
- Parámetros ($N = 609$ casos): $\widehat{\beta_0} = -3{,}911$ $\widehat{\beta_1} = 0{,}652$ $\widehat{\beta_2} = 0{,}029$ $\widehat{\beta_3} = 0{,}342$
- Comparar el riesgo para dos patrones: $\mathbf{x} = (1, 40, 0)$ y $\mathbf{x}' = (0, 40, 0)$
  - $P(C = 1|\mathbf{x}) = P(C = 1|X_1 = 1, X_2 = 40, X_3 = 0)$
    $= \frac{1}{1+e^{-(-3{,}911 + 0{,}652(1) + 0{,}029(40) + 0{,}342(0))}} = 0{,}109$
  - $P(C = 1|\mathbf{x}') = P(C = 1|X_1 = 0, X_2 = 40, X_3 = 0)$
    $= \frac{1}{1+e^{-(-3{,}911 + 0{,}652(0) + 0{,}029(40) + 0{,}342(0))}} = 0{,}060$
- $RR(\mathbf{x}, \mathbf{x}') = \frac{P(C=1|\mathbf{x})}{P(C=1|\mathbf{x}')} = \frac{P(C=1|X_1=1, X_2=40, X_3=0)}{P(C=1|X_1=0, X_2=40, X_3=0)} = \frac{0{,}109}{0{,}060} = 1{,}82$
- Para una persona con 40 años y electrocardiograma normal, el riesgo se multiplica casi por dos al pasar de un nivel de `Colesterol` bajo(0) a uno alto (1)

El modelo logístico

- *odds ratio* $OR(\mathbf{x}) = \frac{P(C=1|\mathbf{x})}{1-P(C=1|\mathbf{x})}$
- Modelo logístico en forma *logit*

$$logit\ (P(C=1|\mathbf{x})) = ln\ OR(\mathbf{x}) = ln\left[\frac{P(C=1|\mathbf{x})}{1-P(C=1|\mathbf{x})}\right]$$
$$= \ldots = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

  - $ln\ OR(\mathbf{0})) = \beta_0$
  - $\mathbf{x} = (1, 40, 0)$, $\mathbf{x}' = (0, 40, 0)$

$$logit\ P(C=1|\mathbf{x})) = \beta_0 + 1 \cdot \beta_1 + 40 \cdot \beta_2 + 0 \cdot \beta_3$$
$$logit\ P(C=1|\mathbf{x}')) = \beta_0 + 0 \cdot \beta_1 + 40 \cdot \beta_2 + 0 \cdot \beta_3$$
$$logit\ P(C=1|\mathbf{x})) - logit\ P(C=1|\mathbf{x}')) = \beta_1$$

El modelo logístico

Teorema: *En un modelo de regresión logística, el coeficiente* β_i *representa el cambio en el logit resultante al aumentar una unidad en la* i*-ésima variable* $X_i (i = 1, \ldots, n)$

Demostración: $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$ y $\mathbf{x}' = (x'_1, \ldots, x'_i, \ldots, x'_n)$ verificando $x_j = x'_j$ para todo $j \neq i$ y $x'_i = x_i + 1$

$$logit\ (\mathbf{x}') - logit\ (\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i x'_i - \left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i \right) =$$

$$\beta_i x'_i - \beta_i x_i = \beta_i (x_i + 1 - x_i) = \beta_i$$

El modelo logístico

- *risk odds ratio* $ROR(\mathbf{x}', \mathbf{x}) = \frac{OR(\mathbf{x}')}{OR(\mathbf{x})} = e^{\sum_{i=1}^{n} \beta_i (x_i' - x_i)}$
  - Mide el riesgo del *odds ratio* de $\mathbf{x}'$ frente al *odds ratio* de $\mathbf{x}$
  - Puede expresar de manera alternativa como:

$$ROR(\mathbf{x}', \mathbf{x}) = \prod_{i=1}^{n} e^{\beta_i (x_i' - x_i)} = e^{\beta_1 (x_1' - x_1)} \cdot \ldots \cdot e^{\beta_n (x_n' - x_n)}$$

Estimación máximo verosimil de los parámetros

- Función de verosimilitud: $L\left((\mathbf{x}^{(1)}, c^{(1)}), \ldots, (\mathbf{x}^{(N)}, c^{(N)}), \beta_0, \beta_1, \ldots, \beta_n\right)$
$$= \prod_{j=1}^{N} P\left(C = 1|\mathbf{x}^{(j)}\right)^{c^{(j)}} \left(1 - P\left(C = 1|\mathbf{x}^{(j)}\right)\right)^{1-c^{(j)}}$$

- Logaritmo de la verosimilitud:
$$lnL\left((\mathbf{x}^{(1)}, c^{(1)}), \ldots, (\mathbf{x}^{(N)}, c^{(N)}), \beta_0, \beta_1, \ldots, \beta_n\right) =$$
$$= \sum_{j=1}^{N} c^{(j)} lnP\left(C = 1|\mathbf{x}^{(j)}\right) + \sum_{j=1}^{N} \left(1 - c^{(j)}\right) ln\left(1 - P\left(C = 1|\mathbf{x}^{(j)}\right)\right) =$$
$$= \sum_{j=1}^{N} c^{(j)} \left[lnP\left(C = 1|\mathbf{x}^{(j)}\right) - ln\left(1 - P\left(C = 1|\mathbf{x}^{(j)}\right)\right)\right] + \sum_{j=1}^{N} ln\left(1 - P\left(C = 1|\mathbf{x}^{(j)}\right)\right)$$
$$= \sum_{j=1}^{N} c^{(j)} ln \frac{P\left(C = 1|\mathbf{x}^{(j)}\right)}{1 - P\left(C = 1|\mathbf{x}^{(j)}\right)} + \sum_{j=1}^{N} ln\left(1 - P\left(C = 1|\mathbf{x}^{(j)}\right)\right) = \ldots$$
$$= \sum_{j=1}^{N} c^{(j)} \left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right) - \sum_{j=1}^{N} ln\left(1 + e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}\right)$$

Estimación máximo verosimil de los parámetros

$\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_n}$, para los parámetros $\beta_0, \beta_1, \ldots, \beta_n$, como soluciones del sistema:

$$\frac{\partial lnL}{\partial \beta_0} = \sum_{j=1}^{N} c^{(j)} - \sum_{j=1}^{N} \frac{e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}}{1 + e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}} = 0$$

$$\frac{\partial lnL}{\partial \beta_1} = \sum_{j=1}^{N} c^{(j)} x_1^{(j)} - \sum_{j=1}^{N} x_1^{(j)} \frac{e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}}{1 + e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}} = 0$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\partial lnL}{\partial \beta_n} = \sum_{j=1}^{N} c^{(j)} x_n^{(j)} - \sum_{j=1}^{N} x_n^{(j)} \frac{e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}}{1 + e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i^{(j)}\right)}} = 0$$

Estimación máximo verosimil de los parámetros

- No es posible obtener una fórmula cerrada para los estimadores de los parámetros
- Método de Newton-Raphson (técnica iterativa):

$$\widehat{\beta}^{nuevo} = \widehat{\beta}^{viejo} + (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{c} - \widehat{\mathbf{p}})$$

$\mathbf{X} \in M(N, n)$ matriz cuyas filas son $\mathbf{x}^{(j)}, j = 1, \ldots, N$

$$\mathbf{W} = \begin{pmatrix} p^{(1)}\left(1 - p^{(1)}\right) & \ldots & \ldots & 0 \\ 0 & p^{(2)}\left(1 - p^{(2)}\right) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & p^{(N)}\left(1 - p^{(N)}\right) \end{pmatrix}$$

$\widehat{\mathbf{p}} \in M(N, 1)$, con $p^{(j)} = \dfrac{e^{\left(\mathbf{x}^{(j)}\widehat{\beta}^{viejo}\right)}}{1 + e^{\left(\mathbf{x}^{(j)}\widehat{\beta}^{viejo}\right)}}$

$\mathbf{c} \in M(N, 1)$ vector de componentes $c^{(j)}, j = 1, \ldots, N$

Test de la razón de verosimilitud

- Compara dos modelos de regresión logística: el *modelo completo* frente al *modelo reducido* –submodelo del completo–
- El test de la razón de verosimilitud compara -2 veces el logaritmo neperiano de un cociente entre las verosimilitudes del modelo completo y el modelo reducido, con el percentil correspondiente de una distribución chi-cuadrado
- Hipótesis nula: los parámetros de las variables que forman parte del modelo completo, pero no del modelo reducido, valen cero

Test de la razón de verosimilitud

`Modelo 1`: logit $P_1(C=1|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2$

`Modelo 2`: logit $P_2(C=1|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

`Modelo 3`: logit

$P_3(C=1|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3$

- `Modelo 1` –modelo reducido– frente al `Modelo 2` –modelo completo–
- `Modelo 2` –modelo reducido– frente al `Modelo 3` –modelo completo–

Test de la razón de verosimilitud

`Modelo 1`: logit $P_1(C=1|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2$

`Modelo 2`: logit $P_2(C=1|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$$H_0 : \beta_3 = 0 \text{ frente a } H_1 : \beta_3 \neq 0$$

- Si X_3 hace que el `Modelo 2` se ajuste mucho mejor a los datos que el `Modelo 1` entonces $\widehat{L_2}$ será mucho mayor que $\widehat{L_1} \Longrightarrow \dfrac{\widehat{L_1}}{\widehat{L_2}} \simeq 0 \Longrightarrow ln\dfrac{\widehat{L_1}}{\widehat{L_2}} \simeq -\infty \Longrightarrow$ $-2ln\dfrac{\widehat{L_1}}{\widehat{L_2}} \simeq +\infty$
- Cuanto mayor sea el valor de $-2ln\dfrac{\widehat{L_1}}{\widehat{L_2}}$ más en contra estaremos de la hipótesis nula $H_0 : \beta_3 = 0$
- $-2ln\dfrac{\widehat{L_1}}{\widehat{L_2}}$ sigue –cuando N es suficientemente grande– bajo la hipótesis nula H_0 una distribución de probabilidad χ^2_r, con r igual al número de parámetros que en el modelo completo deben igualarse a cero para que dicho modelo completo coincida con el modelo reducido

El test de Wald

- Sólo puede ser usado para testar un único parámetro: $H_0 : \beta_j = 0$ frente a $H_A : \beta_j \neq 0$
- Válido para testar el `Modelo 2` frente al `Modelo 1`. No sirve para testar el `Modelo 3` frente al `Modelo 2`
- Estadístico de Wald para la variable X_j: $\dfrac{\widehat{\beta}_j}{\widehat{S}_{\beta_j}}$
- Se verifica que $\dfrac{\widehat{\beta}_j}{\widehat{S}_{\beta_j}} \rightsquigarrow \mathcal{N}(0,1)$ o lo que es equivalente, $\left(\dfrac{\widehat{\beta}_j}{\widehat{S}_{\beta_j}}\right)^2 \rightsquigarrow \chi^2_1$

Modelización

- El test de la razón de verosimilitud como el test de Wald son instrumentos para la modelización
- Si la enumeración completa de todos los modelos posibles es costosa, se utilizan estrategias –secuenciales– de modelización
  a) *Selección hacia adelante*: en cada etapa se añade la mejor variable predictora aún no seleccionada
  b) *Eliminación hacia atrás*: partiendo del conjunto completo de variables predictoras, se va eliminando en cada etapa la peor variable predictora hasta que las variables que quedan en el modelo son todas ellas pertinentes
  c) *Modelización paso a paso*: se combinan las dos estrategias anteriores