Curso

7

Machine learning & big data

Supervised Learning: Clasificador K-Nearest Neighbor

Algoritmos Supervisados

josé nelson zepeda doño

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.

Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda

MIP • V 1.0

San Salvador El Salvador

Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

[Algoritmo KNN 1](#_Toc4542888)

[Limitaciones del KNN 8](#_Toc4542889)

[Aspectos para seleccionar K 9](#_Toc4542890)

[Bibliografía 10](#_Toc4542891)

Capítulo

1

# Algoritmo KNN

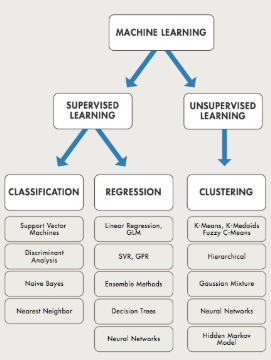
Dime con quién andas y te diré quien eres.

E

l algoritmo KNN (K Nearest Neighbors) fue diseñado en 1951 por Fix y Hodges, pertenece a la familia de algoritmos supervisados y su característica principal es la simplicidad.

Este algoritmo puede recibir diferentes nombres:

* KNN
* Memory based Reasoning
* Example based Reasoning
* Instance based Learning
* Lazy Learning



Fgura 1-1 Familias de algoritmos

La idea detrás del algoritmo es muy simple pero poderosa, el algoritmo clasifica cada dato nuevo en el grupo que corresponda, según K vecinos más cerca de un grupo o de otro.

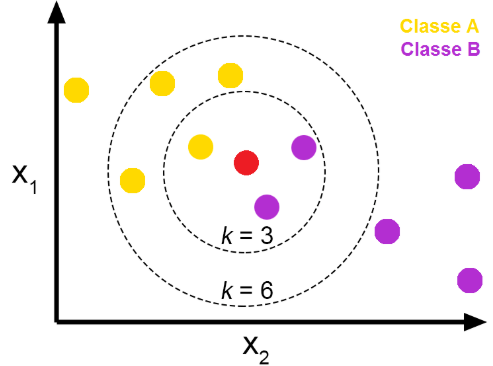


Figura 1-2 KNN Representación[[1]](#footnote-1)

Para llevar a cabo lo descrito en el párrafo anterior, el algoritmo calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias en forma ascendente es decir de menor a mayor para ir seleccionando el grupo al cual pertenece dicha observación. El grupo será por lo tanto el de mayor frecuencia con menores distancias.

Tal como se indicó antes, por ser un algoritmo supervisado, a partir de un conjunto de datos de entrenamiento, se construirá un artefacto cuyo objetivo será el de clasificar correctamente todas las nuevas instancias y vale la pena mencionar desde el inicio que los conjuntos de datos a utilizar pueden incluir atributos tanto numéricos como descriptivos.

Por su forma de aprendizaje también se puede afirmar que el algoritmo pertenece a la familia de *Lazy Learning Methods (Aprendizaje Vago),* el cual es un método de aprendizaje en el que la generalización más allá de los datos de entrenamiento es demorada hasta que se hace una pregunta al sistema, al contrario que en el aprendizaje entusiasta, donde el sistema intenta generalizar los datos de entrenamiento antes de recibir preguntas[[2]](#footnote-2).

Como funciona KNN[[3]](#footnote-3)

El algoritmo de los K Vecinos más cercanos es uno de los algoritmos basados en distancias y se fundamenta en el supuesto de que todas las distancias se encuentran en un espacio de dimensión n.

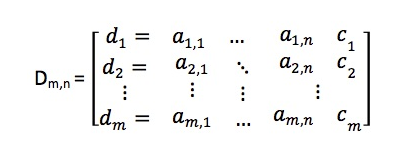
El vecino más cercano de una instancia es definido en términos de la distancia, popularmente se utiliza la Distancia Euclidiana Estándar, pero tal como se ha visto en otros estudios, existen más distancias disponibles para estimar esta métrica.

La distancia Euclidiana se calcula mediante la siguiente ecuación:



Por ejemplo, pensemos en una base de datos en donde cada registro corresponde a un cliente de una entidad financiera, los campos de la base de datos pueden ser edad, ingresos, gastos, patrimonio, saldo, estado civil, etc. Y también para cada cliente tenemos una variable objetivo que es la que queremos predecir: Cliente de Bajo Riesgo Financiero o Cliente de Alto Riesgo Financiero.

Sea D la base de datos, podríamos representar la estructura resultante como:



En donde dm indica cada una de las distancias de los registros de la base de datos, amn indica cada uno de los atributos de cada cliente y Cm indica la clase a predecir.

Luego se fija un valor K que corresponde a la cantidad de vecinos que vamos a evaluar, el algoritmo selecciona las K instancias de la base de datos de entrenamiento más cercanas (de acuerdo con la métrica de similitud utilizada) y se le asigna a nuestra observación la clase más frecuente de entre las K instancias seleccionadas como más cercanas.

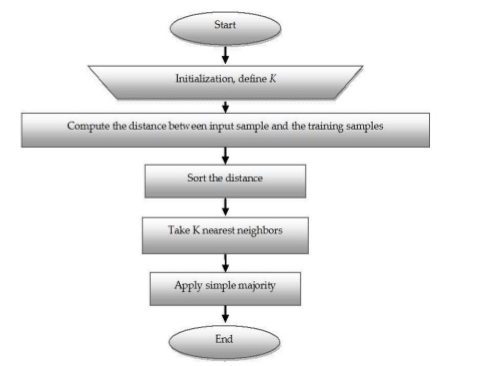


Figura 1-3 Agoritmo KNN

Hay que destacar dos puntos muy importantes:

* El algoritmo es altamente sensible a la variable K, con valores distintos de K podremos obtener también resultados muy distintos en la predicción. El valor de K suele establecerse luego de una serie de pruebas con varias instancias.
* La métrica de similitud utilizada: uno de los pilares de este algoritmo pues influye directamente en las relaciones de cercanía que se irán estableciendo en el proceso de construcción del algoritmo.

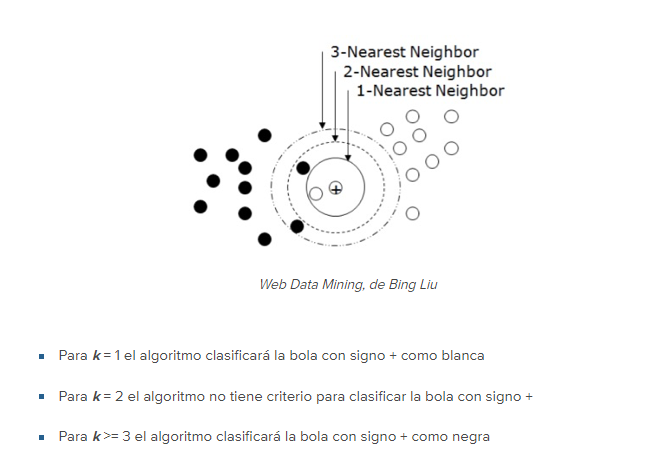


Figura 1-4 Influencia del valor de K

Su mayor debilidad es la lentitud en el proceso de clasificación puesto que su objetivo no es obtener un modelo optimizado, sino que cada instancia de prueba es comparada contra todo el juego de datos de entrenamiento y, será la bondad de los resultados lo que determinará el ajuste de aspectos del algoritmo como el propio valor k, el criterio de selección de instancias para formar parte del juego de datos “D” de entrenamiento o la propia métrica de medida de similitud.

Y para cerrar esta sección, vale la pena mencionar que a diferencia de K-means, que es un algoritmo no supervisado y donde la “K” significa la cantidad de “grupos” (clusters) que deseamos clasificar, en K-Nearest Neighbor la “K” significa la cantidad de “puntos vecinos” que tenemos en cuenta en las cercanías para clasificar los “n” grupos -que ya se conocen de antemano, pues es un algoritmo supervisado.

Tipos de Distancias

El concepto de distancia entre objetos o individuos permite interpretar geométricamente muchas técnicas clásicas del análisis multivariante, equivalentes a representar estos objetos como puntos de un espacio métrico adecuado.

Esta interpretación es posible no solamente cuando se dispone de variables cuantitativas, sino también, y sobretodo, cuando las variables observadas son de tipo más general, o incluso cuando no se dispone de variables propiamente dichas, siempre que tenga sentido obtener una medida de proximidad entre los objetos o individuos[[4]](#footnote-4).

Si Euclides, el padre de la geometría, hubiese conocido Manhattan, posiblemente le habría hecho dudar sobre aquello que decimos constantemente de que «la distancia más corta entre dos puntos siempre es la línea recta». Ya sea en la Gran Manzana o en cualquier otra ciudad, si para diseñar una ruta en una zona donde haya edificios usamos la línea recta para unir el origen y el destino de la misma, nos encontraremos con el pequeño inconveniente de que nuestra ruta debe atravesar algún inmueble.

En la geometría que nos enseñó el gran matemático griego –que es la que aprendemos en los primeros años de la escuela– la distancia entre dos puntos se mide como la longitud del segmento que los une. Esta forma de medir la distancia es conocida como «distancia euclídea» y tiene infinidad de aplicaciones en matemáticas y en nuestra vida cotidiana. Sin embargo, no es la única forma de hacerlo y no siempre es real en según qué problemas queramos resolver: por ejemplo, el de la ruta mínima en medio de una ciudad. Si los puntos que debemos conectar con el camino más corto están cada uno de ellos en un lateral de una manzana de edificios, el camino más corto entre ellos vendrá dado por la suma de los dos catetos de un triángulo rectángulo[[5]](#footnote-5).

Es a esta suma de longitudes a lo que llamamos «distancia Manhattan» (o más formalmente, distancia L1). Es mayor que la distancia euclídea pero también es más real en la práctica. De hecho, la distancia Manhattan entre dos puntos se calcula como la longitud de cualquier camino que los una mediante segmentos verticales y horizontales; todos miden lo mismo.

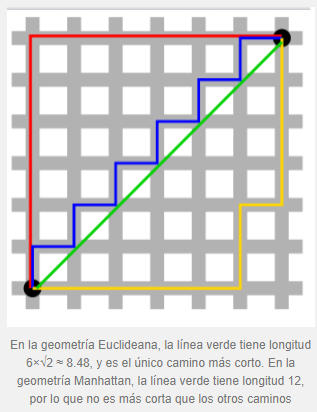
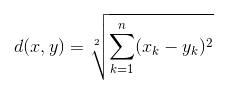


Figura 1-5 Distancia Manhattan

A continuación se muestran las ecuaciones de las distancias antes descritas:

* Distancia Euclidiana: En matemáticas, la distancia euclidiana o euclídea es la distancia "ordinaria" entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras.



* Distancia Manhattan: también conocida como geometría del taxista, distancia rectilínea o longitud de Manhattan

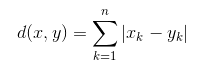
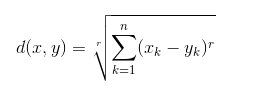




Figura 1-5 Tipos de distancias

* Distancia Minkowski



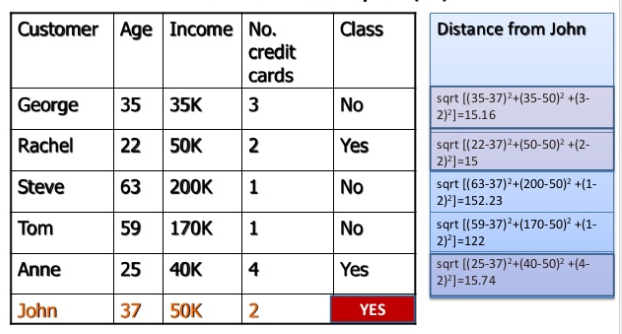


Figura 1-6 ejemplo de cálculo de distancia

## Limitaciones del KNN

* El algoritmo es extremadamente sensible a los atributos irrelevantes y a la multidimensionalidad.
* Sensible al ruido.
* Es lento si hay muchos datos de entrenamiento.
* La función de distancia utilizada puede inferir en el resultado.
* Altos requerimientos de memoria
* KNN no aprende ningún modelo

Por lo antes descrito, para una buena aplicación del algoritmo se vuelve fundamental la estandarización/normalización de los datos numéricos para evitar que los atributos con mucho rango no tengan más peso que los demás.

## Aspectos para seleccionar K

* Si el valor de K es muy pequeño, el resultado será muy sensible al ruido.
* Si el valor de K es grande, trabajara mucho mejor, pero se perderá la idea de localidad.
* Para evitar que los vecinos lejanos tengan mucha influencia, se puede hacer que cada vecino vote de manera inversamente proporcional a la distancia 1/d.
* Se puede utilizar un valor de K impar para deshacer empates.

# Bibliografía

* Machine Learning for Beginners

By Ken Richards, 2017

* Modelado mediante RF de las emisiones de autobuses urbanos en función de los ciclos cinemáticos

By Victor Pita González-Campos , 2017

* R Data Analysis Cookbook

by Kuntal Ganguly,2017

* Making Sense of Data: A Practical Guide to Exploratory Data Analysis

by W Johnson ,2007

1. <https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d> [↑](#footnote-ref-1)
2. <https://es.wikipedia.org/wiki/Aprendizaje_vago> [↑](#footnote-ref-2)
3. <https://www.analiticaweb.es/algoritmo-knn-modelado-datos/> [↑](#footnote-ref-3)
4. Distancias Estadísticas y Escalado Multidimensional, Aurea Grané, Universidad Carlos III Madrid [↑](#footnote-ref-4)
5. <https://metode.es/revistas-metode/secciones/cajon-de-ciencia/les-distancies-a-manhattan.html> [↑](#footnote-ref-5)