Tarea 3 Curso Análisis y Limpieza de datos.

Posgrado en Data Science UDD

Profesor responsable: Jorge Alexis Castillo Sepúlveda

El objetivo de esta tarea es evaluar el conocimiento práctico en lo que respecta a clustering y análisis de componentes principales. Se espera además que los conocimientos sean acumulativos, así que también está considerado evaluar destrezas respecto a la carga, limpieza y análisis descriptivo de datos. También se evaluará el buen uso de markdown para la generación del ipynb, por lo cual se espera que el estudiante presente un trabajo bien redactado y con buena estética ofimática.

Instrucción principal: Generar un archivo .ipynb donde redacte de manera clara y ordenada, todos los desafíos que están enumerados en esta tarea.

En esta tarea usted enfrentará el desafío de segmentar clientes que compran un determinado conjunto de productos en cierto periodo de tiempo. ¿Por qué segmentar clientes? Porque se sabe que los clientes no son todos iguales, y, por tanto, se pueden aplicar ciertas reglas y procesos de negocio diferenciadas de acuerdo a estas categorías de clientes. No es lo mismo, por ejemplo, un supermercado mayorista, que un almacén de vegetales del barrio.

Creando la data

Cada estudiante analizará un set de datos distinto, ya que cada uno/una generará su propia muestra de análisis.

1. Usar la librería numpy y el método random.choice para generar una matriz de datos aleatorios de 500 filas y 6 columnas, en donde las entradas sean números enteros tengan como máximo valor 10000. Esto significará que tiene 500 clientes que compran 6 categorías de productos distintos, en donde cada uno compró entre 1 y 10000 unidades en cierto periodo de tiempo.
2. Las categorías de productos son “vegetales”, “leche”, “abarrotes”, “congelados”, ”limpieza”, “gourmet”. Nombre las columnas en el orden que usted quiera.

Preparando lo datos

3. Normalizar los datos.
4. Remover outliers para cada categoría usando el rango intercuartil.

Componentes principales

5. Aplicar componentes principales a los datos resultantes luego de haber retirado outliers. Ojo: es sin reducir dimensiones aún.
6. Interpretar cada dimensión respecto a las categorías, y las varianzas explicadas
7. Reducir a dos dimensiones y graficar en el plano, en conjunto con todas las componentes. Interpretar estas componentes conjuntas con las direcciones de las dimensiones originales.

Segmentación de clientes

8. Aplicar un método de clusterización a la data reducida, en donde el número de clusters debe ser óptimo de acuerdo al coeficiente de silhouette si usa k-means o alguna extensión de éste.

9. Visualizar con el número de clusters óptimo. Caracterizar cada clúster en el mundo real de acuerdo a los resultados obtenidos.
10. Aplicar componentes principales inverso para recuperar los datos originales.

Se espera que el estudiante vaya redactando e interpretando a medida que va obteniendo sus resultados.

El tipo de entrega es individual. Plazo máximo de entrega: jueves 8 de octubre a las 23:59 hrs., al correo j.castillo@udd.cl