Tecnológico de Monterrey

Herramientas computacionales: el arte de la analítica

TC1002S.222

Actividad Evaluable 4: Patrones con K-means
Equipo 5

Nombre – Apellidos	Matrícula	Iniciales de carrera
Sebastián Íñigo López	A01661179	ITC
Karina Ruiz Tron	A01656073	IRS
María Fernanda Argueta Wolke	A00830194	ITC
José Andrés Rodríguez Ruiz	A01661651	ITC
José Aram Méndez Gómez	A01657142	ITC

Profesor: Sergio Ruiz Loza

Febrero – Junio

Fecha de entrega: 12/05/2022

Análisis

Basado en los centros responde las siguientes preguntas

- ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?

Si lo son, el hecho de que los datos estén organizados en centros no significa que dejen de ser representativos o carezcan de valor. La información obtenida es representada en una línea y en dicha línea se dan las agrupaciones que dan origen a los llamados centros para su consecuente análisis.

- ¿Cómo obtuviste el valor de k a usar?

Si de antemano conocemos cómo se agrupan los datos y claramente se puede observar el número de centros, resulta bastante fácil determinar el valor de k. Pero por otro lado, si no conocemos el número de agrupaciones, en realidad es a base de prueba y error. Se tiene que variar el valor de k y comparar entre dichos valores su variación total. Si graficamos la reducción en la varianza por valor para k nos arroja una gráfica de codo ("Elbow Plot"). Y al tener está gráfica, se puede encontrar k al encontrar el codo en la gráfica debido a que en ese punto existe una gran reducción en la variación con respecto a la variación de k, pero después de eso, la variación no baja tan rápido.

- ¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?

No. Cada vez que agregamos un nuevo grupo(k+n), la variación total dentro de cada grupo es menor que antes. Y cuando solo hay un punto por conglomerado, la variación es igual a 0. De la misma manera, cuando utilizamos un valor más bajo, la variación total tiende a ser bastante alta. Por lo que los centros son más representativos cuando el valor de k es el correcto, es decir el número de centros que existen, ni más, ni menos.

- ¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?

Al analizar la base de datos "avocado.csv" con K-means. Podemos observar un total de 5 centros con un comportamiento bastante particular, donde encontramos 3 centros agrupados en la esquina inferior izquierda contenidos en un rango de $0.5 * 1e^7$ y encontramos similarmente 2 centros agrupados en la esquina superior derecha igualmente contenidos en un rango de $0.5 * 1e^7$. Entre ambos conglomerados de centros, un conglomerado de 3 centros y otro de 2 centros respectivamente, hay una distancia de $0.5 * 1e^7$.

- ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?

Si tuviéramos muchos outliers en el análisis de cajas y bigotes, los centros variarán dependiendo de qué tantos y qué tan lejos se encuentren dichos outliers. Puede que debido a la variación entre los outliers y los datos algunos centros se eliminen o incluso se puede dar el caso de que si se tiene un número importante de outliers, sean considerados como centros. En general, entre más outliers se tenga y más varianza exista entre ellos y los datos, peor será el análisis de cajas y bigotes y los centros terminarán siendo muy inexactos.

- ¿Qué puedes decir de los datos basándose en los centros?

Gracias al algoritmo K-Means, se pudo organizar todos los datos que teníamos sin etiquetar. Se encontraron 5 grupos/centros (clusters) entre los datos crudos y observándolos se puede concluir que, los primeros 3 centros se encuentran interrelacionados debido a la cercanía en la que están y los últimos 2 también.