Distrofia Muscular de Duchenne

Jaqueline Girabel

La *Distrofia Muscular Duchenne* en una enfermedad transmitida genéticamente de una madre a sus hijos: las mujeres descendientes afectadas usualmente no presentan síntomas y pueden ser portadoras "silenciosas" de la enfermedad; mientras que los hombres afectados mueren a una temprana edad.

Aunque quienes son portadoras en general no presentan síntomas físicos, ellas tienden a presentar niveles elevados en determinadas enzimas o proteínas, como lo son la Creatina Quinasa (CK), Hemopexina (H), Lactato deshidrogenasa (LD) o Piruvato Quinasa (PK).

En este trabajo se pretende ajustar un modelo predictivo sobre el set de datos **DMD**, en el que se encuentran los resultados del diagnóstico de la Distrofia Muscular de Duchenne en función de un conjunto de variables explicativas. De estas variables explicativas, cuatro de ellas corresponden a los niveles obtenidos en los cuatro tipos de enzimas características de la *DMD* mencionados anteriormente.

```
##   Edad Mes Año CK     H   PK  LD Diagnostico
## 1   22   6  79 52  83.5 10.9 176      Normal
## 2   32   8  78 20  77.0 11.0 200      Normal
## 3   36   7  78 28  86.5 13.2 171      Normal
## 4   22  11  79 30 104.0 22.6 230      Normal
## 5   23   1  78 40  83.0 15.2 205      Normal
## 6   30   5  79 24  78.8  9.6 151      Normal
```

Observamos valores ausentes correspondientes a las enzimas PK y LD, provenientes de un total de 15 observaciones que serán descartadas.

```
#Descartaremos las filas con valores ausentes
head(DMD[PK==-9999 | LD==-9999,])
```

```
##    Edad Mes Año CK     H   PK    LD Diagnostico
## 13   25  10  77 41  87.3 15.0 -9999      Normal
## 26   31  11  77 29  94.0 11.8 -9999      Normal
## 50   22  12  77 22  85.5 15.0 -9999      Normal
## 59   27  10  77 22  99.0 10.8 -9999      Normal
## 61   26  12  77 28  93.5  7.0 -9999      Normal
## 83   38  12  77 45 108.0 13.7 -9999      Normal
```

En el siguiente gráfico se resumen las distribuciones de los tipos de enzimas por grupo (“Normal” y “Portadora”), así como también la dispersión y ambas matrices de correlación.

```
library(GGally)
ggpairs(DMD[,4:8], aes(colour = Diagnostico))
```

El objetivo es modelar un clasificador que prediga si una mujer es portadora de la DMD o no, en funciÃ³n de los niveles en las cuatro enzimas características.

Lo primero que hacemos es separar un 70% de los datos en un subset de entrenamiento **DMD.T** con el que será ajustado el modelo y el 30% restante, **DMD.V**, estará reservado para testeo.

```
set.seed(1234)
indices<-sort(sample(1:dim(DMD)[1], round(dim(DMD)[1]*0.7, 0 ), replace = FALSE))
DMD.V<-DMD[-indices,]
DMD.T<-DMD[indices,]
```

Proponemos, en principio, un ajuste lineal en los datos con un modelo de regresión logística.

```
modelo.glm <- glm( Diagnostico ~ CK+H+PK+LD, data =DMD.T , family = binomial )
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

El mensaje que se imprime en pantalla nos advierte que un subconjunto de los datos se predice perfectamente, de manera que las estimaciones de los coeficientes del modelo no son únicas (dado que, teóricamente, tienden a infinito). Cuando se detecta esta condición, el algoritmo que optimiza los coeficientes finaliza el proceso de iteración de máxima verosimilitud e informa los resultados de la última iteración. Este problema puede deberse a la falta de datos para ajustar correctamente el modelo.

Si queremos predecir con este modelo.glm la probabilidad de pertenecer a la clase "Portadora" sobre el dataset DMD.V, y asociamos estas predicciones a la clase positiva con el umbral 0.5, el error obtenido posiblemente no sea un buen indicador de la realidad.

```
modelo.glm.probas<-predict(modelo.glm ,type ="response", DMD.V)


modelo.glm.predicciones<-rep("Normal", length(modelo.glm.probas))
modelo.glm.predicciones[modelo.glm.probas>0.5]<-"Portadora"
modelo.glm.predicciones<-as.factor(modelo.glm.predicciones)
```

```
mean(modelo.glm.predicciones != DMD.V$Diagnostico)
```

```
## [1] 0.0862069
```

Recurrimos, entonces, a técnicas de análisis discriminante para modelar este clasificador. Proponemos un modelo lineal **modelo.ADL** con el que se predice el tipo de diagnóstico en el conjunto de datos DMD.V:

```
library(MASS)

modelo.ADL <-lda (Diagnostico~ CK+H+PK+LD ,data= DMD , subset = indices)

modelo.ADL.predicciones<-predict(modelo.ADL, DMD.V)

predicciones<-modelo.ADL.predicciones$class
```

```
table(predicciones, DMD.V$Diagnostico)
```

```
##
## predicciones Normal Portadora
##    Normal        39         6
##    Portadora      1        12
```

De la tabla anterior se puede ver que este modelo clasifica en la clase positiva (que en este caso fue asociada a la etiqueta "Normal") siendo 0.5 el punto de corte. En efecto:

```
sum(modelo.ADL.predicciones$posterior[,1]>0.5)
```

```
## [1] 45
```

El error de predicción obtenido es

```
error<-round(mean(modelo.ADL.predicciones$class != DMD.V$Diagnostico),4)
error
```

```
## [1] 0.1207
```

Comparemos esto con los resultados obtenidos en un modelo cuadrático.

```
modelo.ADQ <-qda (Diagnostico~ CK+H+PK+LD ,data= DMD.T)

modelo.ADQ.predicciones<-predict(modelo.ADQ, DMD.V)
```

El error obtenido en este caso es

```
round(mean(modelo.ADQ.predicciones$class != DMD.V$Diagnostico),4)
```

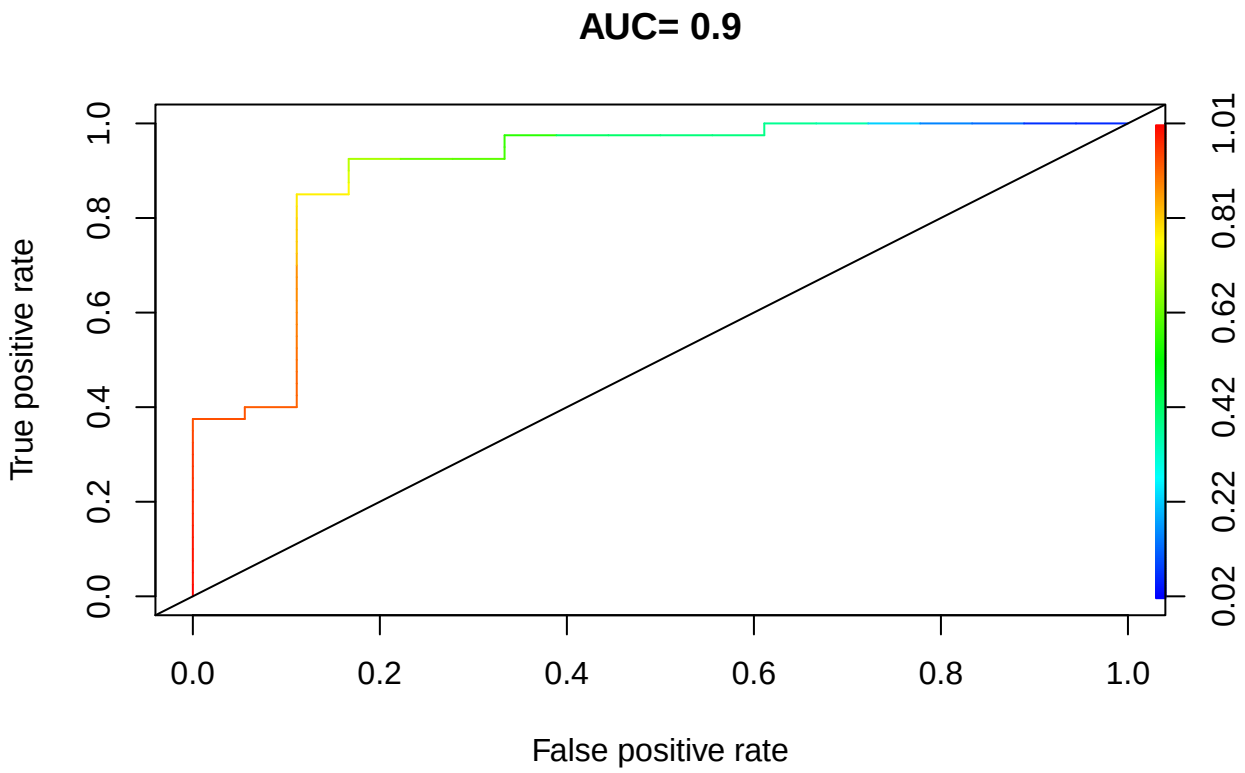
```
## [1] 0.1552
```

El mejor error de predicción se obtiene con el modelo.ADL, con el que trabajaremos en adelante.

Sabemos que nuestro modelo clasifica en la clase “Normal” (clase positiva) tomando 0.5 como punto de corte, pero queremos identificar ahora un umbral con el que se logre un balance entre los indicadores de sensibilidad y especificidad. Recordamos que una tasa alta de sensibilidad significa que prácticamente ningún caso positivo fue mal clasificado, pero posiblemente se haya clasificado mal a algunos de la otra clase. Lo ideal es, entonces, que la capacidad de detectar casos negativos (especificidad) tambiÃ©n sea optimizada.

En el siguiente gráfico podemos ver la performance del modelo ADL sobre una grilla de umbrales.

```
library(ROCR)
predict.roc<-prediction(modelo.ADL.predicciones$posterior[,1], DMD.V$Diagnostico,
                        label.ordering=c("Portadora", "Normal"))
perf.roc<-performance(predict.roc, "tpr", "fpr")
auc.diag<-as.numeric(performance(predict.roc, "auc")@y.values)


plot(perf.roc, colorize=TRUE, type="l", main=paste("AUC=", round(auc.diag,2)))
abline(a=0, b=1)
```

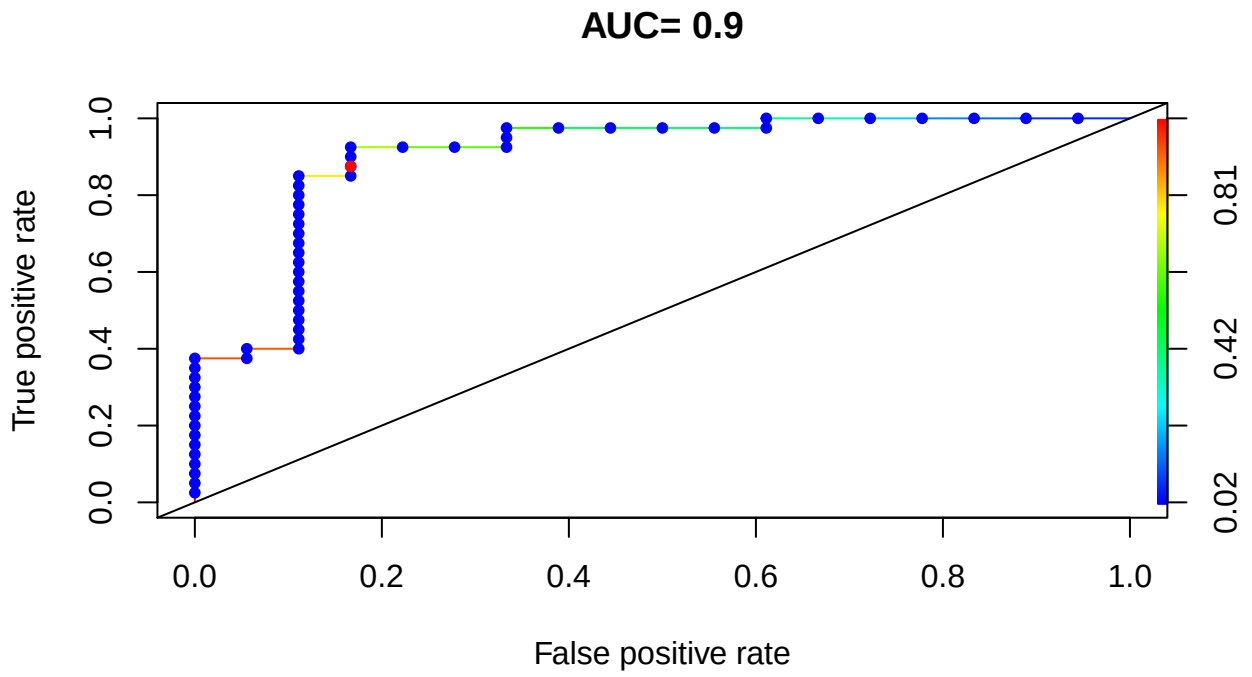


El punto de la curva ROC más cercano al punto $(0, 1)$ se corresponde con el umbral que optimiza los indicadores de sensibilidad y especificidad simultaneamente.

```
especificidad<-function(umbral){
  clasificador<-rep("Portadora", length(DMD.V$Diagnostico))
  clasificador[modelo.ADL.predicciones$posterior[,1] > umbral]<-"Normal"
  clasificador<-as.factor(clasificador)
  tabla<-table(clasificador, DMD.V$Diagnostico)
  return(tabla[4]/(tabla[4]+tabla[3]))
}



sensibilidad<-function(umbral){
  clasificador<-rep("Portadora", length(DMD.V$Diagnostico))
  clasificador[modelo.ADL.predicciones$posterior[,1] > umbral]<-"Normal"
  clasificador<-as.factor(clasificador)
  tabla<-table(clasificador, DMD.V$Diagnostico)
  return(tabla[1]/(tabla[1]+tabla[2]))
}



distancias<-function(probas){
  distancias<-c()
  for (p in 1:length(probas)) {
    distancia<-sqrt(sum((c(1-especificidad(probas[p]),
                           sensibilidad(probas[p]))-c(0,1))^2))
    distancias[p]<-distancia
  }
  return(distancias)
}


#Tomamos la lista de umbrales que proporcionan los datos perf.roc
alphas<-perf.roc@alpha.values[[1]]

distancias<-distancias(alphas)



s<-c()
for (p in 1:length(alphas)) {
  s[p]<-sensibilidad(alphas[p])
}

ne<-c()
for (p in 1:length(alphas)) {
  ne[p]<-1-especificidad(alphas[p])
}
```

```
# Superponemos a la curva ROC los puntos asociados a cada umbral

plot(perf.roc, colorize=TRUE, type="l", main=paste("AUC=", round(auc.diag,2)))
points(ne,s, col="blue", pch=20)
points(ne[40],s[40], col="red", pch=20)
```

```
abline(a=0, b=1)
```



El umbral asociado al punto de la curva más cercano al extremo izquierdo superior (en el gráfico, en color rojo) es 0.7. Calculamos los valores de sensibilidad y especificidad correspondientes a ambos puntos de corte, 0.5 y 0.7 :

```
sensibilidad(0.5)
```

```
## [1] 0.975
```

```
especificidad(0.5)
```

```
## [1] 0.6666667
```

```
sensibilidad(0.7)
```

```
## [1] 0.9
```

```
especificidad(0.7)
```

```
## [1] 0.8333333
```

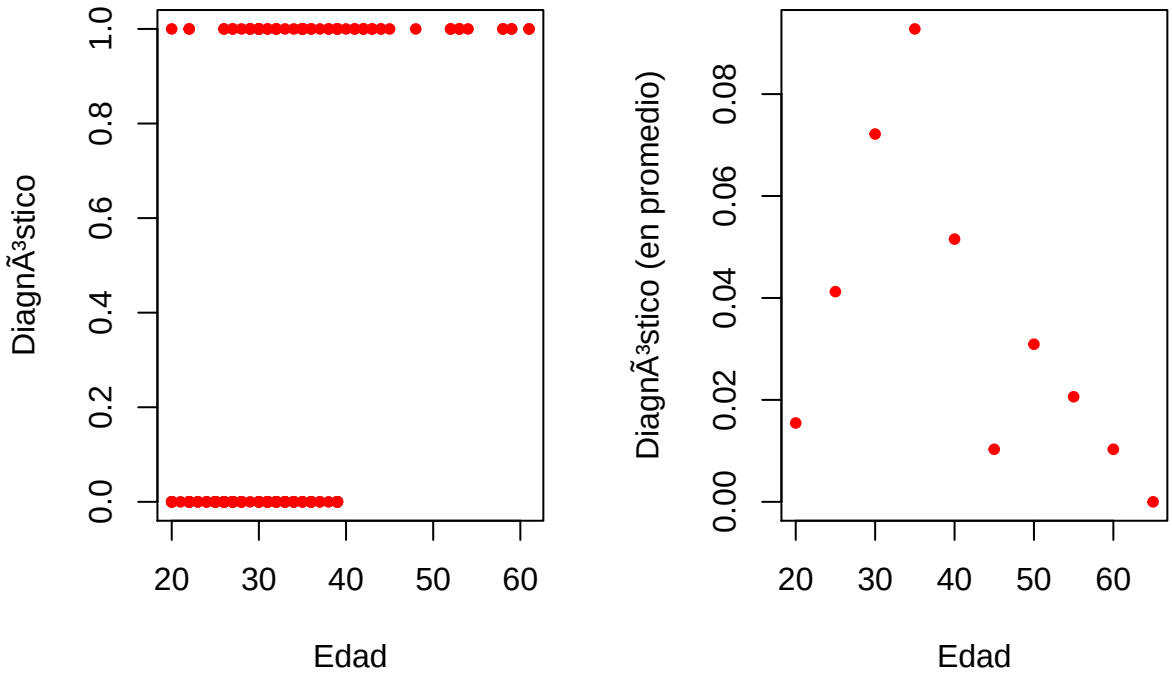
Ajustamos el clasificador del modelo ADL con punto de corte 0.7, y vemos que el error de predicción no sufre modificaciones notables:

```
modelo.predicciones<-rep("Portadora", length(DMD.V$Diagnostico))
modelo.predicciones[modelo.ADL.predicciones$posterior[,1]>0.7]<-"Normal"
modelo.predicciones<-as.factor(modelo.predicciones)
error<-round(mean(modelo.predicciones != DMD.V$Diagnostico),4)

error
```

```
## [1] 0.1207
```

Una variable que no fue tenida en cuenta para determinar un diagnótico es Edad. ¿Qué efecto tiene la edad sobre la variable respuesta? Vemos en el gráfico de la derecha a continuación los promedios de mujeres que resultan portadoras en función de su edad medida en intervalos de 5 años, basado en el dataset DMD.

```
#range(Edad)
promedios.diag<-c()
edades<-seq(from=20, to=65, by=5)
for (i in 1:length(edades)) {
  promedios.diag[i]<-mean(Diagnostico=="Portadora" & Edad>=edades[i] & Edad<edades[i+1])
}

par(mfrow=c(1,2))
plot(Edad, Diagnostico=="Portadora", xlab="Edad", ylab="DiagnÃ³stico", pch=20, col="Red")
plot(edades, promedios.diag, xlab="Edad", ylab="DiagnÃ³stico (en promedio)", pch=20, col="Red")
```



Si bien notamos que los promedios más altos provienen de mujeres entre 30 y 40 años, no hay una tendencia definida considerando al diagnóstico (en promedio) en función de la edad. Pero, además, se sabe que los niveles en las enzimas posiblemente dependan de la edad, por lo que la información relevante proveniente de esta última variable ya es codificada por los indicadores de valores enzimáticos, que son los ultilizados para modelar un predictor del diagóstico de la DMD.

De las cuatro enzimas características que disponemos en el dataset **DMD**, sabemos que las primeras dos (CK y H) pueden medirse de forma bastante económica respecto de los segundos dos marcadores (PK y LD). Queremos ver cómo varía la performance del modelo.ADL si se prescinde de estas últimas, en términos del error de predicción.

```
modelo <-lda (Diagnostico~ CK+H ,data= DMD, subset=indices)
modelo.predicciones<-predict(modelo, DMD.V)

nuevo_error<-round(mean(modelo.predicciones$class != DMD.V$Diagnostico),4)
nuevo_error
```

```
## [1] 0.2931
```

Si bien el error excede el doble del error obtenido antes de descartar las últimas dos variables, ambos valores pueden interpretarse como una buena performance del modelo discriminante lineal a la hora de predecir un diagnóstico.

Para terminar, se sabe que la probabilidad de que una mujer sea portadora es conocida, tomando el valor 1/3200, pero no es este un dato que se haya tenido en cuenta al momento de modelar un clasificador. En este caso es conveniente basarnos en la estimación de dicha probabilidad: si se implementase un clasificador con esta "verdadera" probabilidad a priori, es claro que prácticamente ninguna mujer sería clasificada como portadora, lo cual nos alejaría bastante de la realidad.