Minería de Datos, 2008-2009

Clasificación con Regresión

(Pág. 6, ejerc. 7 Estadística Descriptiva con Excel)

Dado un conjunto de datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ que son observaciones de dos variables (X, Y) con $\mathsf{X} \subset \mathbb{R}^d$ (es decir, posiblemente vectorial) y con $\mathsf{Y} \in \{0, 1\}$, los comentarios siguientes proporcionan una buena justificación y método para utilizar la regresión en la clasificación de nuevos ejemplos $\mathbf{x}$.

1. En general, la regresión de Y sobre X es la curva $m(x) = \mathbb{E}(\mathsf{Y}|\mathsf{X} = \mathbf{x})$ que proporciona el valor medio de Y cuando $\mathsf{X} = \mathbf{x}$.

   Como normalmente no tenemos acceso a toda la población (X, Y) y, en su lugar, disponemos sólo de n datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, tenemos que conformarmos con estimar esa curva $m(x)$ mediante algún método de regresión.

2. El hecho de que $\mathsf{Y} \in \{0, 1\}$ es equivalente a decir que Y es una variable aleatoria de Bernoulli. Este modelo queda totalmente especificado dando la probabilidad de uno de sus valores, por ejemplo, $p = \mathbb{P}(\mathsf{Y} = 1)$ pues la otra es $\mathbb{P}(\mathsf{Y} = 0) = 1 - p$.

   (a) Por ello, una variable de Bernoulli se denota mediante $Be(p)$ y el hecho de que Y sea una variable de Bernoulli se denota mediante $\mathsf{Y} \sim Be(p)$.

   (b) La media de esta variable aleatoria (no la media de un conjunto de sus datos sino la media *poblacional* también llamada *esperanza* –en el sentido de *esperable*) se calcula mediante una media de sus valores ponderada por sus probabilidades:

   $$\mathbb{E}(\mathsf{Y}) = 1 \cdot p + 0 \cdot p = p = \mathbb{P}(\mathsf{Y} = 1) \tag{1}$$

3. Sin embargo, en nuestro problema, la probabilidad p de que un ejemplo con $\mathsf{X} = \mathbf{x}$ pertenezca a la clase $\mathsf{Y} = 1$ puede ser diferente a la que tienen otros ejemplos con diferentes valores de X.

   Por tanto, deberíamos hablar más bien de una variable de Bernoulli $\mathsf{Y}(\mathbf{x})$ para cada $\mathbf{x}$ y de una p para cada $\mathbf{x}$, es decir: $p(\mathbf{x}) = \mathbb{P}(\mathsf{Y}(\mathbf{x}) = 1)$.

   (a) Por tanto, aplicando (1) en cada $\mathbf{x}$ se cumplirá que:

   $$\mathbb{E}(\mathsf{Y}(\mathbf{x})) = 1 \cdot p(\mathbf{x}) + 0 \cdot p(\mathbf{x}) = p(\mathbf{x}) = \mathbb{P}(\mathsf{Y}(\mathbf{x}) = 1)$$

   (b) Pero la variable $\mathsf{Y}(\mathbf{x})$ no puede ser otra que la variable Y condicionada a $\mathsf{X} = \mathbf{x}$, es decir, la variable $\mathsf{Y}|\mathsf{X} = \mathbf{x}$, por lo que lo anterior puede escribirse como:

   $$\mathbb{E}(\mathsf{Y}|\mathsf{X} = \mathbf{x}) = 1 \cdot p(\mathbf{x}) + 0 \cdot p(\mathbf{x}) = p(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = 1|\mathsf{X} = \mathbf{x})$$

   Así pues, la curva de regresión $m(\mathbf{x}) = \mathbb{E}(\mathsf{Y}|\mathsf{X} = \mathbf{x})$ nos proporciona las probabilidades $p(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = 1|\mathsf{X} = \mathbf{x})$.

4. Por tanto, si utilizamos el método de regresión con los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ obtendremos estimaciones de la media condicionada $m(\mathbf{x}) = \mathbb{E}(\mathsf{Y}|\mathsf{X} = \mathbf{x})$ que, por lo anterior, son estimaciones de $p(\mathbf{x}) = \mathbb{P}(\mathsf{Y} = 1|\mathsf{X} = \mathbf{x})$.

   Si para un $\mathbf{x}$ determinado, esa estimación es $\hat{m}(\mathbf{x}) = \hat{p}(\mathbf{x}) = \hat{\mathbb{P}}(\mathsf{Y} = 1|\mathsf{X} = \mathbf{x})$, entonces, un método fundamentado para clasificar ese ejemplo $\mathbf{x}$ sería utilizar la siguiente *regla de clasificación*:

$$g(\mathbf{x}) = \begin{cases} 1 & \text{si } \hat{p}(\mathbf{x}) = \hat{\mathbb{P}}(\mathsf{Y} = 1|\mathsf{X} = \mathbf{x}) > 0.5 \\ 0 & \text{en caso contrario} \end{cases}$$

   (a) Se dice que esta regla de clasificación emula a la *regla de Bayes* (regla que asigna cada $\mathbf{x}$ a su clase más probable, es decir la que maximiza las denominadas *probabilidades a posteriori* $\mathbb{P}(\mathsf{Y}|\mathsf{X} = \mathbf{x})$) pues utiliza su mismo procedimiento sólo que, en lugar de utilizar las verdaderas $\mathbb{P}(\mathsf{Y}|\mathsf{X} = \mathbf{x})$ (las poblacionales), utiliza unas estimaciones (las producidas por la regresión).

   (b) Por esta razón, esta regla de clasificación se denomina la regla *plug-in*: en lugar de las verdaderas probabilidades a posteriori, se *enchufan* en la regla de Bayes unas estimaciones de aquéllas.

5. Nótese que, cuanto mejor se comporte la técnica de regresión, mejor serán sus estimaciones de las probabilidades a posteriori y, por tanto, mejor será la clasificación.