QUIZ 1

1. What is difference between KMeans and KNN ?

K-Means is an Unsupervised learning technique

‘K’ in K-Means is the number of clusters the algorithm is trying to identify/learn from the data. The clusters are often unknown since this is used with Unsupervised learning.

It is typically used for scenarios like understanding the population demomgraphics, market segmentation, social media trends, anomaly detection, etc. where the clusters are unknown to begin with.

---------------------------------------------------------------------------------------------------------------------------

KNN is a Supervised learning technique

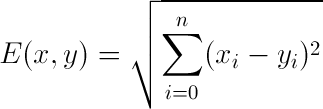
‘K’ in KNN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample

It is used for classification and regression of known data where usually the target attribute/variable is known before hand.

2.What are distance functions used in KNN ?

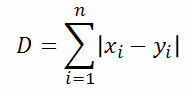
1.Euclidean distance

This algorithm  is only valid for continuous variables.This algorithm is used mostly for regression.  It could be used to measure the "similarity" between two vectors (though you should normalize the data first). The Euclidean is often the "default" distance used in e.g., K-nearest neighbors (classification) or K-means (clustering) to find the "k closest points" of a particular sample point. Another prominent example is hierarchical clustering, agglomerative clustering (complete and single linkage) where you want to find the distance between clusters.



2.Manhattan Algorithm

According to this interesting paper, Manhattan distance (L1 norm) may be preferable to Euclidean distance (L2 norm) for the case of **high dimensional** data in order to improve the results of distance-based algorithms, like clustering



3. Hamming algorithm

**Hamming distance** can be used to measure how many attributes must be changed in order to match one another. We can calculate the ratio to determine the similarity (or difference) between two data points using the **simple matching coefficient**:

4.Cosine distance

If we care about the direction of the data rather than the magnitude, then using the **cosine distance** is a common approach. It computes the dot product of the two data points divided by the product of their magnitude. Cosine distance, together with the term/document matrix, is commonly used to measure the similarity between documents.