Ciencia de Datos: Visualización de la Información

UNAM-IIMAS, Semestre 2023-2

Tarea-05: Filtrado de datos y distancias

Antecedentes

Sea un conjunto de vectores $X = \{\boldsymbol{x}_i | \, \boldsymbol{x}_i \in \mathbb{R}^n \text{ e } 1 \leq i \leq N\}$, donde $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})^T$. Una distancia $d(\boldsymbol{a},\boldsymbol{b})$ entre dos vectores $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ se puede considerar como la *longitud* de la trayectoria que une a ambos vectores $\boldsymbol{a}$ y $\boldsymbol{b}$. Existen varias definiciones de distancia:

Distancia de Manhattan (norma L_1):

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} |a_i - b_i|$$

Distancia Euclidiana (norma L_2):

$$d(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{i=1}^{n} (a_i - b_i)^2 \right)^{1/2}$$

Distancia de Minkowski:

$$d(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{1/p}$$

Actividades

a).- **Filtrado de datos**: Realice la lectura del archivo "titanic3.csv" y responda las siguientes preguntas:

i. ¿Cuántos cuerpos fueron encontrados?
ii. ¿Cuántos de ellos fueron hombres mayores de 40 años?
iii. ¿Cuántas mujeres desaparecieron entre las edades de 15 a 35 años?
iv. ¿Cuántos hombres mayores de 20 años sobrevivieron?
v. ¿Cuántas mujeres menores de 25 años sobrevivieron?

Además, genere una copia del conjunto de datos y rellene los datos faltantes (NA's) con un valor de 0 en el caso de datos numéricos usados como identificador, la palabra "desconocido" en el caso de datos tipo cadena de caracteres y en el caso de variables numericas use el *promedio* de los valores de esa columna (por ejemplo, la *edad* y la *tarifa*).

Finalmente, de los campos "*age*" y "*fare*" agregue columnas al conjunto de datos que contengan los *valores normalizados*. Elija la normalización tipo

$$\frac{x_i - \bar{x}}{\sigma}$$

para el caso de que la variable tenga una *distribución normal* y utilice la *normalización* tipo

$$\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

en cualquier otro caso.

b).- **Distancias**: Utilizando el archivo "movies.csv" construya una o varias funciones que permitan calcular una *matriz de distancias* para los datos numéricos en el *dataFrame*. La función debe permitir construir la matriz de distancia usando las distancias de Manhattan, Euclideana y de Minkowski (para p igual a 3).

Una *matriz de distancia* es una matriz cuadrada qué contiene las distancias entre los elementos de un conjunto (medidas un par a la vez).

Compare sus resultados con los que se obtienen por medio del método **distance_matrix** de **scipy.spatial**.

Además, usando los métodos "*dendrogram*" y "*linkage*" construya un diagrama en forma de árbol (*dendrograma*) para el conjunto de datos en "movies.csv".
Repita el proceso ahora usando algún esquema de *normalización* del rango de los datos.

i. ¿Qué diferencias puede encontrar en los resultados previos?
ii. ¿En qué casos resulta importante llevar a cabo un proceso de normalización del rango de datos?
iii. Consulte los diferentes tipos de distancias que se pueden usar como parámetro en el método "*linkage*", ¿En qué características de los datos se podría basar uno para elegir una determinada distancia?.

Requisitos
a) Crear su propio código.

Datos
Archivos titanic3.csv y movies.csv.

Es obligatorio entregar:
a) Código fuente.
b) Un reporte que explique:
    1. Conceptos usados.
    2. Estructura de su código.
    3. Instrucciones de ejecución y utilización de su programa. En particular los mecanismos implementados para cambiar los puntos de vista y el tipo de proyección.
    4. Explicar razonadamente las elecciones que han tomado y sus resultados.