Optimización

Primer Cuatrimestre 2016

Trabajo Práctico N° 2: Clasificación binaria por regresión logística.

Nos interesa clasificar ciertos objetos de una población, podrían ser personas, empresas, países, etc, en dos categorías, digamos Grupo 1 y Grupo 2. Contamos solo con información de ciertos miembros de la población, digamos que de n de ellos. Para ciertos miembros i, $i = 1, \ldots, n$ de la población contamos con los datos $(\mathbf{x}_i, y_i)$, donde $y_i = 0$ si el miembro i pertenece al Grupo 1, $y_i = 1$ si el miembro i pertenece al Grupo 2 y $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ es un vector donde cada coordenada es cierta variable medida en el miembro i. A la variable y se la suele llamar variable respuesta, o variable dependiente y a las variables que son las coordenadas de $\mathbf{x}$ se las suele llamar covariables, variables predictivas o variables independientes.

Un ejemplo, un tanto artifical, de la situación descripta en el párrafo anterior podría ser el siguiente. La población en cuestión podrían ser las personas con edad entre 40 y 70 años de Buenos Aires. El Grupo 1 está formado por las personas con baja presión arterial y el Grupo 2 por las personas con alta presión arterial. Tomamos una muestra al azar de toda la población de tamaño 100. A cada persona seleccionada en la muestra le medimos su presión arterial y la clasificamos en el Grupo que le corresponda. Para cada persona también registramos el número de cigarrillos que fuma por semana y si realiza o no actividad física. Estas dos últimas variables jugarían el rol de variables predictivas.

Nuestro objetivo es construir una función que usando la información en $\mathbf{x}$ nos de una predicción de y. Llamaremos a esta función regla de clasificación y la notaremos con la letra C. C es una función que a cada vector $\mathbf{x}$, le asigna un número $C(\mathbf{x}) \in \{0, 1\}$. Como no contamos con toda la información sobre la población, sino con solo la de n miembros, no aspiramos a que nuestra regla de clasificación sea perfecta, es decir que para cada miembro de la población con información asociada $(\mathbf{x}, y)$ valga que $C(\mathbf{x}) = y$, si no a tener una regla que tenga un error de clasificación bajo.

Una manera de enfocar este problema es dar un modelo estocástico para la relación entre las covariables y la variable respuesta. Supongamos que $(\mathbf{x}_i, y_i)$ $i = 1, \ldots, n$ son vectores aleatorios independientes y que $y_i|\mathbf{x}_i$ tiene distribución Bernoulli de parámetro $p_i = \mathbb{P}(y_i = 1|\mathbf{x}_i)$. El modelo logístico modela las probabilidades p_i a traves de la relación

$$p_i = p(\mathbf{x}_i) = \frac{1}{1 + exp(-\alpha_0 - \boldsymbol{\beta}_0^T \mathbf{x}_i)},$$

donde α_0 y $\boldsymbol{\beta}_0$ son parámetros a estimar. Por definición de p_i, parece natural clasificar $C(\mathbf{x}_i) = 1$ si $p_i > c$ para cierta constante c y $C(\mathbf{x}_i) = 0$ caso contrario. Claro que para poder aplicar una regla de este tipo, tendriamos que conocer α_0 y $\boldsymbol{\beta}_0$. Como estamos trabajando con la información de solo algunos miembros de la población, nos tendremos que conformar con armar nuestra regla de clasificación utilizando estimaciones $\hat{\alpha}_0$ y $\hat{\boldsymbol{\beta}}_0$ de α_0 y $\boldsymbol{\beta}_0$. Construiremos estas estimaciones a partir de una muestra $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$.

Para $(\alpha, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}$, sean para cada $i = 1, \ldots, n$

$$p_i(\alpha, \boldsymbol{\beta}) = \frac{1}{1 + exp(-\alpha - \boldsymbol{\beta}^T \mathbf{x}_i)}.$$

El estimador de máxima verosimilitud de $(\alpha_0, \boldsymbol{\beta}_0)$ está dado por

$$(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0) = \arg\max_{\mathbb{R}^{p+1}} \prod_{i=1}^{n} p_i(\alpha, \boldsymbol{\beta})^{y_i} (1 - p_i(\alpha, \boldsymbol{\beta}))^{1-y_i}.$$

Equivalentemente

$$\begin{aligned} (\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0) &= \arg\min_{\mathbb{R}^{p+1}} \sum_{i=1}^{n} -\{y_i \log(p_i(\alpha, \boldsymbol{\beta})) + \log(1 - p_i(\alpha, \boldsymbol{\beta}))(1 - y_i)\} \qquad (1) \\ &= \arg\min_{\mathbb{R}^{p+1}} \sum_{i=1}^{n} d_i, \end{aligned}$$

donde $d_i = -2\{y_i \log(p_i(\alpha, \boldsymbol{\beta})) + \log(1 - p_i(\alpha, \boldsymbol{\beta}))(1 - y_i)\}$, se llama la desviación del dato i y mide el ajuste del modelo a ese dato:

- Si $y_i = 1$, la observación pertenece al Grupo 2 y $d_i = -2\log(p_i(\alpha, \boldsymbol{\beta}))$, por lo que la observación tendrá una desviación grande si la probabilidad estimada de pertenecer al Grupo 2 es chica, o sea, cuando la observación está mal explicada por el modelo.
- Si $y_i = 0$, la observación pertenece al Grupo 1 y $d_i = -2\log(1 - p_i(\alpha, \boldsymbol{\beta}))$, por lo que la observación tendrá una desviación grande si la probabilidad estimada de pertenecer al Grupo 1 es chica, o sea, cuando la observación está mal explicada por el modelo.

Ejercicio 1 Utilizando alguno de los métodos vistos en clase, por ejemplo el de Newton, crear una función de MATLAB que tome como entrada datos $(\mathbf{x}_i, y_i)$ $i = 1, \ldots, n$ y devuelva la estimación de $(\alpha_0, \boldsymbol{\beta}_0)$ que se obtiene de resolver (1).

Para resolver el Ejercicio 1, se necesitará un punto inicial de donde comenzar las iteraciones. Típicamente se recomienda utilizar $(\alpha^0, \boldsymbol{\beta}^0) = (1/n \sum_{i=1}^{n} y_i, \mathbf{0})$.

Una vez que contamos con estimaciones $\hat{\alpha}_0$ y $\hat{\boldsymbol{\beta}}_0$, podemos construir nuestra regla de clasificación C, lo único que resta es el elegir el punto de corte c.

En el archivo credit.mat se encuentran 800 observaciones de las siguientes diez variables medidas en ciudadanos alemanes, según número de columna:

1. Edad.
2. ¿Es un trabajador extranjero? 1 = si, 0 = no.
3. ¿Es dueño de su casa? 1 = si, 0 = no.
4. ¿Tiene un trabajo calificado? 1 = si, 0 = no.
5. Monto de dinero pedido.
6. Numero de creditos con los que ya cuenta.
7. Numero de personas que dependen economicamente de la persona.
8. ¿Pide el credito para mejorar su educacion? 1 = si, 0 = no.
9. ¿Pide el credito para empezar o invertir en un negocio? 1 = si, 0 = no.

Los ciudadanos fueron clasificados además en dos categorías de mérito crediticio: 1 que corresponde a un buen mérito y 0 que corresponde a un mal mérito. Los ciudadanos con buen mérito crediticio son aquellos que pagaron sus deudas en tiempo y forma. Esta es la información que aparece en la décima columna.

Ejercicio 2 Usar el programa del Ejercicios 1 para construir un clasificador que usando las nueve variables predictivas descriptas anteriormente o cierta transformacion de cierto subconjunto de las mismas, digamos $\hat{\mathbf{x}}_i$, decida si una persona tendrá buen o mal mérito crediticio.

Cada grupo deberá elegir el punto de corte c y las variables a ser utilizadas para predecir. Una opción para hacer estas elecciones es usar alguno de los criterios vistos en clase.

El banco prestará dinero solo a aquellas personas que, según la regla de clasificacion creada en el Ejercicio anterior, vayan a tener un buen riesgo crediticio. Para los dueños del banco es más costoso prestarle dinero a alguien que luego resultará tener un mal mérito crediticio que no prestárselo a alguien que luego resultará tener un buen mérito crediticio. Para ellos, la función de costos que mide el rendimiento de cierto clasificador C sobre un conjunto de datos $(\hat{\mathbf{x}}_i, y_i)$, $i = 1, \ldots, n$ está dada por

$$\frac{1}{n}\sum_{i=1}^{n}\{2I(C(\hat{\mathbf{x}}_i) = 1, y_i = 0) + 1.1I(C(\hat{\mathbf{x}}_i) = 0, y_i = 1)\}. \tag{2}$$

Los docentes de la práctica retuvieron 200 de las 1000 observaciones originales del conjunto de datos. Los programas elaborados por cada grupo de alumnos serán utilizados para clasificar estas 200 observaciones, de las cuales se conocen el mérito crediticio verdadero. El grupo que construya el clasificador con menor error de clasificación medido segun (2) sobre este conjunto de datos, gana un premio sorpresa.