Laboratorio: Modelos lineales de clasificación con R (II)

Jose Ameijeiras Alonso

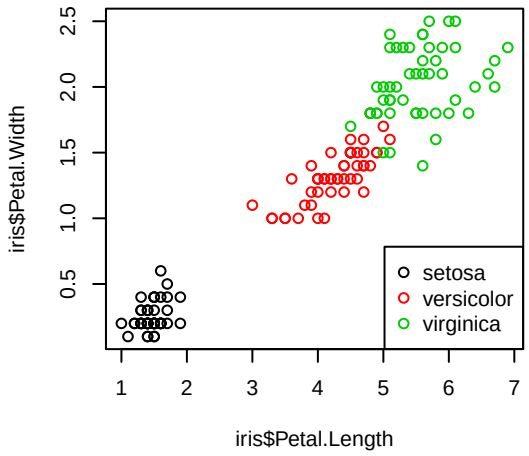
En esta sesión práctica revisaremos el problema de clasificación y veremos como ajustar modelos lineales de clasificación con R. Recordamos que en un problema de clasificación se dispone de un conjunto de observaciones que pueden venir de dos o más poblaciones o clases distintas. El objetivo es clasificar 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 esta práctica se llevará a cabo una clasificación por k vecinos más próximos y un Análisis Lineal/Cuadrático Discriminante.

1 k-vecinos más próximos con R

Ilustraremos el problema de clasificación 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 k-vecinos más próximos en R utilizaremos la función `knn`, que pertence a la librería `class`.

Utilizando la distancia Euclidea clasificaremos a una nueva observación en función de los k puntos que están más cerca de él. Podríamos ver que $k = 4$ puntos quedan más cerca del punto que tiene un `Petal.length=4` y un `Petal.width=1`. Si añadimos el argumento `prob=TRUE` también nos devolverá la probabilidad de pertenece a cada grupo.

```
> library(class)
> matexp <- cbind(iris$Petal.Length, iris$Petal.Width)
> knn.pred <- knn(train=matexp, test = c(4,1), cl = iris$Species,  k=4,  prob = T)
```

En este caso, vemos que a esa nueva planta la clasificaría como `versicolor` con una probabilidad 1 (sus 4 vecinos más próximos también son versicolor).

2 Análisis Lineal Discriminante con R

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.

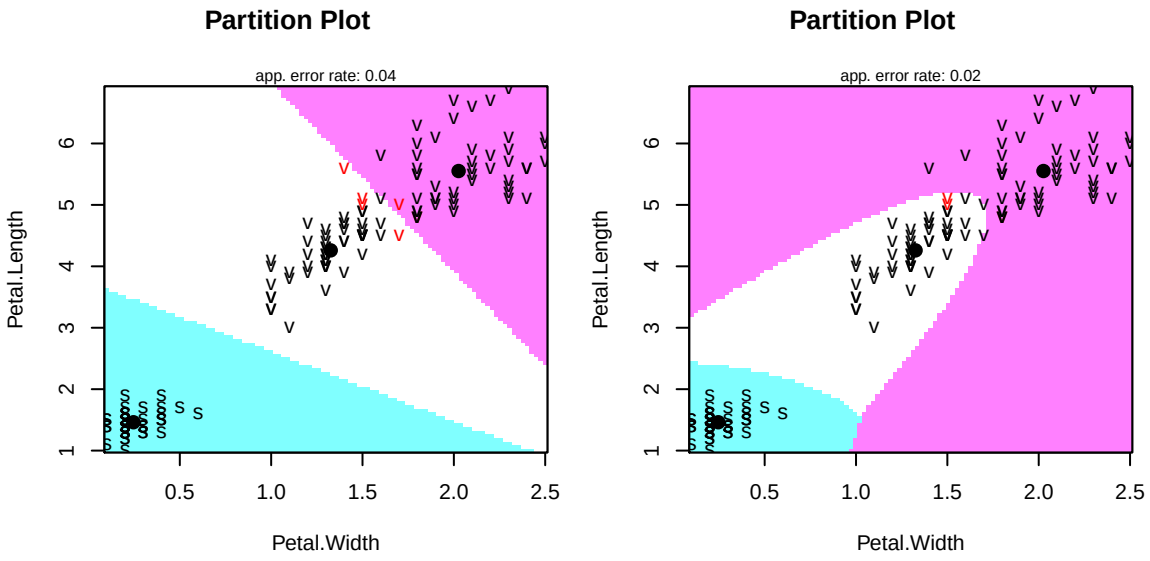
```
> table(lda.pred$class, iris$Species)
```

```
##
##              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 los métodos de clasificación, k-vecinos más próximos, el Análisis Lineal Discriminante y el Análisis Cuadrático Discriminante para los datos de iris utilizando todas las variables predictoras disponibles en el conjunto de datos y analiza los resultados obtenidos.