Aprendizaje Estadístico - FIUBA
Guía 5

Penalización y regularización **1.** Simular datos provenientes de $y = ln(10x + 1) + \epsilon$ con $\epsilon \sim \mathcal{N}(0, 1)$, y realizar ajustes mediante polinomios de grado k. Para cada valor de k simular $n = 100$ muestras y buscar el estimador de mínimos cuadrados para cada muestra y graficarlo sobre el mismo gráfico que contiene a la curva original generadora de datos. Observar como a medida que k aumenta, el sesgo disminuye, pero aumenta la variabilidad en la predicción.

---

2. Simular los datos con tamaño de muestra $n = 100$,
$X_1 \sim U(0; 5)$
$X_2 \sim U(0; 5)$
$X_3 = 2X_1 + U_3$, con $U_1 \sim U(-0.1; 0.1)$
$X_4 = -X_1 + U_4$, con $U_4 \sim U(-0.1; 0.1)$
$Y = 8X_1 - 5X_2 + X_3 + 4X_4 + 5 + \varepsilon$, con $\varepsilon \sim N(0, 1)$
$X_1, X_2, U_3, U_4, \varepsilon$ son variables independientes

Es decir, la respuesta Y depende de dos covariables independientes, X_1, X_2 y las otras dos X_3, X_4 son casi colineales con X_1.

1. Repetir el experimento $B = 1000$ veces. Cada vez, ajustar dos modelos lineales por mínimos cuadrados: uno completo y uno usando regularización Ridge y calcular el desvío estándar de los estimadores, y la proporción de veces que cada coeficiente resultó significativo. Comparar resultados para los dos modelos.

2. Repetir el ítem anterior usando Lasso.

Arboles **3.** Usando el data set *OJ*, del paquete *ISLR*

1. Separar los datos en un set de entrenamiento con 800 observaciones, y set de testeo con el resto.

2. Ajustar un arbol al set de entrenamiento, con *Purchase* como variable respuesta y todo el resto como predictoras. Encontrar el error de entrenamiento (el error al predecir sobre la muestra de entrenamiento)

3. Elegir la cantidad optima de nodos, predecir la respuesta sobre los datos de testeo y calcular el error de predicción

4. Graficar el árbol elegido.

---

4. Aplicar Boosting, Bagging y Random forest a algun set de datos de su elección. Separar los datos en ser de entrenamiento y testeo para evaluar la performance del modelo. ¿Cuan precisos son los resultados comparados con metodos de regresión (lineal o logística según corresponda).

Cluster Analysis **5.** Implemente una función K-medias que dado una matriz X de datos (por fila) y un valor k, efectúa un clustering sobre los datos y devuelva un vector y con el número de clase a la que pertenece.

6.

1. (Simulación ?buena?) (a) Genere 50 datos de la clase A como sigue: $X_i \sim \mathcal{U}(?2, ?1)$, $Y_i \sim \mathcal{U}(1, 2)$.
   (b) Genere50 datos de la clase B como sigue: $X_i \sim \mathcal{N}(0, 1)$,$Y_i \sim \mathcal{N}(0, 1)$.
   (c) Genere 50 datos de la clase C como sigue: $X_i \sim \mathcal{U}(1, 2)$, $Y_i \sim \mathcal{U}(-1, 2)$.

(d) Grafique los tres conjuntos de datos con distintos colores.
(e) Efectuar un clustering K-medias de tres grupos y grafique los resultados del clustering con distintos colores.

2. (Simulación ?mala?) (a) Genere 100 datos de la clase A como sigue: $T_i \sim \mathcal{U}(0, 2\pi)$, $R_i \sim \mathcal{U}(0, 1)$. En base a eso, construir (X_i, Y_i) como sigue: $(X_i, Y_i) = (R_i cos(T_i), R_i sen(T_i)$.
(b) Generar 100 datos de la clase B como sigue: $T_i \sim \mathcal{U}(0, 2\pi)$, $R_i \sim \mathcal{U}(2, 2.5)$. En base a eso, construir (X_i, Y_i) como sigue: $(X_i, Y_i) = (R_i cos(T_i), R_i sen(T_i)$.
(c) Graficar los dos conjuntos de datos con distintos colores.
(d) Efectuare un clustering K-medias de dos grupos. ¿Qué se concluye?

---

7. Considere los datos `productos.txt`. Efectuar un clustering K-medias con dos clusters considerando Precio y Marketing. Considere la conveniencia de hacer un paso previo de estandarización para homogeneizar los datos. Hacer gráficos con los clusters obtenidos y otro con los verdaderos grupos. Realizar el mismo estudio usando Clusters Jerárquicos.

---

8. Considere los datos iris.data. Efectuar un clustering K-medias con tres clusters para los datos ignorando el tipo de especie. Hay alguna semejanza entre el clustering obtenido y las distintos tipos de flores? Realizar el mismo estudio usando Clusters Jerárquicos.