Different Methods to Calculate the Distance in KNN Algorithm

There are a lot of ways to calculate the distance between two points, like Manhattan and Euclidean. Different methods are useful in different ways.

Minkowski Distance:

Minkowski distance is the generalized distance metric. This means, we can use different methods to calculate the difference when setting different values for p.

If P=1, this is Manhattan Distance.

If P=2, this is Euclidean Distance.

If P =∞, this is Chebychev Distance.

Manhattan Distance:

Manhattan Distance is the sum of the absolute difference between coordinates. It finds the elements only in vertical or horizontal dimensions, like analyzing the cheeseboard-thing. Therefore, it is more used in map science and some related fields.

Euclidean Distance:

In most cases, when we talk about the distance, the first method come into our minds is the Euclidean Distance. Therefore, this is the methods that has been widely used in people’s life. When data is dense or continuous, this is the approximately best way to calculate the distance.

This method takes into account all the observations including the outlier or some redundant variables. And, we also need to normalize the data to get more reasonable results. Therefore, we need to be careful when using this method.

Chebychev Distance:

This method compares each variable’s difference and choose the maximum. According to my knowledge, I have took the Project Management and Operation Research courses, it is used to estimate the maximum time or route when preparing the business plan. If I can know the maximum time that I may use for this project or the longest road that I happen or by accident to choose, then I can estimate what might the worst situation be, and I can have a better back-up plan to deal with this situation.

These are methods I have used in Minkowski Distance to calculate the distance.

Jaccard Distance:

The Jaccard Index is measurement to compare the similarity between two sets to check what they have in common and what they have is distinct.

The Jaccard Distance is to measure how the two sets are dissimilar.

In my opinion, it is used useful to do some corrections when doing some experiments or manufacturing. Because we have some different features to compare between the new one and the standard one. And we may find what their differences and can better improve the production procedure or the experiments.

I think it is useful to use this method to compare the feature to get the estimated prices. Because it may consider all the features into consideration and choose the most closed one(or some) as the price.

Cosine Distance:

Cosine Similarity is a metric used to determine how two sets are similar.

This method is more advantageous than some other methods, because it ignores the size of each sets and get more accurate results when size does not matter.

This method is widely used in text or document analysis, for example, to count the frequency of words and find the most frequent word.

Mahalanobis Distance:

x is the vector of the observation.

m is the vector of mean values of independent variables.

c is the inverse covariance matrix of independent variables.

Mahalanobis Distance is used for calculating the distance between two data points in a multivariate space. The advantage of using this method is that it takes into account of the covariance and helps to measure the similarity.

However, in my opinion, this method is not useful in KNN algorithm. Because we can just get each test data’s distance separately, and may not help us to use the minimal distance to find the estimated price.