Laboratorio: Modelos lineales de clasificación con R (I)

Jose Ameijeiras Alonso

En esta sesión práctica revisaremos el problema de clasificación y veremos como ajustar modelos lineales de clasificación con R. Recordamos que en un problema de clasificación se dispone de un conjunto de observaciones que pueden venir de dos o más poblaciones o clases distintas. El objetivo es clasificar una nueva observación a partir de un conjunto de variables predictoras $X = (X_1, \ldots, X_p)$. Para ello contamos con la información de la muestra de entrenamiento, que consiste en observaciones de las variables predictoras junto con la clasificación correspondiente a cada observación.

En la primera parte de esta práctica comentaremos brevemente como ajustar con R un modelo de regresión logística.

1 Ajuste de un modelo de regresión logística con R

En primer lugar veremos como ajustar un modelo de regresión logística con R.

```
> library(ISLR)
> data(Default)
```

Este conjunto de datos contiene información simulada de 10000 clientes de una entidad bancaria. El objetivo es predecir cuando un cliente incurrirá en impago de crédito de la tarjeta. Para ello podemos utilizar la información correspondiente al saldo medio mensual del cliente.

Si representamos los datos, parece razonable pensar que el saldo medio mensual puede influir en la probabilidad de que un cliente incurra en impago de crédito de la tarjeta. Veremos como ajustar un modelo de regresión logística para estudiar esa posible relación. Es decir, supondremos:

$$p(X) = \mathbb{P}(Y = 1/X) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 X)}}$$

donde X representa el saldo mensual del cliente e

$$Y = \begin{cases} 1, & \text{si el cliente incurre en impago,} \\ 0, & \text{si el cliente no incurre en impago.} \end{cases}$$

El modelo de regresión logística se ajusta en R utilizando el comando `glm` (general linear models). Los modelos lineales generalizados son una extensión de los modelos lineales que permiten que la variable dependiente tenga una distribución no normal. La formulación de modelos lineales generalizados permite unificar en un mismo modelo métodos como la regresión lineal y la regresión logística, sin más que especificar la función link o familia de distribución de los errores correspondiente a cada caso.

Por ejemplo, la regresión logística se puede formular como un modelo lineal generalizado en el que la distribución de los errores es binomial (`family="binomial"`)

```
> fit <- glm(default ~ balance, data = Default, family = "binomial")
```

Al igual que se hacía en el análisis de regresión lineal, podremos utilizar las funciones `coef`, `summary`, `residuals`, etc. para obtener información relacionada con el ajuste del modelo. También se puede usar la función `predict` para obtener las predicciones del modelo. Si queremos obtener las predicciones para $\mathbb{P}(Y = 1/X)$, debemos añadir el argumento `type="response"`. Así:

```
> plot(Default$balance, predict(fit, type = "response"))
```

representa gráficamente las predicciones del modelo para los valores de X de la muestra de entrenamiento. Clasificaremos en el grupo de impago a aquellos clientes para los cuales la predicción obtenida para $\mathbb{P}(Y = 1/X)$ sea superior a 0.5.