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 4

1. [40 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 el método $k-$medias para $k = 3$. Modificaremos los atributos de la clase `KMeans(...)` como sigue:

   - `max_iter : int, default: 300`: Número máximo de iteraciones del algoritmo k-medias para una sola ejecución. Para este ejercicio utilice `max_iter = 1000`.

- `n_init : int, default: 10` (Formas Fuertes): Número de veces que el algoritmo k-medias se ejecutará con diferentes semillas de centroides. Los resultados finales serán la mejor salida de `n_init` ejecuciones consecutivas en términos de inercia intra-clases. Para este ejercicio utilice `n_init = 100`.

c) Interprete los resultados del ejercicio anterior usando gráficos de barras y gráficos tipo Radar. Compare respecto a los resultados obtenidos en la tarea anterior en la que usó Clustering Jerárquico.

d) Grafique usando colores sobre las dos primeras componentes del plano principal en el Análisis en Componentes Principales los clústeres obtenidos según k-medias (usando $k = 3$).

e) Usando 50 ejecuciones del método $k-$medias grafique el "Codo de Jambu" para este ejemplo. ¿Se estabiliza en algún momento la inercia inter–clases?

2. [40 puntos] En este ejercicio vamos a realizar k-`medias` 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.

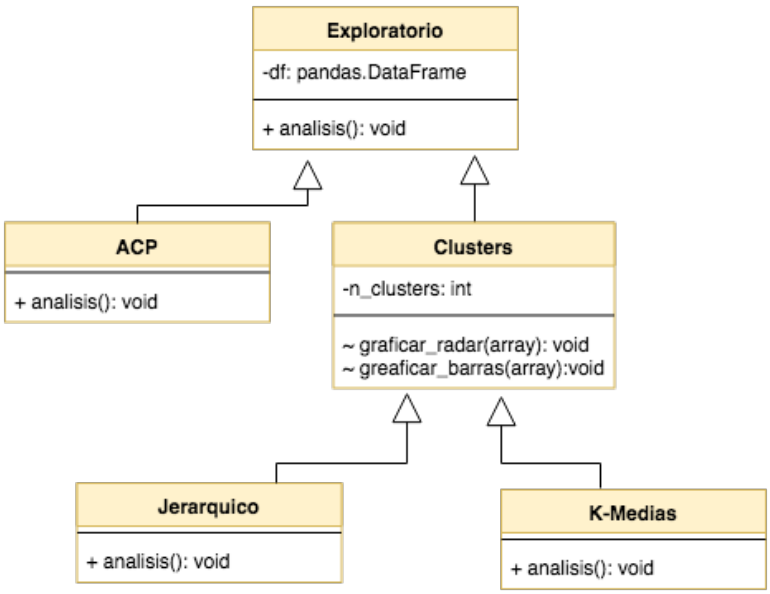
   a) Repita el ejercicio 1 usando $k = 3$ usando esta tabla de datos, usando solo las variables numéricas. Modificaremos los atributos de la clase `KMeans(...)` como sigue:

      - `max_iter : int, default: 300`: Número máximo de iteraciones del algoritmo k-medias para una sola ejecución. Para este ejercicio utilice `max_iter = 2000`.

- `n_init : int, default: 10` (Formas Fuertes): Número de veces que el algoritmo k-medias se ejecutará con diferentes semillas de centroides. Los resultados finales serán la mejor salida de `n_init` ejecuciones consecutivas en términos de inercia intra-clases. Para este ejercicio utilice `n_init = 150`.

b) Repita los ejercicios anteriores pero esta vez incluya las variables categóricas usando códigos disyuntivos completos. ¿Son mejores los resultados?

3. [20 puntos] Programe la jerarquía de clases que se muestra en el siguiente gráfico, la cual fue diseñada especialmente para facilitar los análisis exploratorios de datos vistos hasta ahora en el curso:



La idea es que a través de una instancia de alguna de las clases `Exploratorio`, `ACP`, `Jerárquico` o k-`medias` con solamente ejecutar el método `análisis` automáticamente se despliegan todos los análisis correspondientes a cada caso vistos en el curso (todos hacen por defecto el análisis exploratorio básico). Para esto la clase Base `Exploratorio` tiene un atributo que es una `Data Frame` de `Pandas` y un método `análisis` que realiza al menos los siguiente: Despliega un encabezado de los datos(head), dimensión de la tabla, estadísticas básicas, los percentiles, valores atípicos, boxplot, distribucion de densidad, histogramas y Tests de normalidad. La clase `ACP` agrega al método `analisis` gráficos para el Plano principal, el Círculo de Correlaciones y la inercia acumulada. La clase `Clústeres` agrega un atributo para la cantidad de clústeres y métodos para los gráficos de Radar y de Barras que son usados en Clasificación Jerárquica y k-medias. La clase `Jerárquico` agrega en el método `análisis` los gráficos y análisis vistos en clase. La clase `Jerárquico` agrega en el método k-`medias` agrega los gráficos y análisis vistos en clase para este método.

PRØMiDAT

IBEROAMERICANO

Programa Iberoamericano de
Formación en Minería de Datos