Universidad de la República **Estadística Multivariada Computacional**
Facultad de Ingeniería - IMERL **Primer Semestre 2020**

Práctico 7: Regresión logística

1. *a*) Probar que la curva ROC de un clasificador binario es no decreciente.

   b) Explicar cuando puede ocurrir y que se debería hacer si el AUC de un clasificador es menor a 0,5.

2. Sean X_1 y X_2 dos variables uniformes en $[-4, 5]$ e Y una variable que se quiere predecir a partir de ellas.

   a) Simular una relación entre Y y (X_1, X_2) del tipo:

   ```
   n=100; a=-2; b=2; c=3
   x1=runif(n,-4,5); x2=runif(n,-4,5)
   y=exp(a*x1+b*x2+c + rnorm(n))
   y=y/(1+y) ; y=rbinom(n,1,y)
   ```

   b) Representar graficamente la nube de puntos formada por las variables explicativas, representando los puntos con colores distintos según la modalidad de Y. Representar Y en función de X_1 e Y en función de X_2.

   c) Estimar el modelo de regresión logística a través de la función glm:

   ```
   glm.res=glm(y~x1+x2,family=binomial)
   ```

   Comentar el resultado obtenido con el comando summary(glm.res) ¿Cual es el aporte de cada variable explicativa?

   d) Realizar las predicciones de Y para la muestra de entrenamiento con

   ```
   yhat=predict(glm.res,data.frame(x1=x1,x2=x2),type='response')
   ```

   Convertir las probabilidades a clases y dar los resultados con una matriz de confusión table(yhat,y).

   e) Simular una nueva muestra de tamaño 100. Calcular la sensibilidad y la especificidad para seq(0,1,0.01). Trazar la curva ROC (como función escalera).

   f) Hacer lo mismo usando una sola variable explicativa en el modelo logístico. Superponer ambas curvas ROC y elegir el mejor modelo.

3. En la página del UCI https://archive.ics.uci.edu/ml/datasets.php bajar los datos de Cancer (Breast Cancer Wisconsin). El objetivo consiste en predecir si el tumor es benigno o maligno a partir de varias variables explicativas. Dividir aleatoriamente el conjunto de datos en train/test.

   a) Estimar el modelo completo. Analizar el aporte de cada variable y dar el valor del AIC.

   ```
   glm.modelo=glm(Class~.,family=binomial, data=dlearn)
   ```

   b) Averiguar si el modelo es significativo al 5 %. Debe calcular

   ```
   chi2=glm.modelo$null.deviance - glm.modelo$deviance
   ddl=glm.modelo$df.null-glm.modelo$df.residual
   pvalor=pchisq(chi2,ddl,lower.tail=F)
   ```

   c) Estimar un modelo donde estén presentes las variables significativas al 5 % del apartado anterior.

d) Estimar un modelo simplificado con el método forward.

e) Estimar un modelo simplificado con el método stepwise.

f) ¿Cual es el mejor modelo con el AIC?

g) ¿Cual es el mejor modelo sobre la muestra test?

h) Trazar la curva ROC para cada uno de los modelos. ¿Cuál es el mejor?