



TRABAJO PRÁCTICO 04

Clustering (K-Medias y algoritmos jerárquicos)

Introducción

En este trabajo práctico veremos cómo generar e interpretar clusters a partir de un conjunto de datos con variables de diferente naturaleza. Se espera que al finalizar el trabajo el alumno entienda los principales algoritmos de generación de clusters: K-medias y clustering jerárquicos. Se desarrollarán conceptos de distancia asociados a similitud y disimilitud. Posteriormente, se revisarán los algoritmos y su parametrización a través de la codificación en lenguaje Python y la librería Scikit-Learn.

Consignas

1. Medidas de distancia. Calcule la distancia entre los siguientes puntos y el centroide (4, 4) utilizando las medidas: euclídea, Manhattan y Minkowski (con $p = 3$):

A	B
8	2
15	7
2	9

- a. ¿Encuentra diferencias relativas entre las diferentes métricas utilizadas y el resultado obtenido? Explique el comportamiento de cada una utilizando gráficas.
- b. Implemente en Python las funciones de distancias evaluadas y verifique sus cálculos.



- c. Genere 100 puntos al azar con valores definidos en el intervalo [0, 20]. Genere el histograma de distancias promedios a esos 100 centroides generados de manera aleatoria. ¿Qué puede decir?
2. A continuación, calcule la distancia entre las diferentes variables de tipo categóricas con respecto a la instancia {1, soleado, Frío, alta, leve}:

#	PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO
2	lluvioso	Frío	normal	fuerte
3	nublado	Frío	normal	fuerte
4	soleado	templado	alta	leve
5	soleado	Frío	normal	leve
6	lluvioso	templado	normal	leve
7	soleado	templado	normal	fuerte
8	nublado	templado	alta	fuerte
9	nublado	Calor	normal	leve
10	lluvioso	templado	alta	fuerte

Implemente una función en Python que permita determinar:

- a. ¿Cuáles son las instancias más cercanas a la instancia #1?
 - b. ¿Qué función implementó? ¿Por qué?
3. Implemente en Python su versión del algoritmo k-medias utilizando la medida de distancia euclídea. Su algoritmo solo debe recibir k y el número de iteraciones a realizar de manera fija. Corra su agrupamiento y evalúe con silueta.
 4. K-means. Se provee un dataset¹ sobre las características internas del núcleo de tres clases de trigo diferentes. Cargue el dataset en una de las herramientas de minería de datos provistas y resuelva:
 - a. Utilice el algoritmo k-medias variando la cantidad de centroides a efectos de agrupar los datos de la manera más eficiente.
 - b. ¿Cuál es la cantidad de grupos que permite un mejor agrupamiento de los datos? ¿Mediante cuál métrica puede verificar esto?

¹ Disponible en <https://archive.ics.uci.edu/ml/datasets/seeds>



Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

- c. ¿Cuáles son las características más distintivas de cada uno de los clusters resultantes?
5. **Algoritmos jerárquicos.** Incorpore el dataset `abandono_cuantitativo.csv`, a continuación desarrolle las siguientes actividades:
- a. Realice el agrupamiento de los datos utilizando diferentes parámetros.
 - b. Grafique el resultado y escoja cual es el nivel que mejor agrupa los datos

Bibliografía sugerida

Data Mining: Concepts and Techniques. Jiawei Han & Micheline Kamber. Morgan Kaufmann. Third Edition. 2011. Chapter 7.