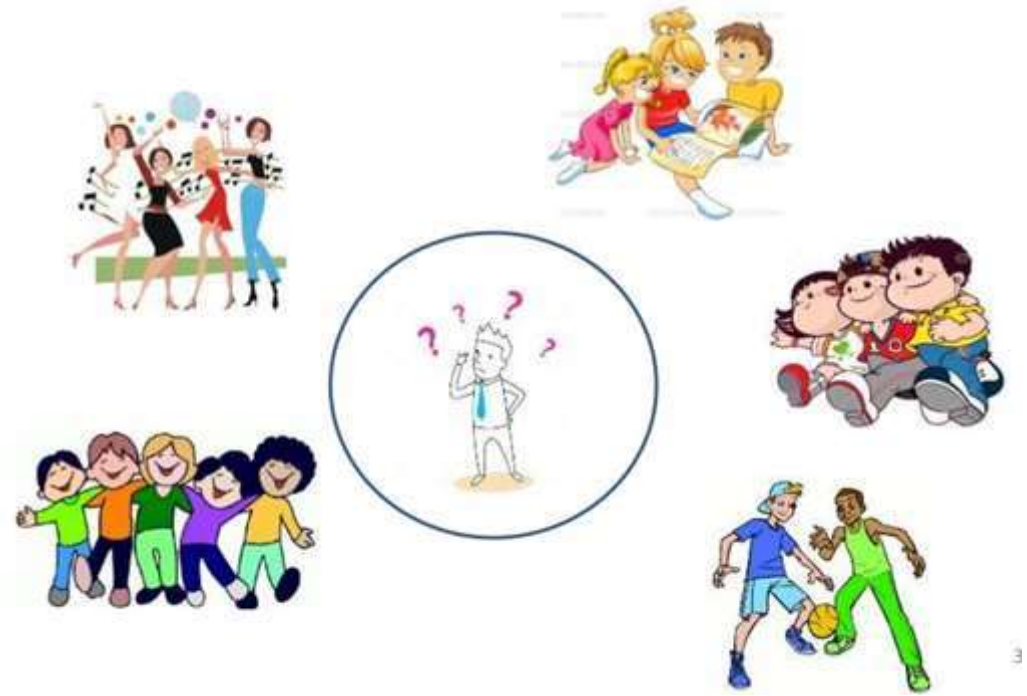
Algorithm: K Nearest Neighbor

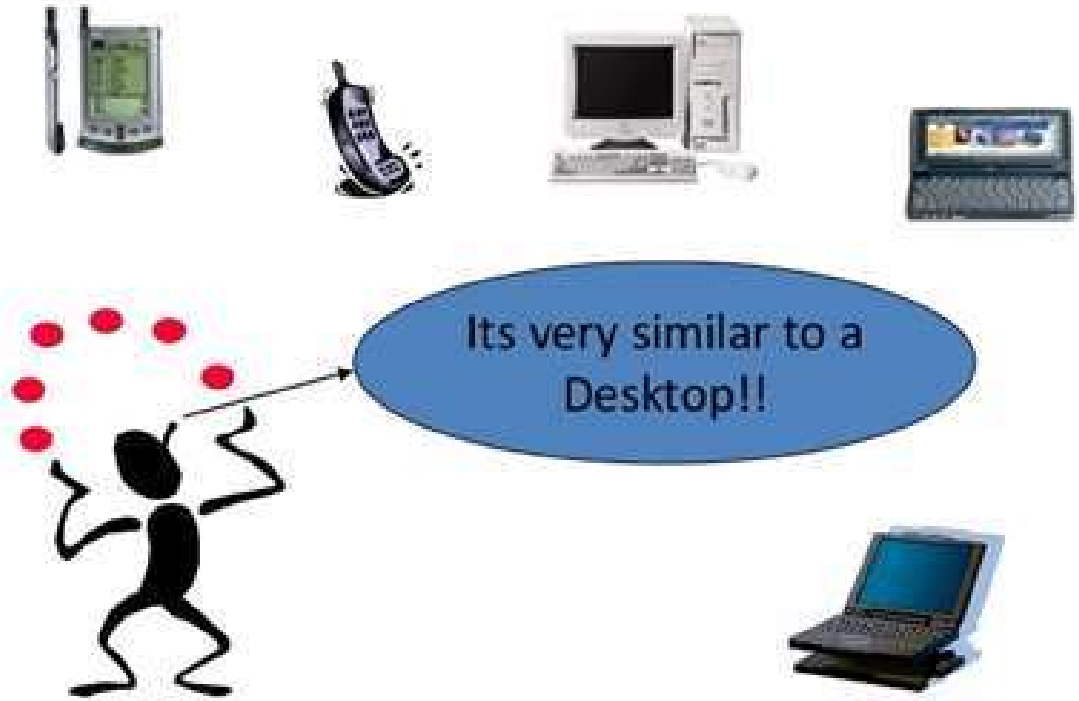
Simple Analogy

Tell me about your friends(who your neighbor are) and *I will tell you who you are*



How We Human Learn

Instance Based learning



What is KNN

- Classification Algorithm
- Distance Based or Similarity based algorithm
- It says check how many neighborhood having positive or negative values.

Distance Measurement

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Hamming Distance :

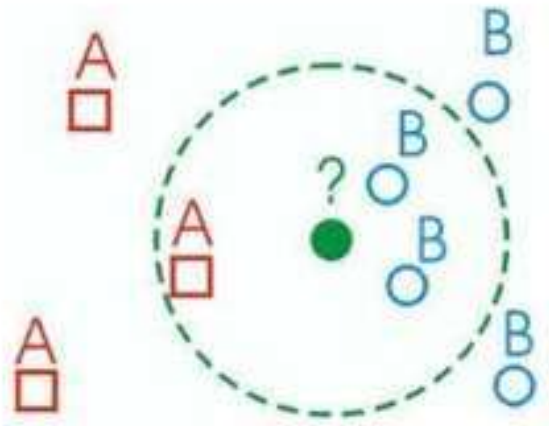
This distance is used to get the distance between categorical variable or strings. This distance is used for same length strings.

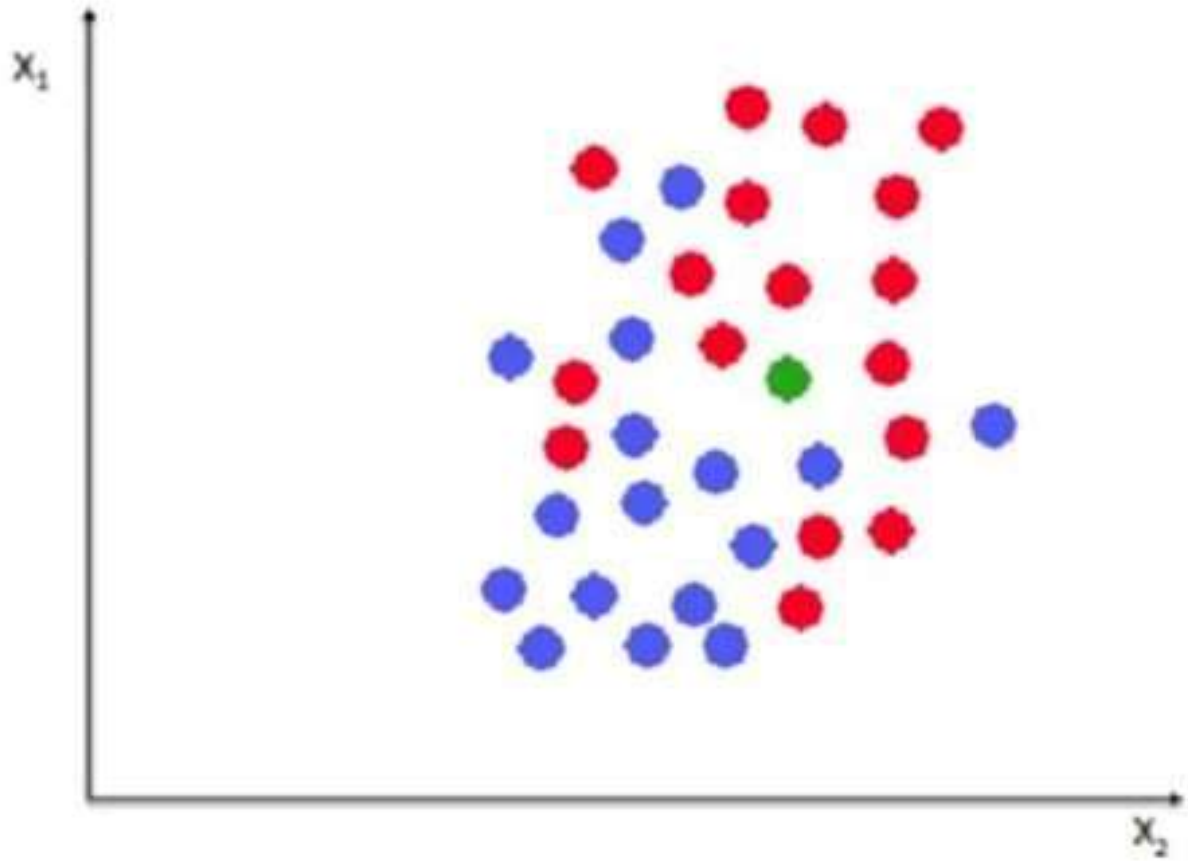
Cosine Distance :

If we have angle between two data points in 2D then we can use cosine similarity/distnace to get the relationship between them

Classification Approach

- Object(New instance) classified based on majority votes from its neighbor.
- The object is classified its most common class among its K Nearest neighbor(Measured by a distance function)



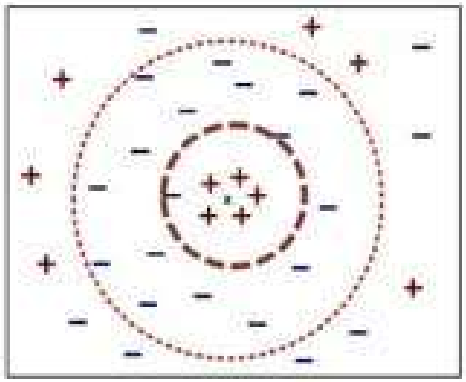

X_1
X_2

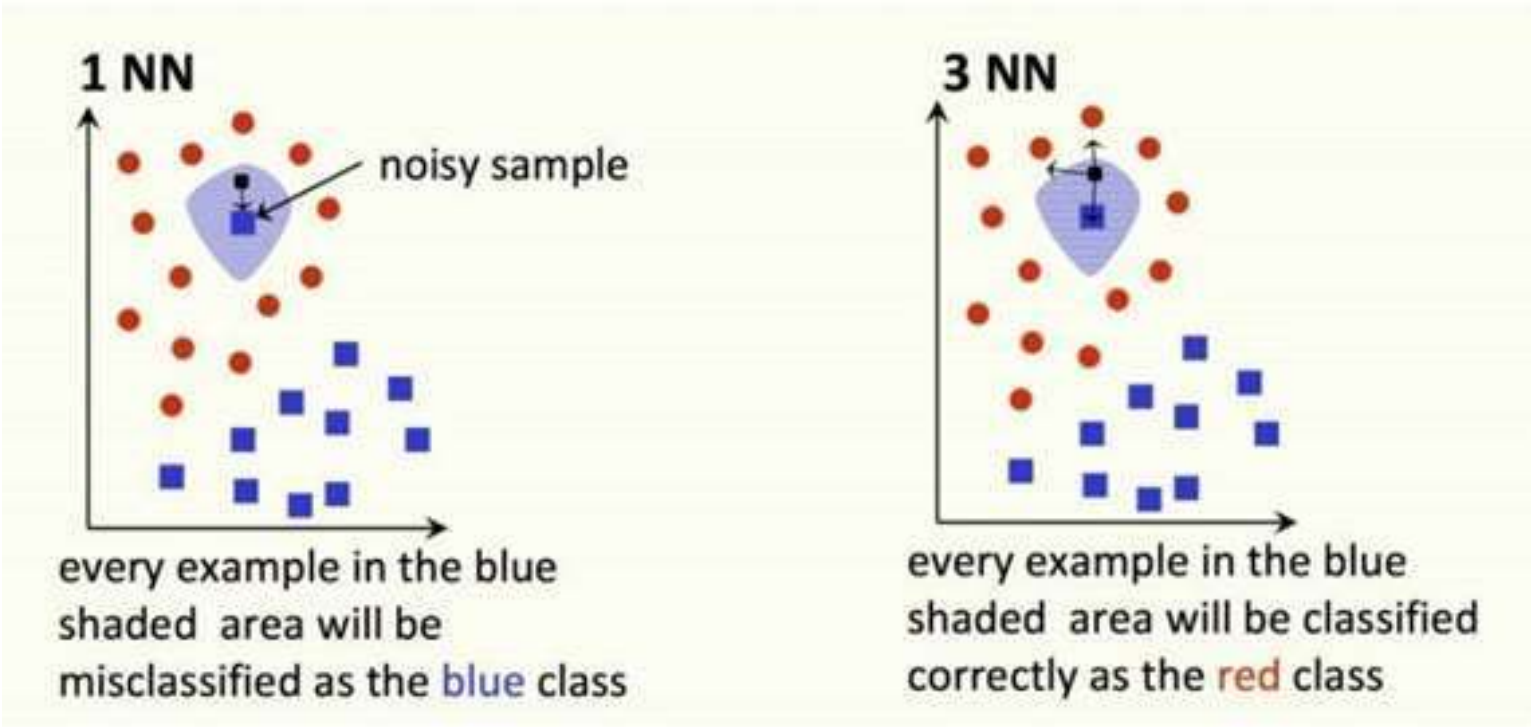
Example of Distance calculation

Customer	Age	Income	No. credit cards	Class	Distance from John
George	35	35K	3	No	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]=15.16$
Rachel	22	50K	2	Yes	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]=15$
Steve	63	200K	1	No	sqrt $[(63-37)^2+(200-50)^2+(1-2)^2]=152.23$
Tom	59	170K	1	No	sqrt $[(59-37)^2+(170-50)^2+(1-2)^2]=122$
Anne	25	40K	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]=15.74$
John	37	50K	2	YES	

How to choose Hyper parameter K

- Hyper parameter tuning : We will do hyper parameter tuning based on cross validation then decide what should be the perfect value of K.
- Small value of K can over fit the model.




1 NN
noisy sample
every example in the blue shaded area will be misclassified as the blue class
3 NN
every example in the blue shaded area will be classified correctly as the red class

Over fitting and under fitting

Decision Surface

- Its is a curve in 2D which separates the positive and negative points
- In 3D it will be a surface
- In nD it will be a hyper plane

Limitations of KNN

- Time complexity is very high
- Space complexity is also very high
- Cannot use in low latency projects

Fact : *In KNN we does not make model or train the data we simply use train data to estimate the class label of given test data*