Taller 9: Agrupamiento K-Means

En esta práctica se verá los elementos necesarios para generar un proceso de agrupamiento usando K-means. El objetivo es que este proceso puede ser aplicado a cualquier caso de prueba que requiera ejecutar este tipo de agrupamiento. En primera instancia se aplicará el proceso para un caso de uso, pero más adelante se ejecutará la misma tarea para otro escenario, indicando los pasos necesarios para reutilizar el proceso creado en una nueva base de datos

En general un proceso de agrupamiento basado en k-means requiere cuatro operadores

Retrieve or Read CSV: Este operador permitirá leer datos desde una fuente. En este caso se usará la BD IRIS https://archive.ics.uci.edu/ml/datasets/iris. Esta fuente de información es quizás la base de datos más conocida que se encuentra en la literatura de reconocimiento de patrones. El artículo de Fisher es un clásico en el campo y se lo menciona con frecuencia hasta el día de hoy. (Ver Duda & Hart, por ejemplo.)

La base de datos contiene la siguiente información:

1. longitud del sépalo en cm
2. ancho del sépalo en cm
3. longitud del pétalo en cm
4. ancho del pétalo en cm
5. clase:
- Iris Setosa
- Iris Versicolor
- Iris Virginica

La idea en esta práctica es tratar de buscar agrupamiento de las flores considerando las variables de la 1 a la 4, sin considerar por supuesto el atributo 5 que representan la clase de flor.

Normalize: Este operador permite normalizar los datos. El objetivo de normalizar es poner a competir a todas las variables numéricas en los mismos términos. Dado que el agrupamiento en k-means se basa en distancias, por tanto lo ideal es que las variables sean aproximadamente las mismas. De modo que una variable que esta medida en una unidad muy grande no domine a otra que esta medida en una unidad más pequeña

El método usado para normalizar será Z-transformation el cual para cada una de las columnas le va a aplicar la media y lo va a dividir para la desviación estándar. Luego de aplicar este proceso todas las medias serán 0 y las desviaciones estándar serán 1. Para observar estos valores use el botón Statistics una vez se ejecute el proceso. De esta manera podemos afirmar que todas las variables están en igualdad de condiciones

K-means: Este operador permitirá ejecutar el proceso de agrupamiento. Es necesario indicar el valor de k, que indica el número de grupos que deseamos hallar. Es difícil conocer a priori cuántos grupos se pueden hallar en los datos, por tanto el objetivo es probar varias opciones hasta encontrar valores de similitud muy cercanos entre sí. Luego de ejecutar este proceso se puede observar que al primer cluster pertenece 50 items y al segundo 100 items

En esta opción se puede observar además la tabla de centroides, lo cual indica las medias de las distintas variables una vez que las hemos separado en dos grupos. Al estar normalizadas las variables lo único que se puede indicar con estos valores es que por ejemplo la media del atributo sepal_length es -1.011 y del segundo cluster es un valor positivo 0.506. No es posible interpretar directamente estos datos

El grafico usando la opción Plot permite diferenciar los valores entre los dos clúster. Por ejemplo aquí se puede visualizar que las flores rojas tienen los pétalos más anchos (petal_width) y altos (petal_height) que las flores rojas.

Cluster Distance Performance: El cual va a permitir evaluar la calidad del agrupamiento, midiendo la distancia promedio de cada una de las flores en el grupo a su centroide de ese cluster. El centroide se convierte como en el representante de cada uno de los grupos.

Del operador de Clustering salen dos terminales:

El Cluster model debe conectarse al Cluster Model en el operador de Performance y el terminal clustered set debe conectarse al terminal example set.

Tareas

1. Pruebe de cambiar el número de grupos de agrupamiento y verifique los resultados obtenidos. Considere que el conjunto de datos contiene 3 clases de 50 instancias cada una, donde cada clase se refiere a un tipo de planta de iris. Una clase es linealmente separable de las otras 2; los últimos NO son linealmente separables el uno del otro
    a. Identifique los valores de los centroides para diferentes grupos y explique las diferencias
    b. Use los gráficos disponibles usando diferentes parámetros para decidir el mejor número de cluster?
2. Pruebe de aplicar este mismo proceso pero para otra fuente de datos. Reporte los cambios que fue necesario ejecutar