A picture containing icon

Description automatically generated

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

ASIGNATURA: SISTEMAS INTELIGENTES

SECCIÓN: 1296

Tarea #4 – Clustering & KNN

PRESENTADO POR:

DANIEL AUGUSTO MORALES ALVARADO CTA: 11941247

JUDÁ AARON PONCE VILLALTA CTA: 11841248

CATEDRÁTICO: Dr. KENNY M. DÁVILA

CAMPUS TEGUCIGALPA 27 DE NOVIEMBRE DE 2022

Índice

[Introducción 1](#_Toc120458584)

[Implementación 2](#_Toc120458585)

[Resultados Clustering 3](#_Toc120458586)

[Resultados K-NN 3](#_Toc120458587)

[Dificultades Encontradas 4](#_Toc120458588)

[Planteamiento 4](#_Toc120458589)

[Conclusiones 6](#_Toc120458590)

Introducción

El siguiente informe mostrará la aplicación de clustering y K-NN siguiente paso a paso lo solicitado y con sus atributos apropiados para evitar problemas graves. Sabemos que existen varios algoritmos para clustering, en esta tarea se implementarán 3, k-means, agglomerative y DBscan respectivamente. El clustering nos ayuda a lograr agrupaciones de conjuntos de datos no etiquetados para hacer subconjuntos, de ahí el termino cluster. Para realizar la segmentación se entregó tres datasets, cada uno con size diferente que serán utilizados para los tres algoritmos de clustering. Para la implementación de k-means se utilizará un valor k, correspondiente al numero de cluster para cada dataset, el valor de k puede tomar 5 valores, del 1 al 5. Para agglomerative se utilizará los mismos valores de k y el valor de umbral, cabe mencionar que, si se usa un valor de k, el umbral es None y viceversa. Para el ultimo algoritmo de clustering (DBScan), se consideran 2 atributos, min simples que puede tomar 3 valores (5,10,15) y 5 valores de eps, este valor es de tipo double y los valores a utilizar serán 0<x<1.

Para la implementación de K-NN se utilizarán los datos que corresponden a la tarea anterior, donde se predijo el gusto particular sobre restaurantes. K-NN sabemos que se utilizar para el aprendizaje supervisado y es fácil de implementar. Los datos al no ser cuantitativos se reemplazarán por 0s y 1s. Para este problema se utilizarán 8 valores k y 4 datos de entrenamiento que al final habran 32 configuraciones diferentes.

Se analizarán los resultados obtenidos para clustering como para K-NN, así como también se mostrarán las gráficas correspondientes para cada configuración de cada problema y con el algoritmo correspondiente.

Implementación

Para esta tarea se realizaron 5 scripts que fueron previamente solicitados uno por cada requisito.

Se utilizaron 6 cantidad de librerías que a continuación se mencionan:

* Scikit-learn
* Pandas
* Pickle
* Matplotlib
* Sys
* Numpy
* Time

Para poder ejecutar el script 1 correspondiente a k-means se coloca el siguiente comando: py .\kmeans.py .\datos\_1.csv 2, el numero 2 corresponde al número de clusters.

Para poder ejecutar el script 2 correspondiente a agglomerative se debe colocar el siguiente comando: py .\jerarquico.py .\datos\_1.csv 2 c ó py .\jerarquico.py .\datos\_1.csv 2 u. La letra “c” corresponde a cluster y el numero que tiene antes corresponde a la cantidad de clusters, lo mismo sucede cuando se coloca la letra “u”, que corresponde a la cantidad del umbral. Cuando se coloca “c” el umbral toma “None” por defecto y viceversa.

Por otro lado para el script 3, correspondiente al DBscan, se ejecuta con el siguiente comando: py .\dbscan.py .\datos\_1.csv 0.5 5, el valor de 0.5 corresponde al valor de eps y 5 al valor del sample.

Para el script 4 se ejecuta el siguiente comando: py .\knearestneighbors.py .\training\_data\_large.csv 3, el numero 3 corresponde al valor de k.

Para el ultimo script se utiliza el siguiente comando: py .\knearestvalidator.py .\ModelosKNN\KNN\_k\_15\_medium.pkl .\validation\_data.csv

Resultados Clustering

1. Mejores Ilustraciones
2. Análisis

1. ¿Qué tipo de clustering considera que funciona mejor en cada dataset y por qué?

2. ¿Cuántas clases reales cree que se usaron para generar los datos en cada dataset?

Resultados K-NN

1. Análisis del efecto del valor de K en cuanto a rendimiento y overfitting.
2. Análisis del efecto del tamaño del dataset en cuanto a rendimiento y overfitting.
3. Comparación contra los resultados de Arboles de Decision y Random Forest.
4. Sorpresas. Detalle si los resultados anteriores fueron consistentes o no con sus propias expectativas sobre el tema.

Dificultades Encontradas

Con la experiencia adquirida con la tarea pasada, esta vez fue más sencillo implementar estos algoritmos, sin embargo tuvimos un obstáculo muy grande que estaba fuera de nuestras manos, y básicamente el problema pasaba al momento de ejecutar el agglomerative por la cantidad de datos y obviamente por el proceso que hace este algoritmo para segmentar los datos, lo que ocasionaba es que la computadora de uno de nosotros no tenia memoria RAM suficiente y este algoritmo la utiliza mucho, la solución inmediata era correrlo en una PC con al menos 16 GB de RAM, quitando eso no encontramos mas dificultades.

Planteamiento

Nuestro nuevo planteamiento tiene un poco de relación con la definición como tal de K-NN, vecinos mas cercanos, es decir que se tomó literalmente la palabra vecinos y como uno de los integrantes de este equipo esta utilizando Google Earth Engine y se sabe que esa herramienta también ayuda a predecir enfermades entonces investigamos si con K-NN se puede lograr eso y, si se pude por lo tanto, lo utilizaríamos para predecir en cuanto tiempo el covid (solo por poner una enfermedad, pero puede ser cualquiera) avanzará a su vecino más cercano. Obviamente tomando en cuenta ciertos atributos que tenga relación con la población de esa vecindad.

Los atributos que se considerarán por ejemplo un índice de contacto de las personas con otras personas o si tuvo alguna interacción, lugares con mucha gente donde no respeten el distanciamiento, cuantas veces se desinfecta las manos o los objetos que toca y si la vecindad es un lugar muy cerrado o con pocos espacios donde fácilmente las personas pueden contagiarse. Básicamente se agruparían por casas, y en una casa por lo general viven entre 2 a 6 personas, pero la cantidad puede variar y pues esto también es una característica que tiene mucho que ver con la propagación de cualquier enfermedad.

Al agrupar los datos por casas, puede que no se obtengan los resultados correctos, pero para este ejemplo se van a considerar de esa manera, pero fácilmente los agrupamientos pueden extenderse a colonias vecinas y agrupar por colonias y no por casas individuales que perteneces a una colonia como tal. También se debe de tomar en cuenta es como una persona infectada toma las recomendaciones necesarias y obligatorias para intentar no contagiar a los de su casa y estos no contagiar a sus vecinos.

Al obtener los resultados con este agrupamiento y los atributos mencionados, mediante algún índice de solubilidad podría predecirse que vecinos están mas propensos a ser contagiados.

Conclusiones

* Se implementaron todos los algoritmos solicitados para esta tarea, cumpliendo cada uno de los requisitos para la obtención correcta de los resultados.
* Al igual que la tarea 3, los modelos entrenados para K-NN fueron almacenados en un objeto de tipo pickle en el cual fue almacenada toda la información necesaria para el cálculo y la extracción de las métricas para la construcción de las tablas.
* Se logró comprender el funcionamiento de clustering y K-N, así como también la diferencia entre los algoritmos de clustering los cuales son necesarios para implementarlos en casos donde sea más beneficioso.
* Finalmente, mediante la implementación de estos algoritmos logramos experimentar cuan importantes son, sin embargo, siempre pueden haber situaciones donde el overfitting se haga presente.