Tema 13. Regresión Logística

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco–Euskal Herriko Unibertsitatea

1 Introducción

En este tema vamos a introducir las ideas fundamentales subyacentes a un paradigma denominado *regresión logística*. La regresión logística se ha convertido en un paradigma muy usado en Ciencias de la Salud para construir modelos predictores. La razón fundamental de esta popularización es debido al hecho de que los parámetros en los que se basa tienen una interpretación en términos de riesgo.

Este tema se ha estructurada de la manera siguiente: En la Sección 2 se introduce el modelo de regresión logística. La Seccíón 3 muestra como estimar los parámetros de los que depende el modelo a partir de sus estimaciones máximo verosímiles. La sección 4 es una interpretación del modelo logístico en términos de riesgo, asi se introducen conceptos tales como *risk ratio*, *odds ratio*, *formulación logit* y *risk odds ratio*. En las Secciones 5 y 6 se introducen respectivamente el *test de la razón de verosimilitud* y el *test de Wald* ambas herramientas útiles para el proceso de *modelización*.

2 El Modelo Logístico

Antes de introducir el modelo de regresión logística, vamos a considerar el modelo lineal de la regresión, el cual consiste en aprender una función real lineal que asigna a cada instancia un valor real. Es la principal diferencia respecto a la clasificación; el valor a predecir es real.

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). Ahora, dado un conjunto de datos de entrenamiento, la pregunta es ¿cómo aprender el vector de los parámetros β de manera que la función $g(\mathbf{x})$ aproxime la salida C?

Una manera razonable para realizar este aprendizaje, considerando una muestra de tamaño N, sería definir una función que mide, para cada valor de β, cómo está de cerca $g(x^{(i)})$ de $c^{(i)}$. Esta función se llama función de coste y viene dada por,

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

Esta función es simplemente la función de coste de mínimos cuadrados que da lugar a un modelo de regresión de mínimos cuadrados ordinarios.

Finalmente, el problema de aprendizaje se reduce a un problema de optimización ya que el objetivo es encontrar β que minimiza $J(\beta)$.

Como se ha señaldo anteriormente, el probelma de clasificación es similar al problema de regresión con la diferencia de que los valores de la variable clase que se quiere predecir sólo pueden tomar un número finito de valores discretos. Para simplificar el análisis, consideramos que la varibale clase es binaria y denotamosla por C. Podriamos abordar el problema de clasificación ignorando el hecho de que la variable clase es discreta y usar un algoritmo de regresión lineal para predecir C dada $\mathbf{x}$. Sin embargo, aúnque esto es posible matemáticamente, nos conduce a resultados absurdos. Por ello, vamos a cambiar nuestra hipotesis sobre la función $g_\beta(\mathbf{x})$, y consideramos la siguiente función:

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

la cual esta acotada entre 0 y 1, y por tanto, puede ser interpreta en términos de probabilidad. Por otra parte esta función verifica las siguientes propiedades:

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

Esta función es conocida como función logística o función sigmoide.

Entonces, asumiendo que la varible clase C es binaria, el paradigma de regresión logística se expresa de la manera siguiente:

$$\begin{aligned} P(C = 1|\mathbf{x}; \beta) &= f_\beta(\mathbf{x}) \\ P(C = 0|\mathbf{x}; \beta) &= 1 - f_\beta(\mathbf{x}) \end{aligned}$$

donde β^T es el vector de los parámetros, que deben ser estimados a partir de los datos, a fijar para tener determinado un modelo concreto de regresión logística.

3 Estimación Máximo Verosímil de los Parámetros

La estimación del vector de los parámetros $\widehat{\beta^T}$ de un modelo de regresión logística se efectúa por medio del método de estimación por máxima verosimilitud. Según dicho método se obtienen los estimadores máximo verosímiles como funciones de la muestra que hacen que se maximice la función de verosimilitud asociada a la nuestra. Denotando por $L\left((\mathbf{x}^{(1)}, c^{(1)}), \ldots, (\mathbf{x}^{(N)}, c^{(N)}), \beta_0, \beta_1, \ldots, \beta_n\right)$ a la función de verosimilitud asociada a una muestra de tamaño N, para un modelo de regresión logística con una función densidad de probabilidad dada por,

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

se tiene que:

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

Por otra parte, teniendo en cuenta que $ln(z)$ es una función creciente estrictamente, y por tanto maximizar la función L se reduce a maximizar su función logaritmo lnL. Desarrollando el logaritmo natural de la función de verosimilitud obtenemos:

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

En algúnos casos, es posible encontrar análiticamente los estimadores máximo verosímiles $\widehat{\beta}$. Pero más a menudo, es necesario recurrir a los métodos numéricos para encontrar los $\widehat{\beta}$. A continuación, se presenta brevemente uno de los métodos más utilizados, el método de Newton-Raphson. Es un método iterativo que genera una secuencia de valores $\beta_0, \beta_1, ...$ que, bajo condiciones ideales, converge hacia el estimador máximo verosímil $\widehat{\beta}$.

El método de Newton consiste en lo siguiente: dada una función real, f, encontrar β tal que $f(\beta) = 0$. $\beta \in \Re$ es un número real. Este método implementa la siguiente regla de adaptación:

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

La generalización de este método al caso multiparámetros ($\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_n)$) se conoce como el método de de Newton-Raphson. Aplicado a nuestro modelo de regresión logística viene dado por:

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

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 $\nabla_{\beta_i} l_\beta = \frac{\partial l_\beta}{\partial \beta_i}$. $i = 1, ..., n$.

Considerando un ejemplo de entrenamiento, $(\mathbf{x}^{(j)}, c^{(j)})$, el gradiente de $l_\beta(\mathbf{x}, C)$ viene dado por:

$$\frac{\partial l_\beta}{\partial \beta_i} = (\frac{c^{(j)}}{f_\beta(x^{(j)})} - \frac{(1 - c^{(j)})}{1 - f_\beta(x^{(j)})}) \frac{\partial f_\beta(x^{(j)})}{\partial \beta_i}$$

Desarrollando la derivada de la función logística ($f'(\mathbf{x}) = f(\mathbf{x})(1 - f(\mathbf{x}))$), tenemos:

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

De lo anterior, y adoptando una notación matricial, tenemos $\nabla_\beta l_\beta = (C - f_\beta(\mathbf{x}))\mathbf{x}$, de allí 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 Newton-Raphson viene dada por:

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

4 Interpretación del modelo logístico en términos de riesgo

Por ejemplo usando los datos proporcionados por Kleinbaum (1994) donde C representa la variable aletoria *Enfermedad Coronaria* con posibles valores (1 si, 0 no), X_1 *Nivel de Colesterol* (1 alto, 0 bajo), X_2 *Edad Continua*, y X_3 el *resultado del Electrocardiograma* (1 anormal, 0 normal), supongamos que obtenemos el modelo de regresión logística -obtenido a partir de $N = 609$ casos siguiente:

$$\begin{aligned}
\widehat{\beta_0} &= -3{,}911 \\
\widehat{\beta_1} &= 0{,}652 \\
\widehat{\beta_2} &= 0{,}029 \\
\widehat{\beta_3} &= 0{,}342
\end{aligned}$$

Si quisieramos comparar el riesgo para dos patrones: $\mathbf{x} = (1, 40, 0)$ y $\mathbf{x}' = (0, 40, 0)$ podríamos comenzar calculando la probabilidad de que $C = 1$ para cada uno de ellos:

$$P(C = 1|\mathbf{x}) = P(C = 1|X_1 = 1, X_2 = 40, X_3 = 0) = \frac{1}{1 + e^{-(-3{,}911 + 0{,}652(1) + 0{,}029(40) + 0{,}342(0))}} = 0{,}109$$

$$P(C = 1|\mathbf{x}') = P(C = 1|X_1 = 0, X_2 = 40, X_3 = 0) = \frac{1}{1 + e^{-(-3{,}911 + 0{,}652(0) + 0{,}029(40) + 0{,}342(0))}} = 0{,}060$$

para posteriormente utilizar el denominado *risk ratio (RR)* de $X_1 = 1$ frente a $X_1 = 0$ definido de la siguiente manera:

$$RR(\mathbf{x}, \mathbf{x}') = \frac{P(C = 1|\mathbf{x})}{P(C = 1|\mathbf{x}')} = \frac{P(C = 1|X_1 = 1, X_2 = 40, X_3 = 0)}{P(C = 1|X_1 = 0, X_2 = 40, X_3 = 0)} = \frac{0{,}109}{0{,}060} = 1{,}82$$

Es decir, 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).

Otro concepto de interés es el de *odds ratio* (OR) de un determinado patrón $\mathbf{x}$ el cual se denota por $OR(\mathbf{x})$ y se define como el cociente entre la probabilidad de que el patrón pertenezca a la clase 1 entre la probabilidad de que el patrón pertenezca a la clase 0. Es decir:

$$OR(\mathbf{x}) = \frac{P(C = 1|\mathbf{x})}{1 - P(C = 1|\mathbf{x})}$$

Para trabajar con $OR(\mathbf{x})$ de manera ágil, es conveniente expresar el modelo de regresión logística en la manera *logit*. Para ello se efectúa una transformación del modelo, de la manera siguiente:

$$logit\ (P(C = 1|\mathbf{x})) = ln\ OR(\mathbf{x}) = ln\ \left[\frac{P(C = 1|\mathbf{x})}{1 - P(C = 1|\mathbf{x})}\right]$$

Sustituyendo en la fórmula anterior las expresiones correspondientes al modelo logístico obtenemos:

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

Tal y como se ha comentado anteriormente, el *odds ratio* (OR) de un individuo con patrón $\mathbf{x}$ se define como el cociente entre la probabilidad de que $C=1$ dado dicho patrón $\mathbf{x}$ y la probabilidad de que $C=0$ dado $\mathbf{x}$. Así un *odds ratio* de $\frac{1}{3}$ para un patrón $\mathbf{x}$ se interpreta diciendo que para dicho patrón la probabilidad de que se dé $C=1$ es una tercera parte de la probabilidad de que $C=0$.

Además se tiene que $ln\ OR(\mathbf{x})) = ln\ \dfrac{P(C=1|\mathbf{x})}{1-P(C=1|\mathbf{x})} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$ y por tanto si $\mathbf{x} = (0,0,\ldots,0)$ entonces $ln\ OR(\mathbf{0})) = \beta_0$.

Siguiendo con el ejemplo anterior relativo a la enfermedad coronaria, si calculásemos el logit para $\mathbf{x} = (1,40,0)$ y para $\mathbf{x}' = (0,40,0)$ obtenemos:

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

Así como:

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

y restando ambos logit se obtiene:

$$logit\ P(C=1|\mathbf{x})) - logit\ P(C=1|\mathbf{x}')) = \beta_1$$

Veamos cómo expresar de manera genérica la constatación anterior.

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

Demostración: Sean $\mathbf{x} = (x_1,\ldots,x_i,\ldots,x_n)$ y $\mathbf{x}' = (x_1',\ldots,x_i',\ldots,x_n')$ dos patrones verificando $x_j = x_j'$ para todo $j \neq i$ y $x_i' = x_i + 1$.
Calculando el cambio en el logit obtenemos:

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

□

Otro concepto que resulta de interés es el de *risk odds ratio* (ROR) de $\mathbf{x}'$ frente a $\mathbf{x}$, el cual mide el riesgo del *odds ratio* de $\mathbf{x}'$ frente al *odds ratio* de $\mathbf{x}$ (OR($\mathbf{x}$)), es decir:

$$ROR(\mathbf{x}',\mathbf{x}) = \frac{OR(\mathbf{x}')}{OR(\mathbf{x})} = \frac{e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i'\right)}}{e^{\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right)}} = e^{\left(\sum_{i=1}^{n} \beta_i (x_i' - x_i)\right)}$$

Obviamente $ROR(\mathbf{x}', \mathbf{x})$ se puede expresar de manera alternativa como:

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

5 Test de la Razón de Verosimilitud

El test de la razón de verosimilitud se basa en comparar el producto entre -2 y el logaritmo neperiano de un cociente entre verosimilitudes con el percentil correspondiente de una distribución chi-cuadrado. Dicho test de la razón de verosimilitud tiene como objetivo el comparar dos modelos de regresión logística, el denominado *modelo completo* (*full model*) frente al que se conoce como *modelo reducido* (*reduced model*). Este segundo modelo puede verse como un submodelo del modelo completo. La hipótesis nula testada en el test de la razón de verosimilitud establece que los parámetros correspondientes a las variables que forman parte del modelo completo, pero no del modelo reducido, valen cero.

Para ver la manera en la que funciona el test de la razón de verosimilitud vamos a considerar los siguientes tres modelos de regresión logística expresados en su formulación logit:

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

Tal y como puede verse, el Modelo 2 es una extensión del Modelo 1, de igual manera que el Modelo 3 constituye una extensión del Modelo 2. En caso de querer comparar el Modelo 2 frente al Modelo 1, este último jugará el papel de modelo reducido, mientras que el Modelo 2 será el modelo completo. De manera análoga, si quisiésemos comparar el Modelo 3 frente al Modelo 2, dicho Modelo 2 sería el modelo reducido, mientras que el Modelo 3 se interpretará como modelo completo.

Vamos a denotar por $\widehat{L_1}, \widehat{L_2}$ y $\widehat{L_3}$ los valores de máxima verosimilitud obtenidos respectivamente por el Modelo 1, Modelo 2 y Modelo 3 en relación con un conjunto de N casos previamente determinado. Debido a que cuanto más parámetros tiene un modelo mejor se va ajustando a los datos, y esta es la situación existente con los tres modelos anteriores debido a sus características jerárquicas, se tiene que:

$$\widehat{L_1} \leq \widehat{L_2} \leq \widehat{L_3}$$

Pero por otra parte, al ser el logaritmo una función creciente, se obtiene que:

$$ln\widehat{L_1} \leq ln\widehat{L_2} \leq ln\widehat{L_3}$$

y por tanto

$$-2ln\widehat{L_3} \leq -2ln\widehat{L_2} \leq -2ln\widehat{L_1}$$

siendo esta la relación existente entre los denominados *log likehood statistics*, a partir de los cuales se va a construir el test de la razón de verosimilitud.

El test de la razón de verosimilitud (LR) tiene en cuenta la resta entre dos *log likehood statistics*, o lo que es lo mismo, el logaritmo neperiano del cociente entre dos verosimilitudes.

Siguiendo con el ejemplo introducido anteriormente y tratando de comparar el Modelo 2 frente al Modelo 1, el test de la razón de verosimilitud plantea como hipótesis

nula el que $\beta_3 = 0$, es decir, que el parámetro de la componente que forma parte del Modelo 2 pero no del Modelo 1 es cero. Por tanto se tiene:

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

La manera en la que funciona el test de razón de verosimilitud es la siguiente: Si la variable X_3 efectúa una gran contribución a la modelización y hace que el Modelo 2 se ajuste mucho mejor a los datos que el Modelo 1, se tendrá que $\widehat{L_2}$ será mucho mayor que $\widehat{L_1}$, y por tanto $\frac{\widehat{L_1}}{\widehat{L_2}} \simeq 0$. Tomando logaritmos neperianos, $ln\frac{\widehat{L_1}}{\widehat{L_2}} \simeq -\infty$, y de ahí que $-2ln\frac{\widehat{L_1}}{\widehat{L_2}} \simeq +\infty$. Por tanto cuanto mayor sea el valor de LR $= -2ln\frac{\widehat{L_1}}{\widehat{L_2}}$ más en contra estaremos de la hipótesis nula $H_0 : \beta_3 = 0$.

Por otra parte, si la contribución de X_3 es escasa, se tendría que $\frac{\widehat{L_1}}{\widehat{L_2}} \simeq 1$ y por tanto $ln\frac{\widehat{L_1}}{\widehat{L_2}} \simeq 0$ y finalmente LR$= -2ln\frac{\widehat{L_1}}{\widehat{L_2}} \simeq 0$.

Se demuestra teóricamente que LR $= -2ln\frac{\widehat{L_1}}{\widehat{L_2}}$ sigue bajo la hipótesis nula H_0 una distribución de probabilidad χ^2_r cuando N, número de casos en la base de datos, es suficientemente grande. El número de grados de libertad de la distribución chi-cuadrado, r, es igual al número de parámetros que en el modelo completo deben igualarse a cero para que dicho modelo completo coincida con el modelo reducido.

Nótese que LR verifica $0 \leq$ LR $< +\infty$.

6 El test de Wald

El test de Wald constituye otra manera de llevar a cabo test de hipótesis acerca de parámetros sin necesidad de usar el test de la razón de verosimilitud. Sin embargo el test de Wald tan sólo puede ser usado para testar un único parámetro, como por ejemplo ocurre al testar el Modelo 2 frente al Modelo 1. Si tratásemos de testar el Modelo 3 frente al Modelo 2, el test de Wald no sería de aplicación.

Para llevar a cabo el test de Wald hay que tener en cuenta el denominado estadístico de Wald para la variable en cuestión, en este caso denotada por X_j. Para dicha j-ésima variable dicho estadístico de Wald es $\frac{\widehat{\beta_j}}{\widehat{S_{\beta_j}}}$, siendo $\widehat{\beta_j}$ y $\widehat{S_{\beta_j}}$ las estimaciones máximo verosímiles de β_j y de su correspondiente desviación estándard.

Se verifica que $\frac{\widehat{\beta_j}}{\widehat{S_{\beta_j}}} \rightsquigarrow \mathcal{N}(0,1)$ o lo que es equivalente, $\left(\frac{\widehat{\beta_j}}{\widehat{S_{\beta_j}}}\right)^2 \rightsquigarrow \chi^2_1$. Esta distribución del estadístico de Wald sirve para aceptar o rechazar la hipótesis nula establecida sobre el j-ésimo parámetro,

$$H_0 : \beta_j = 0$$

$$H_A : \beta_j \neq 0$$

Referencias

1. A. Albert, J.A. Anderson (1984). On the Existence of Maximun Likelihood Estimates in Logistic Models. *Biometrika*, **71**, 1-10

2. R. Christensen (1997). *Log-linear Models and Logistic Regression*, Springer

3. D.W. Hosmer, S. Lemeshov (1989) *Applied Logistic Regression*, Wiley

4. D.G. Kleinbaum (1994) *Logistic Regression*, Springer

5. S. Menard (2000). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, **51**, 1, 17-24