Tema 13: Regresión Logística

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

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

Contenido

- Introducción
- El modelo Logístico
- Estimación máximo verosímil de los parámetros
- Interpretación en términos de riesgo
- Test de hipótesis

Introducción

Clasificación Supervisada

	X_1	$\ldots$	X_n	C
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	$c^{(1)}$
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	$c^{(2)}$
$\ldots$		$\ldots$		$\ldots$
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	$c^{(N)}$
$\boldsymbol{x}^{(N+1)}$	$x_1^{(N+1)}$	$\ldots$	$x_n^{(N+1)}$	???

Introducción

Regresión lineal

Si denotamos por C a la variable a predecir, y por $X_1, \ldots, X_n$ a las n variables predictoras, el paradigma de regresión lineal se expresa de la manera siguiente:

$$g_\beta(\mathbf{x}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n = \beta_0 + \sum_{i=1}^{n} \beta_i X_i = \beta^T \mathbf{x}$$

donde $g_\beta(\mathbf{x})$ es la función lineal y β^T es el vector de los parámetros (llamado también vector de pesos).

Introducción

Regresión lineal: problema de aprendizaje

$$J(\beta) = \frac{1}{2}\sum_{i=1}^{N}[g(x^{(i)}) - c^{(i)}]^2$$

donde $J(\beta)$ es simplemente la función de coste de mínimos cuadrados.

El objetivo es encontrar β que minimiza $J(\beta)$.

El modelo logístico

La regresión logística se basa en la *función logística*:

$$f_\beta(\mathbf{x}) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i X_i)}} = \frac{1}{1+e^{-\beta^T\mathbf{x}}}$$

donde,

$$\begin{aligned}
\lim_{\mathbf{x}\to-\infty} f(\mathbf{x}) &= 0 \\
\lim_{\mathbf{x}\to+\infty} f(\mathbf{x}) &= 1 \\
f(0) &= \frac{1}{2} \\
f'(\mathbf{x}) &= f(\mathbf{x})(1-f(\mathbf{x}))
\end{aligned}$$

Esta función es muy usada en Ciencias de la Salud: los parámetros tienen una interpretación en términos de riesgo

El modelo logístico

Asumiendo que la variable clase C es binaria, el paradigma de regresión logística se expresa de la manera siguiente:

$$P(C=1|\mathbf{x};\beta) = f_\beta(\mathbf{x}) = \frac{1}{1+e^{-\beta^T\mathbf{x}}}$$

$$P(C=0|\mathbf{x};\beta) = 1 - f_\beta(\mathbf{x}) = \frac{e^{-\beta^T\mathbf{x}}}{1+e^{-\beta^T\mathbf{x}}}$$

donde β^T es el vector de los parámetros que deben ser estimados a partir de los datos, y $\mathbf{x}$ es el vector de las variables predictoras.

Estimación máximo verosímil de los parámetros

Para una función densidad de probabilidad dada por:

$$P(C|\mathbf{x};\beta) = f_\beta(\mathbf{x})^C.(1 - f_\beta(\mathbf{x}))^{1-C}$$

La función de verosimilitud asociada a una muestra de tamaño N viene dada por:

$$L(\mathbf{x}, C; \beta) = \prod_{j=1}^{N} f_\beta(\mathbf{x}^{(j)})^{c^{(j)}}.(1 - f_\beta(\mathbf{x}^{(j)}))^{1-c^{(j)}}$$

cuya función logaritmo viene dada por:

$$lnL(\mathbf{x}, C; \beta) = l_\beta(\mathbf{x}, C) = \sum_{j=1}^{N} c^{(j)} ln f_\beta(\mathbf{x}^{(j)}) + \sum_{j=1}^{N} \left(1 - c^{(j)}\right) ln \left(1 - f_\beta(\mathbf{x}^{(j)}\right)$$

Estimación máximo verosímil de los parámetros

El Método de Newton-Raphson

$$\widehat{\beta} := \widehat{\beta} - \frac{f(\beta)}{f'(\beta)}.$$

Aplicando esta regla a nuestro modelo de regresión logística ($f(\beta) = \nabla_\beta l_\beta$), tenemos:

$$\widehat{\beta} := \widehat{\beta} - H^{-1} \nabla_\beta l_\beta(\mathbf{x}, C).$$

donde $l_\beta(\mathbf{x}, C)$ es la función log-verosimil, $H_{ij} = \frac{\partial^2 l_\beta}{\partial \beta_i \partial \beta_j}$ es la matriz Hessiana $i, j = 1, ..., n$, donde n el número del los parámetros. ∇_β es el vector gradiente de valores $\nabla_{\beta_i} l_\beta = \frac{\partial l_\beta}{\partial \beta_i}$. $i = 1, ..., n$.

Estimación máximo verosimil de los parámetros

Para un ejemplo de entrenamiento $(\mathbf{x}^{(j)}, c^{(j)})$, tenemos:

$$\begin{aligned}
\frac{\partial l_\beta}{\partial \beta_i} &= \left(\frac{c^{(j)}}{f_\beta(x^{(j)})} - \frac{(1-c^{(j)})}{1-f_\beta(x^{(j)})}\right)\frac{\partial f_\beta(x^{(j)})}{\partial \beta_i}\frac{\partial l_\beta}{\partial \beta_i} \\
\frac{\partial l_\beta}{\partial \beta_i} &= \left[\frac{c^{(j)}}{f_\beta(x^{(j)})} - \frac{(1-c^{(j)})}{1-f_\beta(x^{(j)})}\right][f_\beta(x^{(j)}).(1-f_\beta(x^{(j)}))]\frac{\partial(\beta^T\mathbf{x})}{\partial \beta_i} \\
\frac{\partial l_\beta}{\partial \beta_i} &= (c^{(j)} - f_\beta(x^{(j)})).x_i
\end{aligned}$$

Adoptando una notación matricial, tenemos:

$$\nabla_\beta l_\beta = (C - f_\beta(\mathbf{x}))\mathbf{x}$$

Estimación máximo verosimil de los parámetros

La matriz Hessiana viene dada por:

$$H = \frac{\partial}{\partial \beta} \nabla_\beta l_\beta = -\mathbf{x}^T \frac{\partial}{\partial \beta} f_\beta(\mathbf{x}) = -\mathbf{x}^T [f_\beta(\mathbf{x})(1 - f_\beta(\mathbf{x}))]\mathbf{x}$$

$$H = -\mathbf{x}^T W \mathbf{x}$$

donde W es matriz diagonal con elementos $f_\beta(x^{(j)})(1 - f_\beta(x^{(j)}))$, $j = 1, ..., N$.

Finalmente, la regla de adaptación de Newoton-Raphson viene dada por:

$$\widehat{\beta} := \widehat{\beta} + (\mathbf{x}^T W \mathbf{x})^{-1}(C - f_\beta(\mathbf{x}))\mathbf{x}.$$

El modelo logístico: Interpretación en términos de riesgo

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:

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

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,
  $\left(\dfrac{\widehat{\beta}_j}{\widehat{S}_{\beta_j}}\right)^2 \rightsquigarrow \chi^2_1$