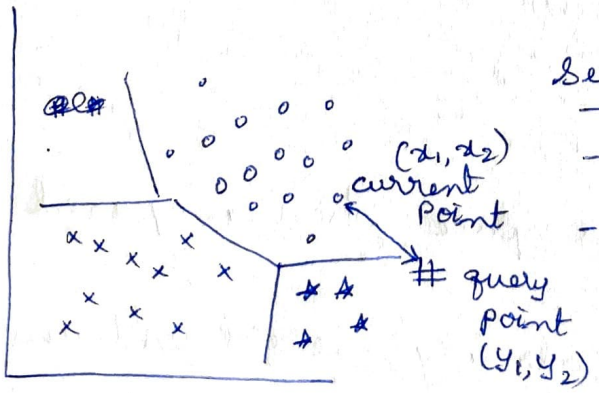
K-Nearest Neighbour Algorithm (K-NN)

K-Nearest Neighbours is a supervised learning approach that stores all available cases and classifies new ~~clas~~ cases based on a similarity measure (eg: distance functions)

Lazy Learning	Eager Learning
(instance-based learning)	
- simply stores training data and waits until test instance is provided	builds classification model once training data is provided
- Prediction involves identifying training instances similar to the test instances	Prediction involves application of classification model to test instance
more time in classification (Prediction)	more time in model building
Eg: K-nearest neighbour	Eg: Decision Trees



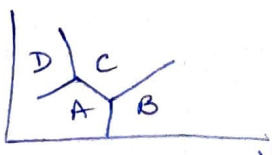
Selection of K

- K must be odd
- K must not be a multiple of number of classes.

$$\text{distance} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

(Euclidean distance)

or manhattan or minkowski's distance



Voronoi partition space

Algorithm

1, Load Data

2, Initialize 'K'

3, For each sample in the training data:

- calculate distance between query point and the current point.
- add the distance and the index of the example to an ordered list.

4, Sort the ordered list of distances and indices from small to large

5, Pick first K entries from the ordered list

6, Obtain the labels of the K entries

7. If regression, return mean of K labels
If classification, return mode of K labels

Note 1, selection of K parameter is crucial for the success of the solution
2, K is data dependent

Advantages

1, has richer hypothesis space when compared with eager learners.
(can handle data not already present in training data

2, simple to interpret

3, can be used for classification & regression

disadvantages

- prediction might be slow for large data
- sensitive to scale of data and irrelevant features
- requires high memory (needs to store all the training data)
- computationally expensive

Applications

- Network Intrusion Detection System
- Movie Recommendation System