Tarea 2. Introducción a Ciencias de Datos 2018

1. Considere tres poblaciones Poisson, con parámetros $\lambda_1 = 10$, $\lambda_2 = 15$ y $\lambda_3 = 20$ respectivamente.

   (a) Establezca la regla óptima de clasificación, basándose en una sóla observación, x.

   (b) Calcule la probabilidad de error asociado a esta regla óptima.

   (c) Escriba un programa (R o Python) para validar, vía simulación, el nivel de error encontrado en el inciso anterior.

   (d) Suponga ahora que, en vez de hacer una sóla observación sobre el objeto a clasificar, se hacen dos observaciones independientes, x_1 y x_2. Encuentre la regla óptima de clasificación pero ahora basada en $\bar{x} = (x_1 + x_2)/2$.

   (e) ¿Qué tanto mejora el procedimiento, comparado con el basado en una sola observación?.

2. Ingrese a la página del repositorio de datos de Machine Learning de la Univ. de California en Irvine:

   `https://archive.ics.uci.edu/ml/index.php`

   - Bajar el conjunto de datos "Wine"
   - Basados en las características químicas de los vinos, construya un clasificador para determinar el origen de los mismos.

3. Suponga que x es un vector aleatorio d-dimensional, con media μ y varianza Σ y se tienen las siguientes particiones:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

   donde x_1 es $d_1 \times 1$ y x_2 es $d_2 \times 1$, con $d_1 + d_2 = d$ (suponga además que Σ_{11} y Σ_{22} son positivas definidas). Considere $y_1 = a_1^T x_1$ y $y_2 = a_2^T x_2$, donde a_1 y a_2 son vectores no aleatorios de dimensiones $d_1 \times 1$ y $d_2 \times 1$ respectivamente. ¿Qué valores de a_1 y a_2 maximizan el cuadrado de la correlación entre y_1 y y_2?

   - Nota 1: Este problema es llamado "'problema de correlación canónica".
   - Nota 2: Recuerde que la correlación entre dos variables aleatorias univariadas se define como

$$\mathsf{Corr}(u, v) = \frac{\mathsf{Cov}(u, v)}{\sqrt{\mathsf{Var}(u)}\,\sqrt{\mathsf{Var}(v)}}$$

   - Nota 3: Si B es positiva definida y a es un vector, entonces

$$\sup_x \frac{(a^T x)^2}{x^T B x} = a^T B^{-1} a$$

     y el supremo se alcanza cuando x es proporcional a $B^{-1}a$. Este resultado es un caso especial del resultado visto en clase acerca de la maximización del cociente de Rayleigh (cociente de dos formas cuadráticas).

   - Nota 4: Sugerencia

$$\sup_{a_1, a_2} (\cdot) = \sup_{a_1} \left\{ \sup_{a_2} (\cdot) \right\}$$

4. Sean $y_1(x), \ldots, y_K(x)$, K funciones lineales de x, esto es $y_j(x) = w_j^T x + w_{j0}$, donde las w_j's son vectores y las w_{j0}'s son escalares. Sean

$$R_j = \{\, x \mid y_j(x) \geq y_r(x), \text{ para todo } r \neq j \,\}, \quad j = 1, \ldots, K$$

Muestre que R_j, $j = 1, \ldots, K$ son conjuntos convexos. En el contexto de clasificadores lineales, esto quiere decir que las regiones de asignación para cada grupo, son regiones convexas. Los procedimientos de clasificación no siempre inducen regiones convexas (e.g. k-nn, como veremos). ¿Es deseable tener regiones convexas? (explicar).

5. Considere un conjunto de datos $x_1, \ldots, x_n$, los cuáles pertenecen a K grupos o poblaciones $C_1, \ldots, C_K$, con $n = n_1 + \ldots + n_K$, donde $n_i = \#\{C_i\}$. Definamos

$$\bar{x}_i = \frac{1}{n_i} \sum_{x \in C_i} x, \qquad \bar{\bar{x}} = \frac{1}{n} \sum_{j=1}^{n} x_j, \qquad S_i = \sum_{x \in C_i} (x - \bar{x}_i)(x - \bar{x}_i)^T, \qquad S_W = \sum_{i=1}^{K} S_i$$

$$S_B = \sum_{i=1}^{K} n_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})^T, \qquad S_T = \sum_{x \in D} (x - \bar{\bar{x}})(x - \bar{\bar{x}})^T$$

Puede verse que la matriz de dispersión total, S_T, es la suma de la matriz de dispersión dentro de grupos, S_W, y la matriz de dispersión entre grupos, S_B, esto es, $S_T = S_W + S_B$. Considere las siguientes cantidades:

$$L_1 = \text{tr}\left(S_T^{-1} S_W\right), \qquad L_2 = \frac{|S_W|}{|S_T|}.$$

Estas han sido usadas como criterios para determinar que tan separados están los grupos. Sean $\lambda_1, \ldots, \lambda_d$ los valores propios de $S_W^{-1} S_B$. **muestre que**:

$$L_1 = \sum_{i=1}^{d} \frac{1}{1 + \lambda_i}, \qquad L_2 = \prod_{i=1}^{d} \frac{1}{1 + \lambda_i}.$$

Ayuda: Mostrar que $S_T^{-1} S_W = (I + S_W^{-1} S_B)^{-1}$ y esto ayuda a resolver el problema.

Fecha de entrega: Jueves 30 de agosto (entregar sólo los problemas 1, 2 y 3).