Actividad 4

Implementación de K-means

Andrea Valeria Pintor Valencia A01067424

Victoria Ramírez Castro A01640824

Carlos Bernabé Rojas Medina A01641668

Juan Diego Salcedo García A01368540

Cesar Simental-Dueñas A01641385

octubre 2023

Herramientas computacionales: el arte de la analítica

Gpo 201

https://github.com/apintorv/Semana_Tec.git
https://colab.research.google.com/drive/1frVQiJf3H1boh035ZJL8aPPMxLVxJheo?usp=sharing

¿Qué para qué variables fue más conveniente usar: modelo lineal o polinomial?

Dado que para ambas aproximaciones usamos las mismas variables (pacientes y positivos) y ambos modelos nos dieron un score muy aproximado a 1 (que es lo óptimo), pudiéramos usar cualquiera entre el modelo lineal y polinomial.

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

La elección de las variables o características utilizadas para la agrupación es fundamental. Si las variables seleccionadas son relevantes y capturan aspectos importantes de la propagación del COVID-19 y el número de pacientes por hospital que ingresaban ya sea por tratamiento o para la realización de pruebas.

¿Cómo obtuviste el valor de k a usar?

Para determinar el valor óptimo de "k", se analizaron los resultados de WCSS y el Silhouette Score. El valor óptimo de "k" generalmente se encuentra en el punto donde el valor de WCSS comienza a aplanarse (el "codo" en el gráfico del codo) y donde el Silhouette Score es más alto.

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

Serían menos representativos usando un valor más bajo o alto para los centros, se observan de manera relativamente bien los centros, por lo que al reducir al valor se podría ver con detalle los centros pero se perdería continuidad. De manera ideal el cálculo anterior de k nos da una buena aproximación al acomodo de los centros por lo que nos podemos quedar con ese valor.

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

En general la distancias son relativamente constantes con pocas variaciones en las distancias , esto teniendo en cuenta la separación por los datos anormales en las gráficas que generan los grupos, se logra apreciar que a pesar de hacer “overlap” no se encuentran en la misma posición.

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

Si existieran muchos datos anómalos en el análisis de cajas y bigotes, los centros de los clusters se verían afectados en su posición. Los valores anómalos podrían influir en la ubicación de los centros distorsionando la verdadera representación de los grupos identificados por el algoritmo de clustering.

¿Qué puedes decir de los datos basándose en los centros de los clusters y en los modelos?

Los datos muestran una fuerte relación lineal entre el número de pacientes hospitalizados y el número de positivos a COVID. Al agrupar los datos en clusters, se identifican principalmente dos grupos: uno con valores bajos en ambas variables, y otro con valores altos. Esto sugiere que a mayor número de pacientes hospitalizados, mayor número de casos positivos a COVID detectados. Los clusters representan los diferentes niveles de estas variables presentes en los datos.