

Tarea 2: Clustering

Plazo: Martes 19 de Julio 11:59 p.m.

En este problema ustedes abordarán el problema de segmentación de clientes usando el algoritmo K-Means. Para ello deberán primero practicar con el demo colgado en curso “demo_clustering.ipynb”; en el se muestra la aplicación del mismo a un dataset pequeño.

En esta tarea usaran la base de datos “Cliente_segmentacion_v0.csv”, subida en Web de curso. Esta base de datos consiste en 851 registros de clientes con los campos: Id, Edad, NivEdua (Nivel educativo), AnhosEmpleo, IngresoMensual, DeudaTarjeta, OtraDeuda, Quiebra (es decir si alguna vez fue declarado en quiebra), Direccion, RatioIngresoDeuda. La mayoría de campos es autodescriptivo.

En este laboratorio debe realizar las siguientes tareas, así como responder las cuestiones:

1. Al procesar los datos, una forma de saber cuales son las variables mas correlacionadas es mediante el coeficiente de correlación. Indique los 2 pares mas correlacionadas y comente si le parecen razonables estos pares.
2. Indique subjetivamente los 5 pares de variables (ejm: AnhosEmpleo e Ingreso Mensual) que usted cree que son las más relevantes para definir patrones de los clientes. Puede incluir el resultado anterior. ¿Cuál es el número de clusters de cada par?. ¿Cómo lo eligió?.
3. Al aplicar K-Means con el número de clusters elegidos: ¿Cuál par tiene los patrones mas definidos?. ¿Cuál par tiene los patrones menos definidos?. Comente sobre los resultados.
4. Al mejor par de variables, ahora agregue una variable de forma subjetiva. ¿Cuál es el número de clusters de cada par?. ¿Cómo lo eligió?. ¿Qué opina de los clusters obtenidos?. (2 puntos)
5. Si un compañero le indicara que Id es una variable útil para ser agregada al algoritmo K-Means. ¿Le parece correcto?. Justifique su respuesta.

- El laboratorio tiene puntaje máximo de 7, además de un puntaje base de 1.
- El máximo de integrantes es de 3 alumnos.
- Las copias están penalizadas según el reglamento de UNAB con 1.0.

Enviar respuestas a correo en formato PDF y respuestas en notebook dentro de zip únicamente a: billy.peralta@unab.cl. De preferencia considere un zip con 1 (con todas las preguntas) o 5 notebooks (1 para las 5 primeras preguntas). Considere el uso de tablas para facilitar la lectura. Espero que tarea les sirva para entender mas de esta técnica. ¡Buena suerte!.

Billy