Material Imprimible

Curso de Data Science

Módulo 5 - Clustering

Contenidos:

- Aprendizaje No supervisado: Clustering
- Algoritmos de Clustering
- KMeans
- Evaluación de resultados

Aprendizaje No supervisado: Clustering

Un problema no supervisado de clustering es una técnica de Machine Learning que se utiliza para encontrar patrones y conexiones ocultas en un conjunto de datos sin etiquetas predefinidas. A diferencia de los problemas supervisados, en los que se dispone de un objetivo etiquetado, en el clustering no se dispone de información sobre las categorías o etiquetas a las que pertenecen los datos.

El objetivo del clustering es agrupar un conjunto de datos en subconjuntos o clusters, donde cada cluster contiene datos que son similares entre sí y diferentes de los datos en otros clusters. Los datos se agrupan en función de sus similitudes y diferencias en algún espacio de características (también llamado espacio de atributos). La similitud entre dos puntos puede medirse utilizando una variedad de métricas, como la medida de la silueta, el método del codo, entre otros.

El clustering puede ser útil en diversas aplicaciones, como la segmentación de clientes, la agrupación de imágenes por temas, la identificación de patrones en el análisis de datos biológicos, la clasificación de documentos, entre otras. Existen diversos algoritmos de clustering, como k-means, clustering jerárquico, DBSCAN, mean-shift, entre otros, que utilizan diferentes técnicas para agrupar los datos.

Es una herramienta útil para descubrir patrones y estructuras ocultas en los datos en una amplia variedad de aplicaciones.

Algoritmos de Clustering

Los algoritmos más famosos para problemas de Clustering son:

- K-Means: Es uno de los algoritmos de clustering más simples y populares. Este algoritmo divide los datos en un número fijo de clusters, representados por sus centros, y luego asigna cada punto de datos al cluster más cercano.

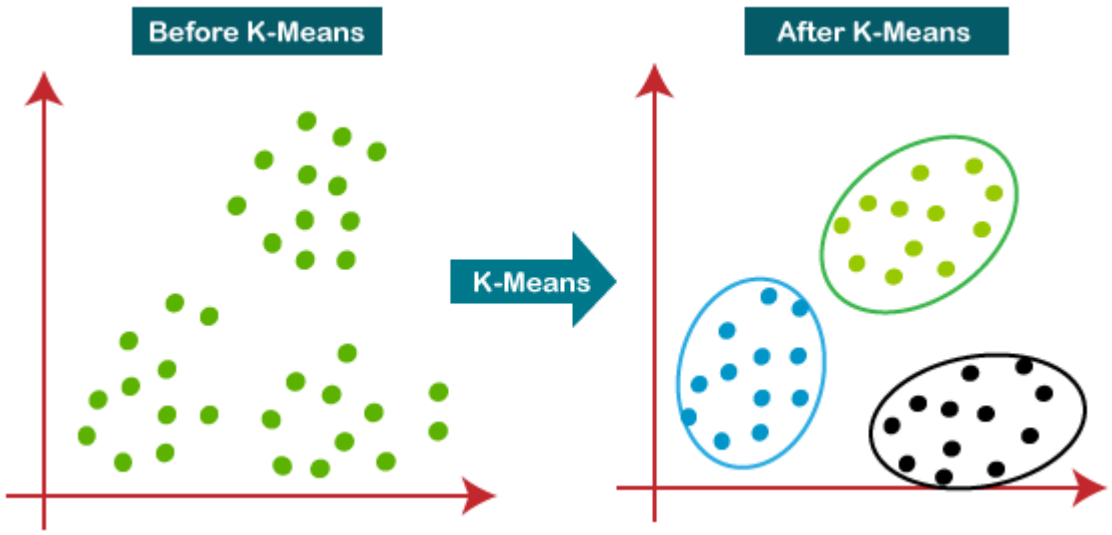
- Clustering Jerárquico: Este algoritmo crea una jerarquía de clusters de manera recursiva, dividiendo los datos en subgrupos más pequeños. Puede ser aglomerativo (comenzando con clusters individuales y fusionándolos) o divisivo (comenzando con todos los puntos en un cluster y dividiéndolos en subgrupos más pequeños).

- DBSCAN: Este algoritmo es útil para encontrar clusters de formas arbitrarias y puede manejar ruido y valores atípicos. Agrupa los puntos de datos que se encuentran cerca unos de otros en clusters y separa los puntos de datos aislados en clusters separados.

- Affinity Propagation: Este algoritmo es capaz de encontrar el número de clusters automáticamente y es útil para encontrar clusters de formas complejas y tamaños diferentes. Es un método iterativo que utiliza una matriz de afinidad para encontrar los puntos de datos más representativos, llamados "ejemplares", y luego agrupa los demás puntos de datos alrededor de ellos.

KMeans

Link: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html



K-means es uno de los algoritmos de clustering más simples y populares que se utiliza para agrupar datos en subconjuntos o clusters en función de sus similitudes y diferencias en algún espacio de características.

El objetivo de este algoritmo es minimizar la suma de las distancias entre cada punto de datos y el centroide de su cluster correspondiente. La convergencia se alcanza cuando la suma de las distancias no disminuye significativamente con cada iteración.

El resultado del algoritmo K-means es una partición de los datos en K clusters, donde cada punto de datos pertenece a un solo cluster. Cada clúster está representado por su centroide, que es la media de los puntos de datos asignados a ese clúster.

Aunque es un algoritmo simple y fácil de implementar, puede haber limitaciones en su uso. Por ejemplo, es sensible a la elección inicial de los centroides, y puede haber problemas cuando los conjuntos de datos (filas y columnas) son de un tamaño grande. También puede ser difícil seleccionar el número óptimo de clusters en algunos casos.

Evaluación de resultados

En problemas predictivos de clustering no supervisado no tenemos un set de datos para evaluar contra la realidad, por ende, el proceso de medición e interpretación puede resultar más dificultoso que en problemas de regresión o clasificación. No obstante, hay una serie de métricas que nos pueden servir como punto de partida para ver que tan bien están diferenciados los clusters en el set de datos X. La medida más utilizada es la de la Silueta.

La medida de la silueta es una métrica de evaluación de clustering que se utiliza para determinar la calidad de la agrupación o clustering. Indica qué tan bien un punto de datos se ajusta a su propio cluster en comparación con otros clusters.

La medida de la silueta se calcula para cada punto de datos y varía de -1 a 1. Los puntos de datos con una puntuación alta de silueta están bien ajustados a su cluster y están alejados de otros clusters. Los puntos de datos con una puntuación baja de silueta están mal ajustados a su cluster y podrían estar más cerca de otros clusters.

Los valores que están cercanos a 1, indican que los datos se diferencian bien de un cluster de otro en su mayoría, y por el contrario, cuanto más cercano a -1 estaremos hablando de clusters totalmente difusos. Normalmente se esperan valores que oscilen de 0.25 a 0.75 para clusters buenos y aceptables, y muy buenos aquellos que superan los 0.75, aunque no hay una regla definida y depende de la tolerancia de cada científico de datos y problema de negocio.