K-Nearest Neighbour

Introduction

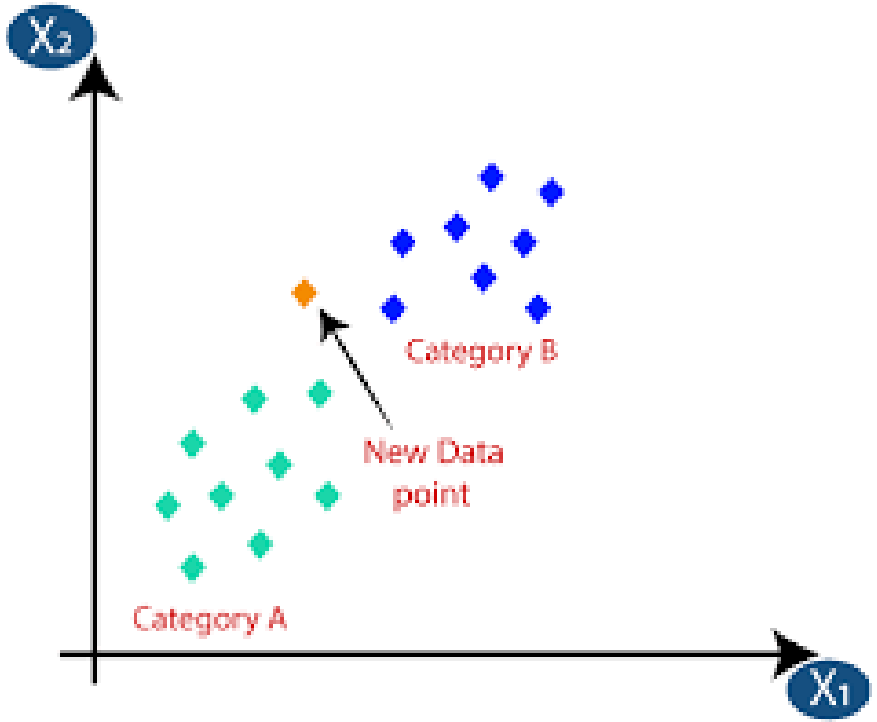
- K - Nearest neighbors is a lazy learning instance based classification(regression) algorithm which is widely implemented in both supervised and unsupervised learning techniques

.

- One of the top data mining algorithms used today.

KNN Classification Approach

- ***A new instance is classified by a** majority **of votes from its neighboring classes***
- The instance is assigned to the most common class amongst its K nearest neighbors



Principle of KNN Classifier

- It is lazy Learner as it doesn't learn from a discriminative function from training data but memorizes training dataset

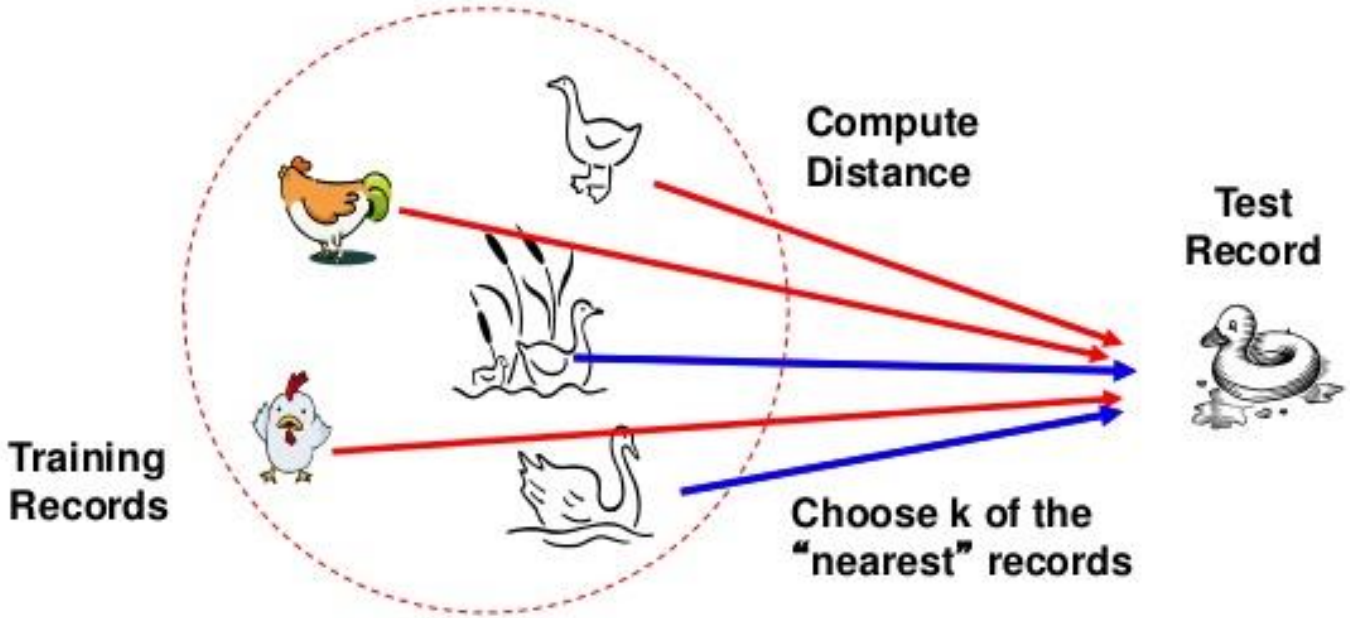
- This kind of technique implements classification by considering majority of vote among the “k” closest points to the unlabeled data point.

.

- It works on unseen data and will search through the training dataset for the k-most similar instances

- Euclidean distance / Hamming distance is used as metric for calculating the distance between points

Distance Measure



Types of Distance Measuring's

- . The Euclidean distance between two points in the plane with coordinates (x, y) and (a, b) is given by

$$dist((x, y), (a, b)) = \sqrt{((x - a)^2 + (y - b)^2)}$$

- Similarly we can use Manhattan distance,Minkowski,Hamming Distance.

Types of Distance Measuring's

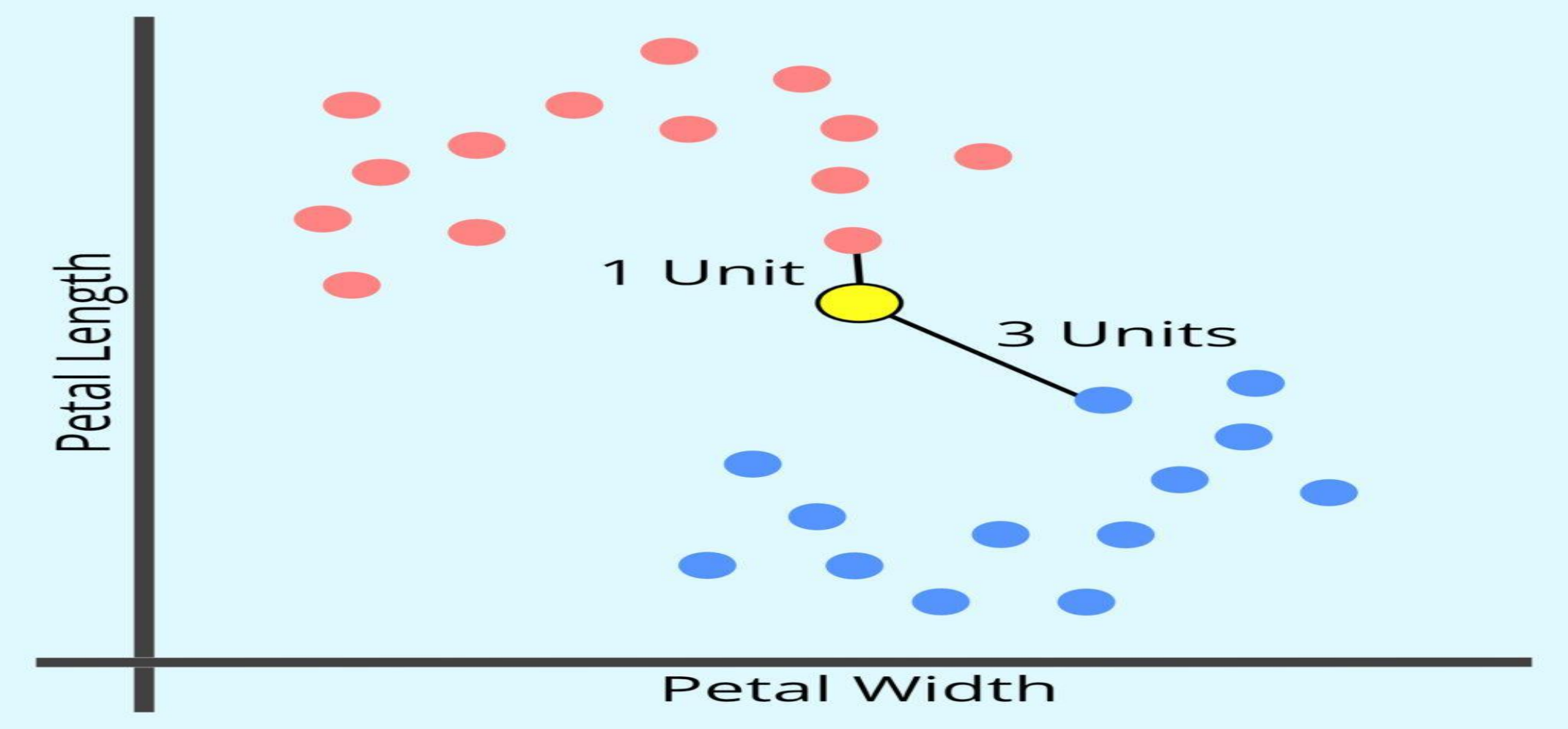
- . The Euclidean distance between two points in the plane with coordinates (x, y) and (a, b) is given by

$$dist((x, y), (a, b)) = \sqrt{((x - a)^2 + (y - b)^2)}$$

KNN Algorithm

- All the instances correspond to points in an n-dimensional feature space.
- Each instance is represented with a set of numerical attributes.
- Each of the training data consists of a set of vectors and a class label associated with each vector.
- Classification is done by comparing feature vectors of different K nearest points

- Select the K-nearest examples to E in the training set
- Assign E to the most common class among its K-nearest neighbors.



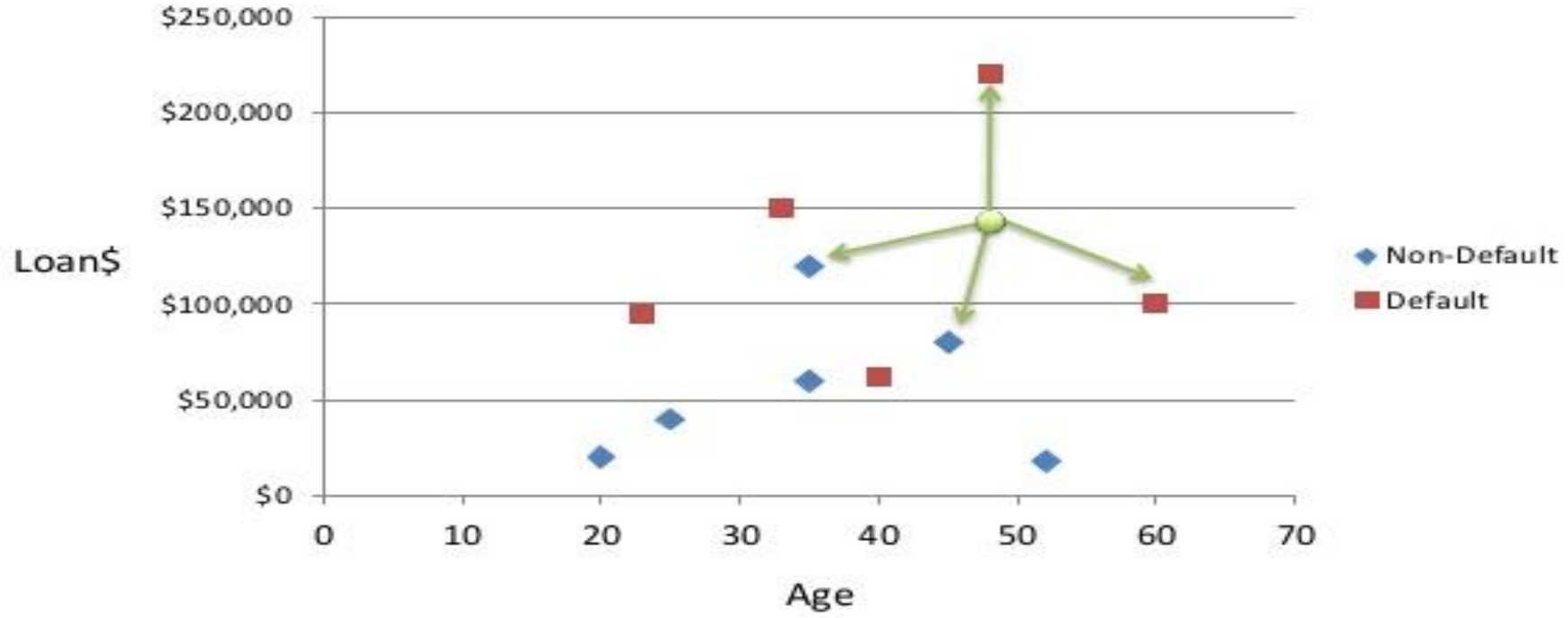
How to Choose K?

- If K is too small it is sensitive to noise points.
- Larger K works well. But too large K may include majority points from other classes.
- Rule of thumb is K < sqrt(n), n is number of examples

KNN Feature Weighting

- Scale each feature by its importance for classification
- Can use our prior knowledge about which features are more important
- Can learn the weights wk using cross-validation

KNN Classification



Applications

- Recommender Systems
- Medicine
- Finance
- Text mining
- Agriculture

Pros

- Non complex and Very easy to understand and implement
- Useful for non linear data as No assumptions about data.
- High accuracy (relatively), but not competitive compared to Supervised learning algorithms.

- Can be used both for classification or regression

- Best used where where the probability distribution is unknown

Cons

- Computationally expensive
- Lot of space is consumed as all the data points are stored
- Sensitive to irrelevant features and the scale of the data
- Output purely depends on K value chosen by user which can reduce accuracy for some values.

Conclusion

- K-Nearest neighbor classification is a general technique to learn classification based on instance and do not have to develop an abstract model from the training data set. However the classification process could be very expensive because it needs to compute the similarly values individually between the test and training examples. K-nearest neighbor classifier also suffers from the scaling issue. It computes the proximity among the test example and training examples to perform classification. If the attributes have different scales, the proximity distance might be dominated by one of the attributes, which is not good.