Clustering

Comparación de métodos

Christian Oliva Moya

Dpto. de Ingeniería Informática, Escuela Politécnica Superior

Universidad Autónoma de Madrid

28049 Madrid, Spain

v1 - 28/02/2024

Consideraciones finales (1)

- Cuidado con la alta dimensionalidad. Cada nuevo atributo hace que los elementos estén más alejados. Las medidas de distancia acaban siendo inútiles

- Alternativas:
  - Preprocesamiento de los datos
  - Aplicar algún algoritmo de reducción de dimensionalidad (PCA)
  - Realizar una fase de selección de características o atributos significativos

Consideraciones finales (2)

- No hay un método cerrado. Encontrar la solución óptima con un algoritmo de clustering es realmente difícil.

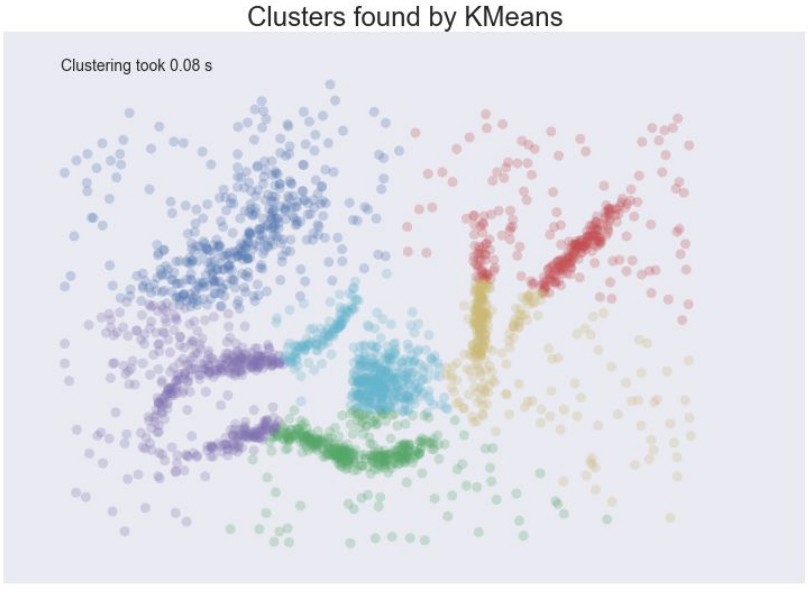
- Seamos creativos. Hemos visto que podemos implementar cualquier heurística o estrategia para resolver un problema en particular.

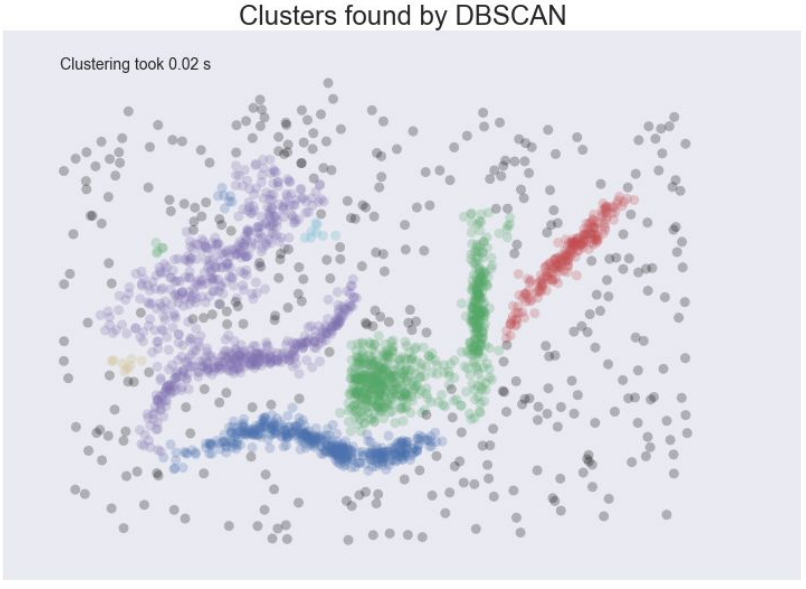
Consideraciones finales (3)

- Compara los resultados de diferentes algoritmos. No todos los algoritmos de clustering funcionan bien para todos los problemas.

- ¿Cómo comparamos los algoritmos de clustering?

Comparación de Clustering (1)





Comparación de Clustering (2)

- Tenemos que basarnos en las siguientes propiedades:
  - Cohesión intra-cluster: un buen algoritmo de clustering tiene puntos agrupados (zonas con alta densidad)
  - Separación inter-cluster: un buen algoritmo de clustering separa mucho los clusters
- Ejemplo con K-Means:
  - ¿Cómo de cerca están los puntos de un cluster a su centroide?
  - ¿Cómo de lejos están los centroides unos de otros?

Comparación de Clustering (3)

- Si para un punto x definimos:
  - intra(x) como el promedio de la distancia de x a los miembros de su propio cluster
  - inter(x) como el promedio de la distancia de x a los miembros del cluster diferente más cercano
- Podemos medir la similitud intra- y inter-cluster para ese punto x como:

$$s(x) = \frac{inter(x) - intra(x)}{max\{inter(x), intra(x)\}}$$

Comparación de Clustering (4)

$$s(x) = \frac{inter(x) - intra(x)}{max\{inter(x), intra(x)\}}$$

- Índice de Silhouette: Media de s(x) para todos los puntos

$$sil = \frac{\sum_i^n s(x_i)}{n}$$

- El promedio proporciona una medida de coherencia general de los clusters.
  - Valores cercanos a +1 indican que los clusters están bien definidos.
  - Valores cercanos a 0 indican superposición de clusters.
  - Valores cercanos a -1 indican un clustering malo.

Comparación de Clustering (5)

- Notebook *06_dbscan.ipynb*