Obligatorio Machine Learning 2022

(Fecha de entrega: 14/6. Todos los ejercicios valen igual por un total de 40 puntos. No se olviden de setseed().).

1. En este ejercicio vamos a simular datos. Usar la función rnorm() de R para crear un predictor X y una muestra de N=100.
    a. Genere una Y=1+X+X^2+X^3+e, donde e es creado usando rnorm() también.
    b. Use el comando regsubsets() para seleccionar el mejor modelo que contenga posiblemente hasta la decima potencia de X (X^10).
    c. Repita b pero usando selección hacia adelante. ¿Como se comparan las respuestas?
    d. Ahora ajusten un Lasso usando hasta la decima potentia de X. Use cross-validation para elegir el valor de lambda el parámetro de ajuste. Grafique.
    e. Ahora genere una respuesta W=1+X^7+e. Haga el criterio óptimo de selección de subconjuntos y el LASSO para hasta la decima potencia de X y discuta los resultados.
2. Use la base de datos del curso Auto data set.
    a. Produzca un diagrama de dispresion que incluya todas las variables en el data set.
    b. Compute la matriz de correlaciones de todas las variables usando la función cor(). Va a tener que excluir la variable name ya que no es cuantitativa.
    c. Use el comando lm() para hacer una regresión lineal multiple para mpg incluyendo todos los regresores excepto name. Use la función summary() para ver los resultados. Comente.
    d. Use el comando plot() para ver el fit del modelo. Comente sobre problemas que vea.
    e. Use el símbolo * y : para ajustar un modelo lineal con interacciones. ¿Alguna de ellas es significativa?
    f. Testee la hipótesis de que Beta1=Beta2=0.
3. Esta pregunta debe ser respondida usando el Weekly data set parte del paquete ISRL o ISRL2.
    a. Produzca medidas de resumen graficas de esta base de datos. ¿Ve algún patrón?
    b. Use la base de datos para ajustar un modelo logístico con Direction como la variable Y y los 5 rezagos mas el volumen como las X. Use la función summary() para ver los resultados. Interprete.
    c. Compute la matriz que muestra los falsos positivos y los falsos negativos.
    d. Ahora ajuste un modelo logístico usando una base de datos de training entre 1990 y 2008, solo usando como regresor Lag2. Repita el punto c para el data set de testeo.
    e. Repita el punto d usando discriminante lineal.
    f. Repita el punto d usando discriminante cuadrático.
    g. Repita el punto d usando KNN.
    h. ¿Qué método ajusta mejor?
4. En este ejercicio van a tener que lidiar con datos del mundo real, no datos limpios del libro del curso. Como grupo, elijan una base de datos de interés que no este en la web del curso.

a. Dedíquense con estos datos a seleccionar el mejor modelo para predecir una variable Y cuantitativa de su elección. Justifiquen la hipótesis de por que ese es el modelo.
b. Justifique por que es un problema de interés.
c. Reporten los resultados del mecanismo de selección del modelo.
d. Reporten los resultados finales del modelo seleccionado. Interprete a la luz del punto a y b.