UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Evaluación con curvas ROC del análisis discriminante

Máster de Bioinformática y Bioestadística

Francesc Carmona

4 de junio de 2020

```
if (!(require(faraway))) install.packages("faraway", dep=TRUE)
if (!(require(ROCR))) install.packages("ROCR", dep=TRUE)
```

Ejercicio 6

Utilizar las curvas ROC para evaluar el análisis lineal discriminante.

Cargamos los datos y ajustamos el modelo para el grupo de entrenamiento.

```
library(faraway)
library(MASS)
data(wbca, package="faraway")
attach(wbca)
set.seed(123)
train <- sample(1:nrow(wbca), 400)
lda.w2 <- lda(wbca[,-1], Class, subset=train)
```

Calculamos las predicciones para el grupo de test.

```
pred.test <- predict(lda.w2, newdata=wbca[-train,-1])
```

y de aquí nos interesan las probabilidades a posteriori.

```
pred.test.posterior <- as.data.frame(pred.test$posterior[,2])
```

Las curvas ROC (*receiver operating characteristic*) dibujan la tasa de verdaderos positivos (sensibilidad) en el eje vertical frente a la tasa de falsos positivos (1 - especificidad) en el eje horizontal para todos los valores de corte (*threshold*) posibles. Estas curvas permiten evaluar la calidad de la clasificación en función del punto de corte y, entre otras cosas, hallar el óptimo.

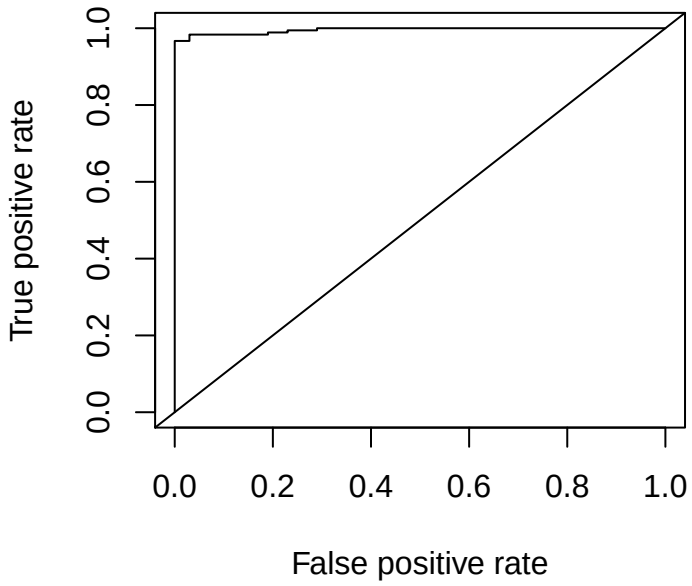
Ahora vamos a preparar los elementos necesarios para dibujar la curva ROC.

En primer lugar preparamos los datos en un objeto `prediction` y calculamos la curva ROC con `measure = "tpr"` (*true positive rate*) y `x.measure = "fpr"` (*false positive rate*). El objeto `prediction` está formado por un conjunto de *slots* y cada uno de ellos es una lista.

```
library(ROCR)
prediction.test <- prediction(pred.test.posterior, Class[-train])
roc.perf <- performance(prediction.test, measure = "tpr", x.measure = "fpr")
```

Con este objeto ya se puede dibujar la curva.

```
plot(roc.perf)
abline(a=0, b=1)
```



Una medida de la calidad de la clasificación es el área bajo la curva (AUC).

```
auc.test <- performance(prediction.test, measure = "auc")
(auc.test.value <- as.numeric(auc.test@y.values))
```

```
## [1] 0.9955801
```

Un clasificador aleatorio tiene un área bajo la curva de 0.5, mientras que el AUC para un clasificador perfecto sería 1. Aquí el AUC es 0.9956.

Otro elemento importante es el *punto de corte óptimo*. Se trata del punto donde la curva tiene la distancia mínima al punto $(0, 1)$, ya que este punto tiene un 100 % True Positive Rate y un 0 % False Positive Rate, lo que significa que tiene una separación/predicción perfecta.

```
cost.perf <- performance(prediction.test, measure = "cost")
# coordenadas del punto de corte óptimo
x <- roc.perf@x.values[[1]][ which.min(cost.perf@y.values[[1]]) ]
y <- roc.perf@y.values[[1]][ which.min(cost.perf@y.values[[1]]) ]
c(fpr.opt=round(x,3), tpr.opt=round(y,3))
```

```
## fpr.opt tpr.opt
##   0.000   0.967
```

También podemos calcular los valores de sensibilidad y especificidad de la tabla de confusión[1] del método LDA para el grupo test.

[1] La función `confusionMatrix()` del paquete `caret` calcula estos valores y otros muchos más.

		Realidad	
		Enfermos (P)	Sanos (N)
Predicción	Enfermos (P)	TP	FP Error tipo I
	Sanos (N)	FN Error tipo II	TN

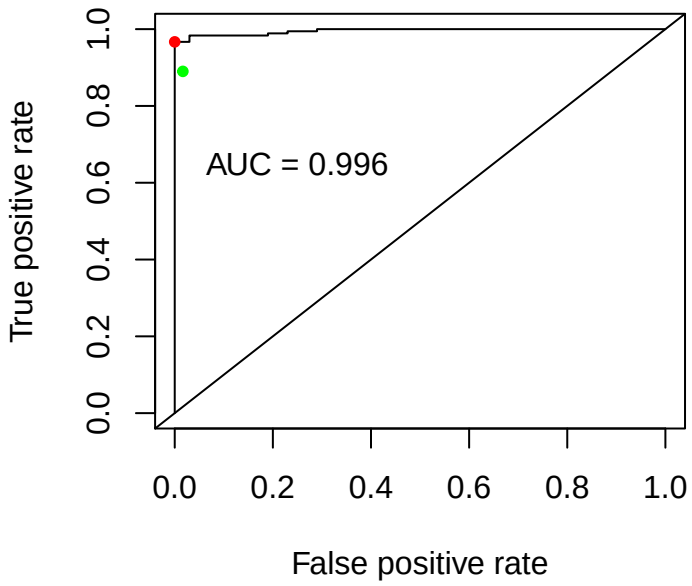
```
# Class == 0 (maligno)
# Class == 1 (benigno)
(tt <- table(pred=pred.test$class,true=Class[-train]))
```

```
##     true
## pred   0   1
##    0  89   3
##    1  11 178
```

```
TP <- tt[1,1]
FP <- tt[1,2]
FN <- tt[2,1]
TN <- tt[2,2]
sensibilidad <- TP /(FN + TP)  # 0.89
especificidad <- TN /(FP + TN) # 0.9834254
```

Ahora dibujamos los resultados.

```
plot(roc.perf)
abline(a=0, b=1)
points(x, y, pch=20, col="red") # punto óptimo
points(1-especificidad, sensibilidad, pch=20, col="green") # LDA test
text(x = .25, y = .65 ,paste("AUC = ", round(auc.test.value,3), sep = ""))
```



Para el grupo de test, el LDA da un 1.7 % de falsos positivos (mujeres con un tumor maligno a las que se predice un tumor benigno) y un 89 % de verdaderos positivos, es decir, un 11 % de mujeres a las que se les puede dar el susto de tener un tumor maligno cuando en realidad es benigno.