Propuesta recibida en mayo 2024 y aceptada para publicación en mayo 2024.

Maching Learning KNN Y K-MEANS

Maching Learning KNN Y K-MEANS

Robinson Restrepo Restrepo
Institución Universitaria ITM

Sebastian Tangarife Salazar
Institución Universitaria ITM

Resumen

El lenguaje de aprendizaje automatico (Machie learning) es un campo de la inteligencia artificial que está enfocado en desarrollar algoritmos que permite a las maquinas aprender a partir de datos. Estos algoritmos son muy esenciales para realizar predicciones o tomar algunas decisiones basada en los patrones que estemos observando en los datos. Dentro de este campo hay muchos hay muchos tipos de algoritmos algunos de ellos son los de clasificación y agrupamiento los cuales son muy esenciales para extraer información significativa de grandes conjuntos de datos. En este documento vamos a explorar dos algoritmos esenciales KNN (K-Nearest Neighnors) y K-Means.

KNN en un algoritmo de clasificación supervisada no parametrizada que utiliza la proximidad para hacer clasificaciones o predicciones. Funciona comparando un punto de datos nuevo con sus vecinos mas cercanos en el conjunto de entrenamiento, la cual utiliza una métrica de distancia y numero de vecinos la cuales se determinan por una K las cuales son demasiado cruciales para el rendimiento del algoritmo. Este es muy fácil de implementar y se adapta muy fácil a los datos que tienen mucho ruido. Sin embargo tiene desventajas muy significativas, como la dificultad para escalar conjuntos de datos demasiado de grandes debido al costo computacional que se necesita para encontrar los vecinos mas cercanos y la tendencia a sobre ajustarse cuando el número de características crece.

K-menas es un algoritmo de clustering no supervisado ampliamente utilizado debido a su simplicidad y eficacia. Este algoritmo agrupa datos no etiquetados en un numero predefinido de los clústeres (K), cada es representado por un centroide. El proceso de K-means incluye la selección inicial de los centroides, la asignación de cada punto de datos al centroide mas cercano y la recirculación de los centroides basados en los puntos asignados. Este proceso lo puede repetir varias veces hasta que las asignaciones ya no cambien significativamente. Este algoritmo es computacionalmente eficiente y el cual puede manejar grandes cantidades de datos de alta dimensionalidad, aunque este se puede ver muy afectado con esto al momento de este seleccionar los centroides de inicialización y su suposición de los clústeres son esféricos y de tamaño similar.

Palabras clave

KNN, k-means, algoritmo, centroide, clustering

Abstract

Machine learning language (Machie learning) is a field of artificial intelligence that is focused on developing algorithms that allow machines to learn from data. These algorithms are very essential to make predictions or take some decisions based on the patterns we are observing in the data. Within this field there are many there are many types of algorithms some of them are classification and clustering algorithms which are very essential to extract meaningful information from large data sets. In this paper we will explore two essential algorithms KNN (K-Nearest Neighnors) and K-Means.

KNN in a non-parameterised supervised classification algorithm that uses proximity to make classifications or predictions. It works by comparing a new data point with its nearest neighbours in the training set, which uses a distance metric and number of neighbours which are determined by a K which are too crucial for the performance of the algorithm. It is very easy to implement and adapts very easily to noisy data. However, it has significant disadvantages, such as the difficulty in scaling very large datasets due to the computational cost needed to find the nearest neighbours and the tendency to overfit when the number of features grows.

K-mines is a widely used unsupervised clustering algorithm due to its simplicity and effectiveness. This algorithm groups unlabelled data into a predefined number of clusters (K), each represented by a centroid. The K-means process includes the initial selection of centroids, the assignment of each data point to the nearest centroid and the recirculation of centroids based on the assigned points. This process can be repeated several times until the assignments no longer change significantly. This algorithm is computationally efficient and can handle large amounts of high-dimensional data, although it can be greatly affected by this when selecting initialisation centroids and its assumption that the clusters are spherical and of similar size.

Keywords

Algoritmo, KNN, K-means y Maching Learning

1. Introducción

El Maching Learning (aprendizaje automatico) es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos que le permites a las maquinas aprender a partir de datos o instrucciones que la persona le pide que realice, ya sea para realizar predicciones o tomar algunas decisiones basadas en patrones observados. En el vasto campo de aprendizaje automático, los algoritmos de clasificación y agrupamiento desempeñan un papel fundamental en la extracción de información significativa a partir de grandes conjuntos de datos. En este documento vamos a explorar dos algoritmos fundamentales que son KNN (K-Nearet Neighbors) y K-Means. Los algoritmos de KNN es catalogado como un clasificador basado en instancias. El cual sirve para clasificar, comparar las instancias no vistas con aquellas etiquetas del conjunto de entrenamiento utilizando similitud.Maillo et al. (2018). Este es uno de los diez algoritmos de clasificación que es el mas relevante. Sin embargo, todos los vecinos igual de importantes en la clasificación son considerados, la condición de frontera que se tiene entre clases las cuales están perfectamente definidas. Estas propuestas se basan mucho en la teoría de los conjuntos difusos que abordan el problema que estamos analizando en el momento. Los algoritmos de KMeans es el algoritmo de clustering más conocido y utilizado ya que es eficaz a pesar de su simple aplicación. Son algoritmos no supervisados que se utilizan para problemas de clustering, el cual también sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado numero de K de clústeres que previamente se le determinan. El nombre de K-means represente cada uno de los clusters por la media de los puntos que están establecidos, es decir por su centroide.Alberto et al. (2016).

Este informe se propone explorar en detalle los fundamentos teóricos, las aplicaciones prácticas y las ventajas y desventajas de los algoritmos KNN y K-Means. A través de ejemplos y estudios de caso, se demostrará cómo estos algoritmos pueden ser implementados y optimizados para resolver problemas complejos en diversas disciplinas.

2. Marco Teorico

2.1. Aprencizaje Inductivo Supervisado

2.1.1. ***KNN o K-vecinos más cercanos***

Los algoritmos de k vecinos más cercanos, también conocidos como KNN o K-NN es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual, que por vecindad están basadas en la búsqueda de un conjunto de prototipos ya clasificados que se encuentran mas cercanos al elemento a clasificar. Para poder a cabo esta clasificación de datos primero se debe especificar una métrica para poder llevar a cabo la medición de proximidad entre los vecinos y el elemento a clasificar, generalmente se utiliza la distancia euclidiana.Madariaga et al. (2022).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(a_i(x) - a_i(y))^2} \tag{1}$$

El valor de n en la ecuación es el numero de atributos totales y a es el valor del atributo en las instancias de x (tupla de prueba) e y. Para la clasificación en este algoritmo se puede ponderar la contribución de cada vecino de acuerdo a la distancia entre el vecino y el ejemplar a ser clasificado (en nuestro caso x) y dando mayor peso a los vecinos mas cercanos. Estos cálculos se muestran en la siguiente ecuación.

$$c(x) = \arg\max \sum_{c \in C=1}^{k} \delta(c, c(y_i)) \tag{2}$$

Distancia Manhattan (p=1): Esta es también otra de las métricas de distancia popular que utiliza estos algoritmos, que mide el valor absoluto entre dos puntos. Tambien se conoce como distancia taxi o distancia cuadrada de la ciudad, ya que comúnmente se visualiza con una cuadricula, que ilustra como se puede navegar de una dirección a otra.

$$d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|\right) \tag{3}$$

Distancia Minkowski: Esta medidad es la forma generalizada de las métricas de distancia Euclidiana y Manhattan, donde el parámetro p, en la formula a continuación permite la creación de otras métricas de distancia.

$$\left(\sum_{i=1}^{m}|x_i - y_i|\right)^{\frac{1}{p}} \tag{4}$$

Distancia de haming: Esta técnica se usa típicamente con vectores booleanos o de cadena, identificando los puntos donde los vectores no coinciden. Como resultado, también se la conoce como la métrica de superposición lo cual lo

podemos ver en la siguiente ecuación:

$$D_H = \left(\sum_{i=1}^{k} |x_i - y_i| \right) \tag{5}$$

2.2. Ventajas y desventajas del algoritmo KNN o K-vecinos cercanos

2.2.1. ***Ventajas***

- Es facil de implementar.
- Se adapta facil a lo que se necesita que realicemos en el programa o lo que estemos buscando.
- Es tolerante al ruido.
- Se pueden aprender conceptos complejos usando funciones sencilllas como aproximaciones locales.

2.2.2. ***Desventajas***

- No realiza bien el escalamiento
- El coste de encontrar los K mejores vecinos es grande
- Es muy propenso a sobre ajustarse
- El rendimiento de este tipo de algoritmos baja si el número de descriptores crece

2.3. Aprencizaje Inductivo no Supervisado

2.3.1. ***K-means***

El algoritmo K-means es un algoritmo no supervisado de clustering más conocido y utilizado ya que es eficaz a pesar de su simple aplicacióny se uyiliza cuando tenemos un montón de datos sin etiquetar. Este sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado numero de K de clústeres, previamente determinado. Este algoritmo representa cada uno de los clusters por la media de los puntos, es decir, por su centroide, que es uno para cada grupo que se analiza. Estos centros se deben colocar de manera astuta porque la ubicación de estos si es diferente provoca unos resultados muy diferentes. Por lo tanto, la mejor opción para colocar estos puntos es lo mas lejos posible entre sí. Torres (2023).

Los objetivos se representan con vectores reales de d dimensiones (x1,x2,. . . ,xn) y el algoritmo K-means construye K grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo S=S1,S2,. . . ,Sk, a su centroide . El problema se puede formular de la siguiente manera;

$$\text{mín}\, E(\mu_i) = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{6}$$

Donde la **S** es el conjunto de datos cuyos elementos son los objetos $\mathbf{X_j}$ representados por los vectores, donde cada uno de sus elementos representa una característica o atributo. Tendremos **k** grupos o clusters con correspondiente centrode $\mu_\mathbf{i}$.

2.4. Ventajas y desventajas del algoritmo K-means

2.4.1. ***Ventajas***

- Es un algoritmo muy eficiente computacionalmente y adecuado para conjunto de datos demasiado grandes, debido a que el algoritmo solo requiere unas pocas operaciones sencillas en cada iteración.
- Es demasiado fácil de entender e implementar, ya que no requiere conocimientos matemáticos o estadísticos avanzados.
- Este puede manejar datos con un gran número de dimensiones, K-means es capaz de encontrar patrones y estructuras en datos de alta dimensionalidad.

2.4.2. ***Desventajas***

- El algoritmo depende de la selección inicial de los centroides, lo que puede afectar los resultados finales de los agrupamientos de los datos.
- Este asume que los centroides son esféricos, lo que puede llevar a asignaciones incorrectas cuando los clústeres no los son.
- Este cuenta con dificultades para identificar clusteres de tamaños y densidades variables, esto se debe a que el algoritmo asigna puntos de datos al centroide más cercano.

2.5. Eleccion de los K y los centroide en K-means

- Si la K es muy pequeña, se agruparán en distintos grupos
- Si la L es muy grande, hay centros que pueden quedar huérfanos o sin agruparce
- El valor de la K puede determinarse según la heurística

Por lo consiguiente, para poder lograr un **k** optimo o una aproximación concluyente, se optara por realizar algunas pruebas con los datos que se van a utilizar, para así poder analizar los resultados, lograr estimar de mejor manera la variable de **k**.Ahmed et al. (2020).

Longitude	Latitude	Bedrooms
-122.23	37.88	129.0
-122.22	37.86	1106.0
-122.24	37.85	190.0
-122.25	37.85	235.0
-122.25	37.85	280.0
-122.25	37.85	213.0
-122.25	37.84	489.0
-122.25	37.84	687.0
-122.26	37.84	665.0
-122.25	37.84	707.0
-122.26	37.85	434.0
-122.26	37.85	752.0
-122.26	37.85	474.0
-122.26	37.84	191.0

Cuadro 1: Algunos valores de los datos que se analizaron

3. Resultados de los algoritmos

3.1. Resultados del agoritmo de KNN

Lo cual en este se utilizo la matriz de confuion, tambien conocida como la matriz de error la cual nos permitio la visualizacion del rendimiento de nuetro algoritmo, el cual se trataba de un aprendizaje supervisado. Cada fila de la matriz representa las instancias de la clase predicha, mientras que las columnas representa las instancias de la clase real.(Cuadro 2)

En esta grafica podemos observar el resultado de los clientes que si pagaron y cuales no pagaron su crédito, y también podemos ver la edad de los clientes que realizaron esta actividad, lo cual nos ayuda a hacernos una idea de cuales fueron las personas que hicieron bien el pago de su crédito. (Imagen 1).

	Realmente es positivo	Realmente es negativo
Predicho como positivo	Verdaderos Positivos(VP)	Falsos Positivos(FP)
Predicho como negativo	Falsos Negativos(FN)	Verdaderos Negativos(VN)

Cuadro 2: Matriz de confusión

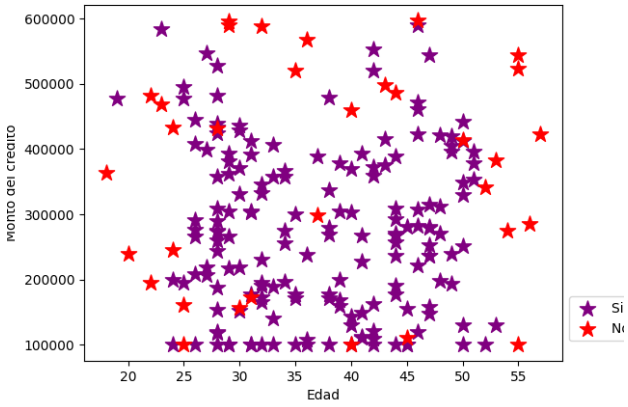

Figura 1: Los clientes que pagaron y cuales no pagaron el credito

3.2. Resultado del algoritmo de K-means

En la siguiente tabla podemos observar alguno de los campos que analizamos en la base de datos que utilizamos de la cual solo tomamos una cierta cantidad de datos que pueden observar en la siguiente tabla:(Cuadro 1).

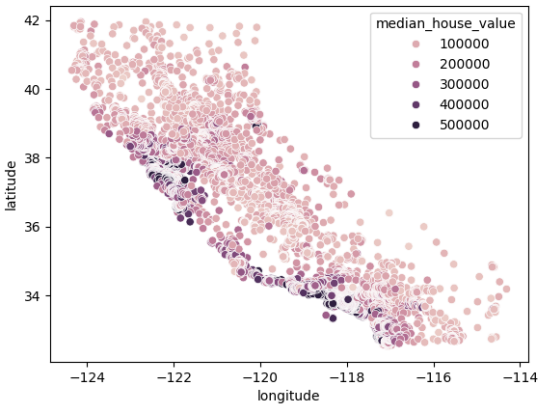

Figura 2: Coordenadas de longitud y latitud de los datos de las viviendas

Con el total de los datos que tenía la tabla que acabamos de ver anteriormente, pudimos sacar de esta base de datos, los resultados que obtuvimos en la grafica, del mapa el cual no lo arrojo el código que implementamos para calcular estos puntos.(Imagen 2)

Después de graficar este gráfico del coeficiente de silueta para diferentes valores de k, podemos analizar los resultados para determinar el número óptimo de clústeres para nuestros datos.(Imagen 3)

Finalmente, obtuvimos las etiquetas de los clusteres y los centroides, las cuales nos sirvieron para la configuracion de las etiquetas del grafico,

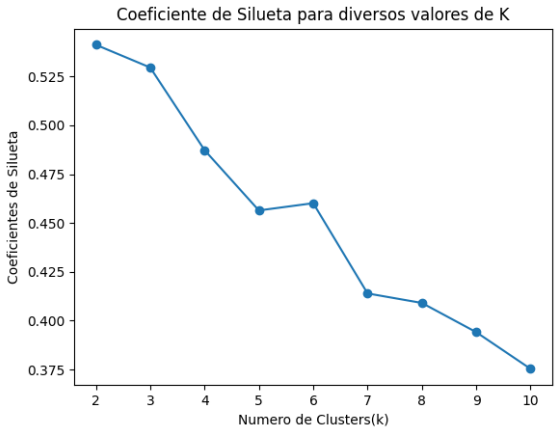

Figura 3: Podemos obserrvar los valores que podemos utilizar de K

las cuales nos ayudaron para mostrar el muestreo de nuentro grafico y de los datos que analizamos, esto lo podemos observar en la siguiente grafica.(Imagen 4)

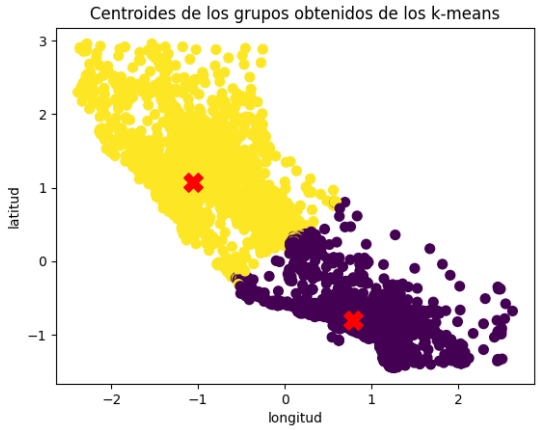

Figura 4: Observamos la configuracion de las etiquetas del grafico

4. Conclusiones

En la elaboración de ambos algoritmos pudimos observar la gran diferencia que tienen estos dos algoritmos, los cuales tienen sus ventajas y desventajas en ciertos aspectos, los cuales pueden resolver problemas en diversas disciplinas. En los KNN vimos que son una herramienta muy poderosa para tareas de clasificación, y es muy útil cunado se necesita una implementación simple y rápida. Sin embargo, el rendimiento de este algoritmo se puede degradar con conjuntos de datos demasiados grandes y de alta dimensionalidad. Por otro lado, tenemos los algoritmos K-mean que son ideales para problemas no supervisados, proporcionando una forma eficiente de descubrir patrones y estructuras en datos sin etiquetar. Su simplicidad y eficacia lo hacen popular, pero su sensibilidad a la inicialización de centroides y suposiciones sobre la forma de los clústeres limitan su aplicabilidad en algunos casos, al seleccionar las K y la inicialización de los centroides de los K-means son muy cruciales para obtener buenos resultados. Pruebas y análisis de los datos específicos son muy necesarios para poder determinar los valores óptimos con los cuales va a funcionar bien el algoritmo.

Referencias

Ahmed, Mohiuddin, Raihan Seraj & Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. 10.3390/electronics9081295

Alberto, Felipe, Cifuentes Ramos, Profesor Guía, Rodrigo Alfaro Arancibia & Wenceslao Palma Muñoz. 2016. Clasificación automática de tweets utilizando k-nn y k-means como algoritmos de clasificación automática, aplicando tf-idf y tf-rfl para las ponderaciones

Madariaga, Carlos Jesus, Yosvani Orlando Lao, Dagnier Antonio Curra & Rafael Martin. 2022. Empleo de algoritmos knn en metodología multicriterio para la. *Universidad de Holguín, Cuba* 1–21

Maillo, Jesus, Julián Luengo, Salvador García, Francisco Herrera & Isaac Triguero. 2018. Un enfoque aproximado para acelerar el algoritmo de clasificación fuzzy knn para big data.

Torres, Raul. 2023. ClasificaciÓn de clientes y prediccIÓn de deserciones usando algoritmos k-means y regresiÓn logÍstica. *PONTIFICIA UNIVERSIDAD CATO**********LICA DEL ECUADOR* 1–38