Laboratorio: Modelos lineales de clasificación con R

Beatriz Pateiro López

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 y como llevar a cabo un Análisis Lineal Discriminante. En la segunda parte de la práctica, se propone un pequeño estudio de simulación para comparar el comportamiento de ambos métodos en distintos escenarios.

1 Ajuste de un modelo de regresión logística con R

En primer lugar veremos como ajustar un modelo de regresión logística con R. Trabajaremos con los mismo datos que se analizaron en la sesión teórica.

```
> 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 como hemos hecho en la sesión de teoría, 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"))
```

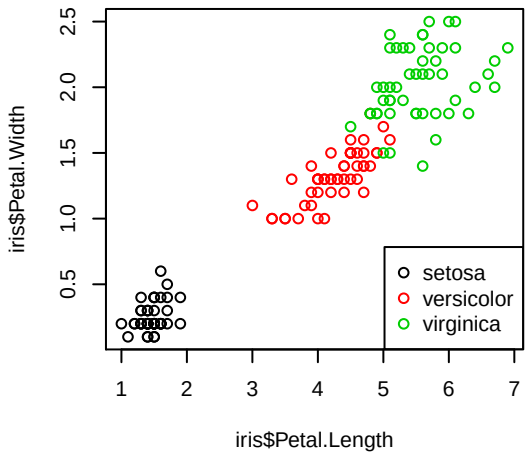
representa gráficamente las predicciones del modelo para los valores de X de la muestra de entrenamiento. Clasificaremos en el grupo de impago a aquellos clientes para los cuales la predicción obtenida para $\mathbb{P}(Y = 1/X)$ sea superior a 0.5.

2 Análisis Lineal Discriminante con R

Ilustraremos el problema de clasificación mediante Análisis Lineal Discriminante con el conjunto clásico de datos de Iris. Este conjunto nos da la medida en cm. de las variables longitud y anchura de sépalo y longitud y anchura de pétalo para un total de 150 flores de tres especies diferentes de iris (iris setosa, versicolor y virginica).

Para comenzar nos centraremos en las variables longitud y anchura de pétalo. Ya sabemos como representar gráficamente los datos correspondientes a estas variables.

```
> data(iris)
> plot(iris$Petal.Length, iris$Petal.Width, col = iris$Species)
> legend("bottomright", levels(iris$Species), pch = 1, col = 1:3)
```



Para llevar a cabo un Análisis Lineal Discriminante en R utilizaremos la función `lda`, que pertence a la librería `MASS`.

```
> library(MASS)
> lda.fit <- lda(Species ~ Petal.Length + Petal.Width, data = iris)
```

Recuerda que el Análisis Lineal Discriminante es un modelo generativo, es decir, calcula la probabilidad a posteriori $\mathbb{P}(Y = k/X = x)$ mediante el teorema de Bayes:

$$\mathbb{P}(Y = k/X = x) = \frac{\mathbb{P}(X = x/Y = k)\mathbb{P}(Y = k)}{\mathbb{P}(X = x)} = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

En la expresión anterior π_k denota la probabilidad a priori de que una observación provenga de la clase k, es decir, $\pi_k = \mathbb{P}(Y = k)$. Por otro lado, $f_k(x) = \mathbb{P}(X = x/Y = k)$ representa la densidad de probabilidad de X en la clase k.

En Análisis Lineal Discriminante se asume además que las funciones de densidad de probabilidad de cada clase $f_k(x)$ son distribuciones Normales de media μ_k y con la misma matriz de covarianzas Σ para $k = 1, \ldots, K$. Una vez calculadas las probabilidades a posteriori, la regla de clasificación consiste en asignar cada observación a la clase para la cual $\mathbb{P}(Y = k/X = x)$ es mayor.

En la práctica, para llevar a cabo la clasificación, tendremos que estimar las probabilidades a priori π_k, así como los parámetros de la densidad de probabilidad Normal correspondiente a cada clase.

La salida de la función `lda` nos muestra, entre otras cosas, las estimaciones de π_k. En este caso,

```
> lda.fit$prior
```

```
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
```

Es decir, como en este caso hay 50 observaciones dentro de cada especie, se tiene que 1/3 de las observaciones pertenecen a la especie setosa, 1/3 de las observaciones pertenecen a la especie versicolor y 1/3 de las observaciones pertenecen a la especie virginica ($\hat{\pi}_k = 1/3, k = 1, 2, 3$).

También se muestran en la salida la longitud y anchura de pétalo media de cada especie (estimaciones de μ_k):

```
> lda.fit$means
```

```
##            Petal.Length Petal.Width
## setosa            1.462       0.246
## versicolor        4.260       1.326
## virginica         5.552       2.026
```

La predicción se lleva a cabo como es habitual con la función `predict`. A continuación evaluamos la función `predict` en la muestra de entrenamiento

```
> lda.pred <- predict(lda.fit)
```

Observa que en `lda.pred$class` se indica la especie asignada a cada observación por la regla de clasificación. Por otro lado, en `lda.pred$posterior` se muestra, para cada observación, el valor estimado para la probabilidad a posteriori de cada especie.

A continuación se muestra un resumen del resultado de la clasificación en la muestra de entrenamiento. Observa que todas las flores de la especie setosa han sido correctamente clasificadas, se han cometido 4 errores de clasificación en las de la especie versicolor y 2 errores de clasificación en las de la especie virginica. En resumen tenemos una tasa de error de 0.04.

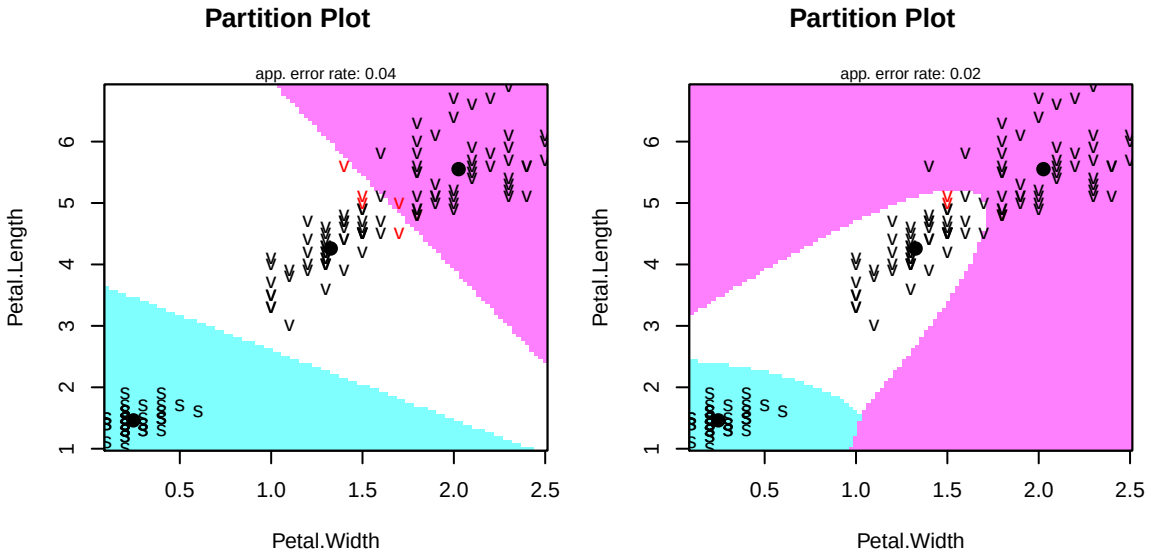
```
> table(lda.pred$class, iris$Species)
```

```
##
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         4
##   virginica       0          2        46
```

Debemos recordar que la regla de decisión que determina el Análisis Lineal Discriminante se basa en la suposición de la normalidad de las observaciones y en la de que las matrices de covarianzas en las clases son iguales. Si mantenemos que la densidad de probabilidad de cada clase es normal pero no podemos asumir la igualdad de las matrices de covarianzas, entonces la regla de decisión deja de dar lugar a un modelo de clasificación lineal. Estaremos en ese caso ante un Análisis Cuadrático Discriminante (QDA). Para llevar a cabo un Análisis Cuadrático Discriminante en R utilizaremos la función `qda`, que pertence a la librería `MASS`.

Por último, como en el ejemplo que hemos utilizado a lo largo de esta sección teníamos únicamente dos variables predictoras, podríamos visualizar las regiones determinadas por la regla de decisión. Para ello usaremos la librería `klaR`, que también nos permite llevar a cabo un Análisis Lineal Discriminante y ofrece más herramientas de visualización. Visualizamos al mismo tiempo los resultados de un Análisis Cuadrático Discriminante

```
> library(klaR)
> partimat(Species ~ Petal.Length + Petal.Width, data = iris, method = "lda")
> partimat(Species ~ Petal.Length + Petal.Width, data = iris, method = "qda")
```



En el gráfico se observa que las fronteras de las tres zonas de clasificación con LDA están delimitadas por rectas (es un modelo de clasificación lineal). También podemos ver señaladas en rojo las 4 observaciones de la muestra de entrenamiento que han resultado mal clasificadas con LDA.

Repite el Análisis Lineal Discriminante para los datos de iris utilizando todas las variables predictoras disponibles en el conjunto de dato y analiza los resultados obtenidos.

3 Comparación de los métodos de clasificación

En esta sección, vamos a comparar los resultados obtenidos por el método de regresión logística, Análisis Lineal Discriminante y Análisis Cuadrático Discriminante en distintos escenarios.

Para cada uno de los escenarios que se describen a continuación simula 100 muestras de entrenamiento, ajusta cada uno de los tres métodos de clasificación citados en cada muestra y evalua los resultados de cada uno de ellos calculando la tasa de error en una muestra test. Puedes representar los resultados obtenidos mediante un diagrama de cajas. En todos los escenarios se utilizan $p = 2$ variables predictoras y dos posibles clases ($K = 2$).

- **Escenario 1**: Genera una muestra de entrenamiento con 20 observaciones en cada una de las clases. Las observaciones dentro de cada clase son normales con una media diferente en cada clase (por ejemplo, $\mu_1 = (0, 0)^t$ y $\mu_2 = (1, 1)^t$).
- **Escenario 2**: Exactamente igual que en el escenario 1, pero la correlación entre las variables X_1 y X_2 es $\rho = -0.5$.
- **Escenario 3**: Generamos observaciones de X_1 y X_2 siguiendo una distribución t-Student con 50 observaciones en cada clase (función `rt`) y de forma que $\mu_1 = (0, 0)^t$ y $\mu_2 = (1, 1)^t$. Las curvas de la densidad de una distribución t-Student son simétricas y con forma similar a la distribución normal estándar. Sin embargo, las colas de las distribución t-Student disminuye más lentamente a cero que las colas de la distribución normal (para grados de libertad pequeños).
- **Escenario 4**: Generamos observaciones de X_1 y X_2 siguiendo una distribución normal de forma que la correlación entre las variables X_1 y X_2 es $\rho = 0.5$ en la clase 1 y $\rho = -0.5$ en la clase 2 (esto se corresponde al caso en que las dentro de cada clase la distribución es normal pero con matrices de covarianzas distintas).

Puedes ver los resultados del estudio completo (con más escenarios y comparando más métodos de clasificación) en el Capítulo 4 del libro *An Introduction to Statistical Learning with Applications in R*.