MACHINE LEARNING
KNN

(K-Nearest Neighbors)

Trainer : Sujata Mohite
Email: sujata.mohite@sunbeaminfo.com

Overview

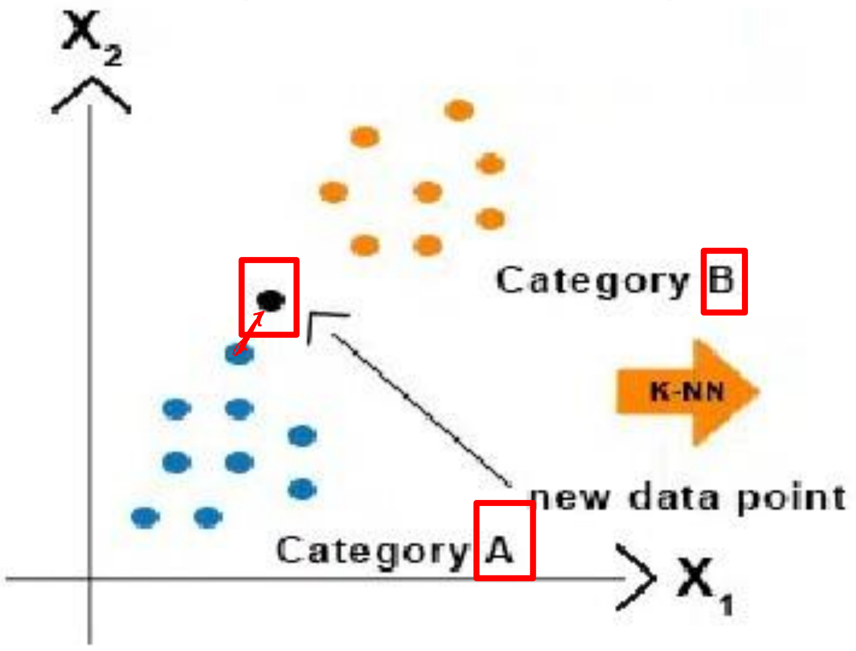
- The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems
- However, it is more widely used in classification problems in the industry
- It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection

nearest

- The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

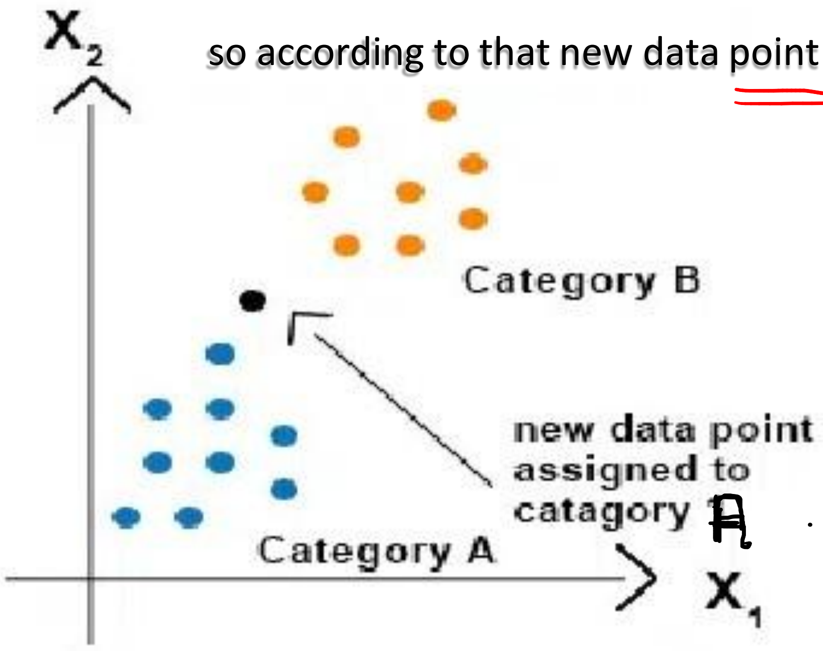
How does it work?

Initially there are two categories
New data point value is to be predicted



Distance from New data point to category A nearest value is calculated
Distance from New data point to category B nearest value is calculated

so according to that new data point assigned to category A



How does it work?

K-NN

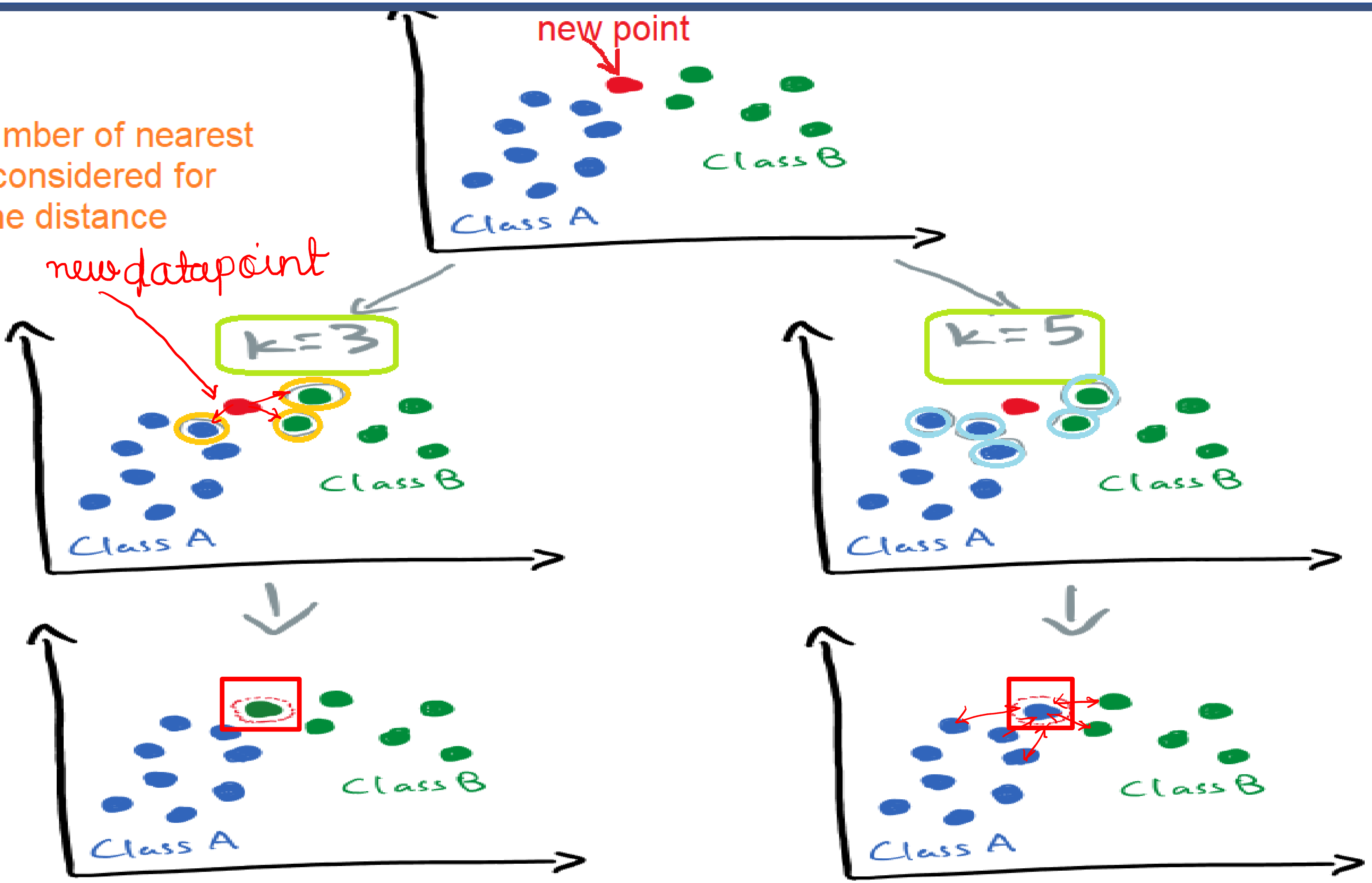
K decides number of nearest points to be considered for calculating the distance

K = 3

We have to find the three closest data points (three nearest neighbors) to the new (red) data point

K = 5

We have to find the five closest data points (five nearest neighbors) to the new (red) data point



How does it work ?

- A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function
- If K = 1, then the case is simply assigned to the class of its nearest neighbour.
- Note: all three distance measures are only valid for continuous variables.

Distance functions

Euclidean

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

Manhattan

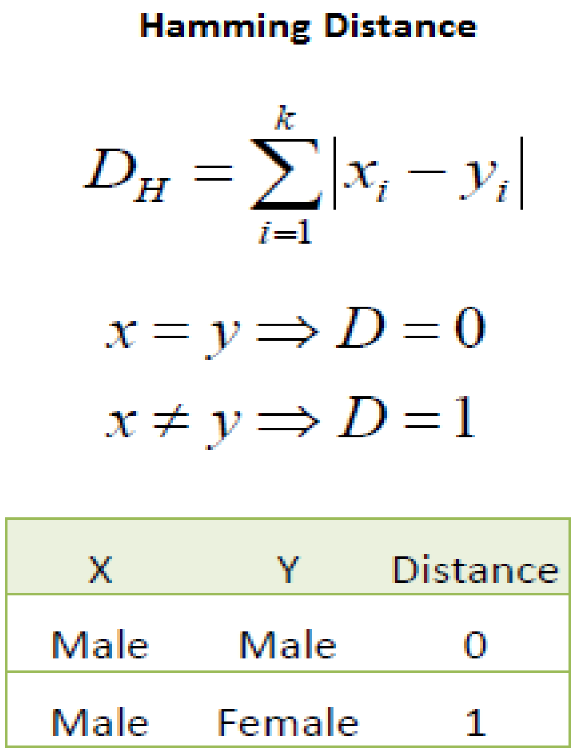
$$\sum_{i=1}^{k}\left|x_i - y_i\right|$$

Minkowski

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

How does it work?

- In the instance of categorical variables the Hamming distance must be used



Hamming Distance

$$D_H = \sum_{i=1}^{k} \left| x_i - y_i \right|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

How does it work ?

- Choosing the optimal value for K is best done by first inspecting the data
- In general, a large K value is more precise as it reduces the overall noise but there is no guarantee

by setting the hyper parameters

- Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value
- Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

range

Advantages

- No assumptions about data — useful, for example, for nonlinear data
- Simple algorithm — to explain and understand/interpret
- High accuracy (relatively) — it is pretty high but not competitive in comparison to better supervised learning models ← SVM
- Versatile — useful for classification or regression

① ②

Disadvantages

- Computationally expensive — because the algorithm stores all of the training data
- High memory requirement
- Stores all (or almost all) of the training data
- Prediction stage might be slow (with big N)
- Sensitive to irrelevant features and the scale of the data
- Difficult to choose K since there is no statistical way to determine that.
- ✓ Slow prediction for large datasets.
- ✓ Computationally expensive since it has to store all the training data

Applications of KNN

- Recommender system
- ✓ Relevant document classification
- OCR(Optical character recognition)

Thank You!!