INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1263)
Available online at: https://www.ijariit.com

What kind of algorithms are applied in search engines?

Kush Sharma
kushsharma024@gmail.com
Ryan International School, Ghaziabad, Uttar Pradesh

ABSTRACT

This paper aims to make readers understand about types of algorithms used in different search engines and e-commerce websites

Keywords— *Computer Science, Data Science, K-NN, K Nearest Neighbor Algorithm, Predictions*

1. INTRODUCTION

This paper is about how recommendation engines work and what type of algorithms are used by them in order to suggest us which movie to watch, which music to listen, what book to read and what clothes to buy. This phenomenon can be observed by us when we listen to music on applications such "Spotify" or stream movies on "Netflix". We often wonder what makes these websites know, what we will like or what we will prefer to listen. One such question struck to my mind when I was feeling nostalgic and streamed an old song on Spotify and then I observed that one after another it started recommending all the songs that I used to love as a child. After 3 hours of Google search (which is also a recommendation as well as search engine) I found out that all these recommendation engines are a part of Machine Learning and are generally based on Algorithms such as K-nearest Neighbor Algorithm. This paper however talks about recommendation engines but to understand about that more deeply we will have to focus on "K-nearest neighbor algorithm" and during that we are going to see how this algorithm was created and by whom it was crated , how it is being used by us in our general life and what effects it withholds on us, what are some of the positive sides of using this algorithm, including technical advantages and time efficiency and after that we will learn about some of the cons of this algorithms with some alternatives of this algorithm and after that we will be learn about its functioning by manually implementing it on a random data set as part of sample project.

2. ALL ABOUT K-NEAREST NEIGHBOR ALGORITHM

The K-nearest algorithm was developed by Thomas M. Cover (7th Aug, 1938- 26th Mar, 2012) who was an information theorist and professor in the jointly Departments of Electrical Engineering and Statistics at Stanford university as a non-parametric method of classification and regression, here Non parametric are statistics based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified. Till now what I understood about k-nearest algorithm was that it basically helps you in finding the K number of nearest neighbors of a reference a reference point.

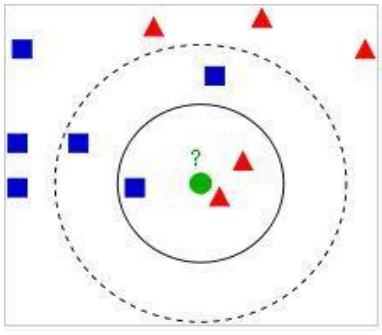

Image 1

For example purposes let us refer to image(1) In this image we have chosen reference point as the green dot(g), so if we point out find 3 nearest neighbour of they will be 1 blue square and 2 red triangles which are confined in the solid circle but if we are asked

to point out 5 nearest neighbors of green dot(g) we will find out 2 red triangles and 3 blue squares and just like that using K-nearest neighbors algorithm we can find the nearest value of our input value that is a user defined constant and just like that it can be implemented on plenty numbers of database as well as situations. Here for distance metric Euclidean distance is used however for discrete variables for such as text classification another metric can be used such as overlap metric or what we call Hamming distance

3. USES OF K- NEAREST NEIGHBOR (KNN) ALGORITHM

KNN or k-nearest neighbour algorithm is used in many of the apps that we use in our daily life, however it works in mysterious ways that we don't understand and to explain that only I have written this paper, so exactly what those apps are which uses KNN? Answer to this question is very easy, apps that use recommendation system on the basis of present data or apps that helps the user in finding the closest results available according to the input parameters. So basically, apps or websites like Spotify, Netflix and Prime videos uses some type of KNN algorithm for recommending different songs, shows and movies to their users and for this they uses your previous choices, we can understand this situation by an example: let us assume that there are two different users "A" and "B". A is new to Netflix however B is an old user, now in this given situation we will find that B is most likely to see a show or movie of their choice however on the other hand A is won't be able to see any such huge diversity of shows or movies of their choice. And this is because Netflix doesn't know anything about user A and don't know their preferences however over time when Netflix gather's data about user A's preferences. They will observe that they are getting content of their choice.

Image 2: Netflix Logo

Now the following parameters can used to sort the content for the user:

* Timeline of the content searched by the user
* Type of content (for example)
    1) Action
    2) Fiction
    3) Documentaries
    4) Comedy
    5) Romantic content
* Length of the content

And many more hundreds of parameters are used to sort hundreds and thousands of contents for billions of users. If we talk about specific use of the above mentioned parameters then we can say that Timeline of the content can help the algorithm to find the content of some specific time period for the user, type of content helps in the finding the content that suits the genre that the user like and length of content can be used to determine the content that is of the length which user generally likes as some people like short films however others like long documentaries. Uses of K-NN are not just limited to suggesting movies and music, it can also be used in banking system to predict if a person is likely to be a bank defaulter.

Image 3: Spotify Logo

As there is so much applicability of K-NN it all must be due to advantages/ features it provides, so in the next chapter we will be discussing about the various features/advantages that K-NN provides

4. ADVANTAGES OF K-NN OVER OTHER DATA ANALYTIC ALGORITHMS

Using K-NN algorithm has some of its own advantages such as its simplicity, flexibility with decision boundaries and its speed, not only this much K-NN has other different advantages but we will describe them separately. So first of all talking about simplicity, K-NN is very easy/simple to use and implement and intuitive to understand and in addition to that K-NN doesn't require any training time and all its calculations and predictions takes place during its execution which also makes it time efficient and not only this during regression and classification we can easily change the boundaries just by adjusting the value of 'K' and since there is a single hyperparameter 'K' its tuning also becomes easy also as we have mentioned earlier since there is no explicit training so we add

more data to the existing data without changing the constant or retraining a new model and there are other additional advantages such as privilege of choosing distance matrix from various distance matrices such Euclidean, Manhattan, Minkowski, Hamming distance and so on.

In addition to the above mentioned advantages K-NN provides high accuracy which is one of the main requirement of any data analysing algorithm, but the thing to remember is that if the value of K approaches 1 it becomes very unstable, so value of K should always be high but if there are increasing numbers of errors then it means that value of K is being pushed very high. While having many advantages K-NN algorithms has its many disadvantages which acts as a hindrance in its practical applicability and we are going to discuss those disadvantages in the next chapter

5. DISADVANTAGES OF K-NN

Just like there are two sides of coin there are both advantages and disadvantages of using K-NN but since we've discussed about its advantages let's discuss about its disadvantages such as speed, however we may have mentioned that K-NN is fast but it lacks speed while dealing with a very large data set and not only speed there can be compromise with the accuracy of the of results or predictions. There can be several reasons behind the accuracy issue of K-NN and it is not only limited to size of data set but is also depended on number of dimensions, K-NN struggles to with deal a huge number of dimensions or variables and hence providing inaccurate results, now coming onto the scale part as we have already mentioned that K-NN is flexible to use with many types of distance matrix such as Euclidean, Manhattan and etc, but along with that there is a restriction of inserting the input value only in the distance matrix that is used to build the program, for example: if the program is built using the Euclidean distance matrix then the initial values should also in the form Euclidean distance.

There are several other disadvantages of using K-NN such as sensitivity to outliers, K-NN is sensitive to outliers, noisy data and missing values as it is mainly based on one dimension that is the distance of neighbouring values from "K" so we are required to manually impute missing values and remove the outliers from the data and keeping all these disadvantages aside one of the main cause of headache caused due to use of K-NN is deciding the value of "K" since it plays the major role in giving the predictions/ results.

In-spite of all these disadvantages K-NN is quite solid to use, so in the next chapter we will make a sample project using K-NN .

6. MANUALLY IMPLEMENTING K-NN

Note: In order to make this project/program one should have a basic understanding of mathematics.
Aim: To predict the size of jeans of Kush with Height 175cms and weight 71kgs
Steps:
Step 1: Make a random data spreadsheet.

Height(in cms)	Weight(in kgs)	Size of jeans
180	80	L
172	72	L
169	68	M
152	70	S
143	65	S
124	75	XS
164	68	M
162	74	L
168	68	M
173	63	M

Step 2: Using the Euclidean distance formula, find the distance.
Where the formula is:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Height(in cms)	Weight(in kgs)	Size of jeans	Euclidean distance
180	80	L	10.29563014
172	72	L	3.16227766
169	68	M	6.708203932
170	69	S	5.385164807
171	70	S	4.123105626
172	76	L	5.830951895
173	75	L	4.472135955
174	73	L	2.236067977
175	74	M	3
176	75	L	4.123105626

Step 3: Select a value of K. For example, here we have taken K=5
Step 4: Rank the distance values in ascending order.

Height(in cms)	Weight(in kgs)	Size of jeans	Euclidean distance	Rank
180	80	L	10.29563014	
172	72	L	3.16227766	3
169	68	M	6.708203932	
170	69	S	5.385164807	
171	70	S	4.123105626	4
172	76	L	5.830951895	
173	75	L	4.472135955	5
174	73	L	2.236067977	1
175	74	M	3	2
176	75	L	4.123105626	

Step 5: Count the votes of all the K neighbours

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Height(in cms)		Weight(in kgs)		Size of jeans		Euclidean distance		Rank								
2	180		80		L		10.29563014										
3	172		72		L		3.16227766		3								
4	169		68		M		6.708203932			FOR K=5							
5	170		69		S		5.385164807			LET US RANK THE VALUES IN ASCENDING ORDER							
6	171		70		S		4.123105626		4	SINCE MAXIMUM NUMBER OF THE SIZES OF JEANS ARE LARGE FOR VALUE OF K= 5							
7	172		76		L		5.830951895			HENCE ACCORDING TO K-NN KUSH SHOULD BUY A JEANS LARGE IN SIZE							
8	173		75		L		4.472135955		5								
9	174		73		L		2.236067977		1								
10	175		74		M		3		2								
11	176		75		L		4.123105626										

Observation: Since most of the sizes of jeans are of size large for K=5, Kush should but a jeans of large size.

7. CONCLUSION

As of now I believe and expect that all of you, who have read this paper have understood the basic concept of K-NN, but to be honest you'll be able to fully understand the concept of K-NN algorithm once you try/apply it practically by yourself. I wish you all the best for your future endeavours.

