Name: Ankit Sharma

Email: kumarankitx022@gmail.com

Mob no.: +91-7677241423

Week: 04

ASSIGNMENT – 10

TOPIC: K-NEAREST NEIGHBORS (K-NN)

Q.1. Distances which we can use in K-NN model and also learn about mathematical formula for those distances.

Q.2. Alter the n_neighbors value and see the difference in the output and also change the type of distance by altering the p value of the distance metric.

Ans. 1) A number of Machine Learning Algorithms - Supervised or Unsupervised, use Distance Metrics to know the input data pattern in order to make any Data Based decision. A good distance metric helps in improving the performance of Classification, Clustering and Information Retrieval process significantly.

Below are the commonly used distance metrics –

Minkowski Distance: Minkowski distance is a metric in Normed vector space. What is Normed vector space? A Normed vector space is a vector space on which a norm is defined. Suppose X is a vector space then a norm on X is a real valued function $||x||$ which satisfies below conditions -

1. **Zero Vector-** Zero vector will have zero length.
2. **Scalar Factor-** The direction of vector doesn't change when you multiply it with a positive number though its length will be changed.
3. **Triangle Inequality-** If distance is a norm then the calculated distance between two points will always be a straight line.

The distance can be calculated using below formula -

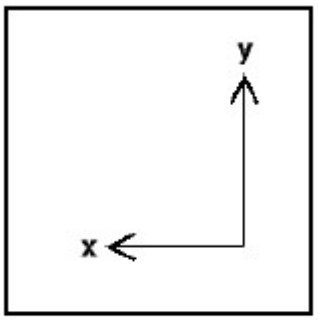
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Minkowski distance is the generalized distance metric. Here generalized means that we can manipulate the above formula to calculate the distance between two data points in different ways.

As mentioned above, we can manipulate the value of *p* and calculate the distance in three different ways-
p = 1, Manhattan Distance
p = 2, Euclidean Distance
p = ∞, Chebychev Distance

Manhattan Distance:
We use Manhattan Distance if we need to calculate the distance between two data points in a grid like path. As mentioned above, we use Minkowski distance formula to find Manhattan distance by setting p's value as 1.

Let's say, we want to calculate the distance, d, between two data points: x and y.



Manhattan

Distance d will be calculated using an absolute sum of difference between its cartesian co-ordinates as below :

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

where, n- number of variables, xi and yi are the variables of vectors x and y respectively, in the two dimensional vector space. i.e.

x = (x1,x2,x3,...) and y = (y1,y2,y3,…).

Now the distance d will be calculated as-

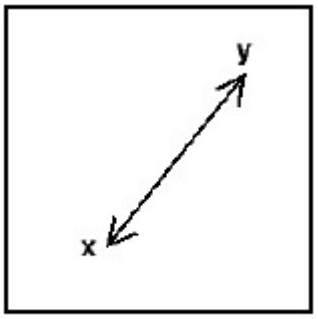
(x1 - y1) + (x2 - y2) + (x3 - y3) + … + (xn - yn).

Euclidean Distance:
Euclidean distance is one of the most used distance metric. It is calculated using Minkowski Distance formula by setting ***p's*** value to ***2***. This will update the distance ***'d'*** formula as below :

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

Let's stop for a while! Does this formula look familiar? Well yes, we just saw this formula above in this article while discussing ***"Pythagorean Theorem".***

Euclidean distance formula can be used to calculate the distance between two data points in a plane.



Euclidean

Cosine Distance:
Mostly Cosine distance metric is used to find similarities between different documents. In cosine metric we measure the degree of angle between two documents/vectors(the term frequencies in different documents collected as metrics). This particular metric is used when the magnitude between vectors does not matter but the orientation.

Cosine similarity formula can be derived from the equation of dot products :-

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Now, you must be thinking which value of cosine angle will be helpful in finding out the similarities.

$$\text{Cos } 0' = 1 \qquad \text{Cos } 90' = 0$$
$$\text{Cos } 180' = -1$$

Now that we have the values which will be considered in order to measure the similarities, we need to know what do 1, 0 and -1 signify.

Here cosine value 1 is for vectors pointing in the same direction i.e. there are similarities between the documents/data points. At zero for orthogonal vectors i.e. Unrelated(some similarity found). Value -1 for vectors pointing in opposite directions(No similarity).