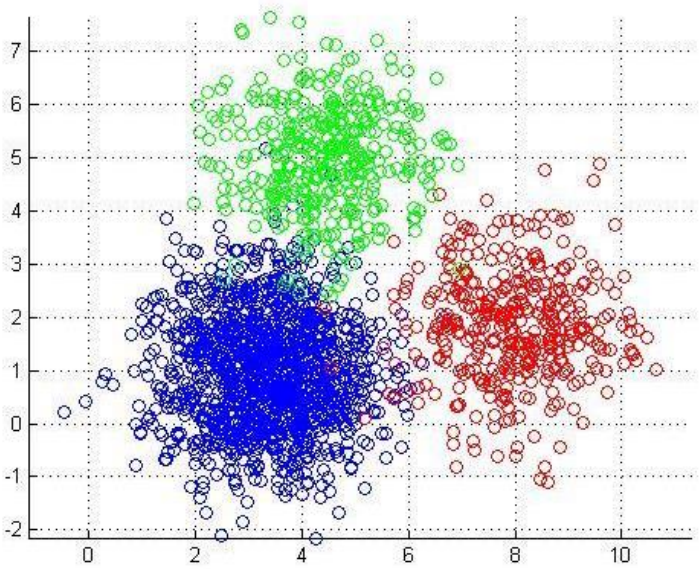
Inteligencia Computacional

Unidad 2: Modelos a partir de datos



El problema del agrupamiento puede definirse como sigue: dados n puntos en un espacio n-dimensional particionar los mismos en grupos tales que los puntos dentro de un grupo son más similares que cada uno a los de los otros grupos, dicha similaridad se mide atendiendo a alguna función distancia (función de disimilaridad) o alguna función de similaridad. Para el mejor funcionamiento de los algoritmos de agrupamiento es importante la detección de ruido para evitar la influencia negativa de éstos en la formación de los grupos y la estimación del número correcto de grupos a determinar.

1. Identifique las etapas del proceso de **Reconocimiento de Patrones**.

2. ¿Qué es **agrupar**? Dé al menos 5 ejemplos.

3. ¿Qué significa "evaluar la distancia entre dos patrones"? ¿Qué medidas de **distancia** existen para variables continuas y nominales?

4. Mire el siguiente video: https://www.youtube.com/watch?v=9991JlKnFmk y responda: ¿Cuáles son las áreas de aplicación del algoritmo Kmedias? ¿Cuál es el propósito de su aplicación? ¿Cuáles son las etapas de este proceso de agrupamiento?

5. Algoritmos de **agrupamientos *kmeans*, *max-min* y *fuzzy-c-means***:

a) Realice una tabla comparativa donde se describan brevemente sus **pasos principales**.
b) ¿Qué **hiperparámetros** requieren del usuario?
c) ¿Cómo se inicializan y qué implicaciones tiene la **inicialización**?

6. **Variación de los resultados con los centros iniciales**

Dado el conjunto de datos *kmeansdata* en MATLAB, o uno similar:

a) Aplicar el algoritmo *kmeans* al conjunto en la variable *p* tomando K=2 y K=3.
b) Repetir el algoritmo unas 25 veces para cada K y mostrar cómo varían los resultados según los centros iniciales resultantes.
c) ¿Cómo se evalúa la calidad del agrupamiento? ¿En qué casos se usa cada método o índice?
d) Utilice las métricas *Silhouette* y *Davies-Bouldin* para encontrar el valor del hiperparámetro K que brinda la mejor calidad de agrupamiento.
e) En base a todos los ensayos realizados ¿Cuál sería el mejor K? Justifique.
f) Los *clusters* o grupos obtenidos, ¿son los esperados al observar la distribución de los puntos?

7. **Variación de los resultados con la normalización**

Dado el conjunto de datos de MATLAB *fisheriris*:

a) Descríbalo indicando formato, dimensiones y variables.
b) Visualice los datos. ¿Qué cantidad de grupos observa?
c) Utilice los datos sin normalizar, con normalización y con estandarización.
d) Compare los resultados obtenidos con los diferentes métodos y tipos de normalización.
e) Indicar qué **medidas intra-clase e inter-clases** y sus **parámetros**.
f) Evalué los agrupamientos realizados con las medidas mencionadas.
g) Justifique los resultados obtenidos.

8. En la base de datos del repositorio UCI:
https://archive.ics.uci.edu/ml/datasets.htmldatasets/seeds
se almacenan diversos descriptores de la geometría de semillas de tres clases diferentes de trigo (Kama, Rosa y Canadiense).

a) Indique la dimensión y tipo de datos.
b) Describa la información recolectada y el objetivo de su estudio.
c) Ensaye las técnicas de agrupamiento. Para ello:
- Analice los grupos formados utilizando las métricas que considere adecuadas.
- ¿Qué información permite obtener cada una?
d) Varíe los hiperparámetros de los algoritmos para elegir la implementación que presenta el menor error. ¿Qué medidas utiliza? Justifique.

9. ¿En qué se basa el método DBSCAN? ¿Qué ventajas presenta?
Material: http://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf