PROMiDAT Iberoamericano
Machine Learning con Python II
Métodos NO Supervisados

- Las tareas tienen fecha de entrega una semana después a la clase y deben ser entregadas antes del inicio de la clase siguiente.
- Cada día de atraso en implicará una pérdida de 10 puntos.
- Las tareas son estrictamente de carácter individual, tareas iguales se les asignará cero puntos.
- En nombre del archivo debe tener el siguiente formato: `Tarea1_nombre_apellido.pdf`. Por ejemplo, si el nombre del estudiante es Luis Pérez: `Tarea1_luis_perez.pdf`. Para la tarea número 2 sería: `Tarea2_luis_perez.pdf`, y así sucesivamente.
- Todas las preguntas tienen el mismo valor.
- Esta tarea tiene un valor de un 25 % respecto a la nota total del curso.

Tarea Número 3

1. [25 puntos] En este ejercicio vamos a usar la tabla de datos `SpotifyTop2018_40_V2.csv`, que contiene una lista de 40 de las canciones más reproducidas en Spotify en el año 2018. Los datos incluyen una serie de características importantes del audio de cada canción.

   La tabla contiene 40 filas y 11 columnas, las cuales se explican a continuación.

   - `danceability`: Describe qué tan apta para bailar es la canción
   - `denergy`: Representa una medida de intensidad y actividad.
   - `dloudness`: Sonoridad general de la pista en decibelios.
   - `dspeechiness`: Detecta la presencia de palabras en la canción.
   - `dacousticness`: Indica qué tan acústica es la canción.
   - `dinstrumentalness`: Indica si la canción contiene o no voces.
   - `dliveness`: Detecta la presencia de público en la grabación.
   - `dvalence`: Describe la positividad musical transmitida por la canción.
   - `dtempo`: Es el tempo estimado general de una pista en beats por minuto.
   - `dduration_ms`: Es la duración de la canción en milisegundos.
   - `dtime_signature`: Especifica cuántos beats hay en cada barra o medida.

   Nota: Todas son variables numéricas y no tienen NA. Realice lo siguiente:

   a) Cargue la tabla de datos `SpotifyTop2018_40_V2.csv`

   b) Ejecute un Clustering Jerárquico con la agregación del Salto Máximo, Salto Mínimo, Promedio y Ward. Grafique el dendograma con cortes para dos y tres clústeres.

   c) Usando tres clústeres interprete los resultados del ejercicio anterior para el caso de agregación de Ward usando gráficos de barras y gráficos tipo Radar.

d) Grafique usando colores sobre las dos primeras componentes del plano principal en el Análisis en Componentes Principales los clústeres obtenidos según la clasificación Jerárquica (usando tres clústeres).

2. [25 puntos] En este ejercicio vamos a realizar un `Clustering Jerárquico` para la tabla `SAheart.csv` la cual contiene variables numéricas y categóricas mezcladas. La descripción de los datos es la siguiente: Datos Tomados del libro: The Elements of Statistical Learning Data Mining, Inference, and Prediction de Trevor Hastie, Robert Tibshirani y Jerome Friedman de la Universidad de Stanford. Example: South African Heart Disease: A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of coronary heart disease. Many of the coronary heart disease positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their coronary heart disease event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal. Below is a description of the variables:

   - `sbp`: systolic blood pressure (numérica)
   - `tobacco`: cumulative tobacco (kg) (numérica)
   - `ldl`: low densiity lipoprotein cholesterol (numérica)
   - `Adiposity`: (numérica)
   - `famhist`: family history of heart disease (Present, Absent) (categórica)
   - `typea`: type-A behavior (numérica)
   - `Obesity`: (numérica)
   - `alcohol`: current alcohol consumption (numérica)
   - `age`: age at onset (numérica)
   - `chd`: coronary heart disease (categórica)

   Las dos variables categóricas se explican como sigue: `famhist` significa que hay historia familiar de infarto y que la variable `chd` significa que la persona murió de enfermedad cardíaca coronaria.

   a) Efectúe un Clustering Jerárquico usando solo las variables numéricas y dé una interpretación usando 3 clústeres.

   b) Efectúe un Clustering Jerárquico usando las variables numéricas y las variables categóricas (recuerde recodificar las categóricas usando código disyuntivo completo). Luego dé una interpretación usando 3 clústeres.

   c) Explique las diferencias de los dos ejercicios anteriores ¿Cuál le parece más interesante? ¿Por qué?

3. [25 puntos] Dada la siguiente matriz de disimilitudes entre cuatro individuos $A1$, $A2$, $A3$ y $A4$, construya "a mano" una Jerarquía Binaria usando la agregación del Salto Máximo, Salto Mínimo y del Promedio, dibuje el dendograma en los tres casos:

$$D = \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 2 & 1 & 0 & \\ 3 & 7 & 6 & 0 \end{pmatrix}$$

Verifique los resultados con `scipy.cluster.hierarchy`.

4. [25 puntos] Realice lo siguiente:

   a) Se define la distancia de *Chebychev* como sigue:

   $$d(i,j) = \max|x_{ik} - x_{jk}| \text{ para } k = 1, 2, \ldots, p.$$

   Programe una clase en `Python` que tiene un atributo tipo `DataFrame`, además de los métodos usuales que tiene toda clase, tendrá un método que calcula la matriz de distancias, para esto usará la distancia de *Chebychev* entre dos vectores que se definió arriba.

   b) Calcule la matriz de distancias usando la distancia de *Chebychev* para la tabla de datos `EjemploEstudiantes.csv`.

   c) Para la tabla de datos `EjemploEstudiantes.csv` ejecute un Clustering Jerárquico usando la distancia de *Chebychev* programada por usted y la agregación Ward, compare el resultado respecto a usar distancia *euclidiana* y agregación de Ward. (Debe investigar cómo usar una distancia propia en `scipy.cluster.hierarchy`.

Entregables:

1. Suba en el Aula Virtual en el `Script` generado.
2. Genere desde Jupyter Notebook un documento autoreproducible con la solución de la tarea y súbalo en el Aula Virtual.

PR0MiDAT
IBEROAMERICANO
Programa Iberoamericano de Formación en Minería de Datos