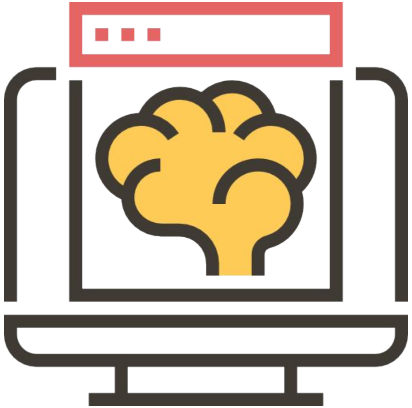
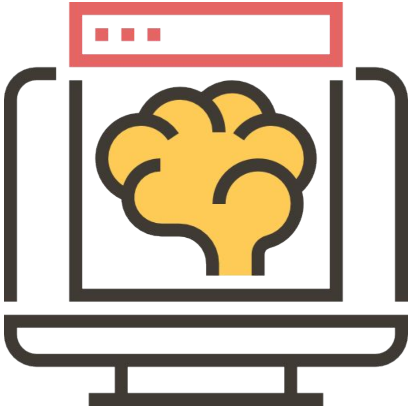


Supervised Learning Classification



K Nearest Neighbour

Agenda

- Introduction to K Nearest Neighbour
- Uses and applications of KNN
- KNN Working
- Optimum value of Factor K in KNN
- Pros and Cons of KNN

Introduction to KNN

- KNN stands for K-Nearest Neighbors
- KNN is a model that classifies data points based on the points that are most similar to it.
- The model representation for KNN is the entire training dataset.
- KNN is an algorithm that is considered both non-parametric and an lazy learning.
- KNN belong to Supervised learning method.

Uses and applications of KNN

- KNN is often used in simple recommendation systems, image recognition technology, and decision-making models.
- KNN can be used for both classification and regression.

KNN Working

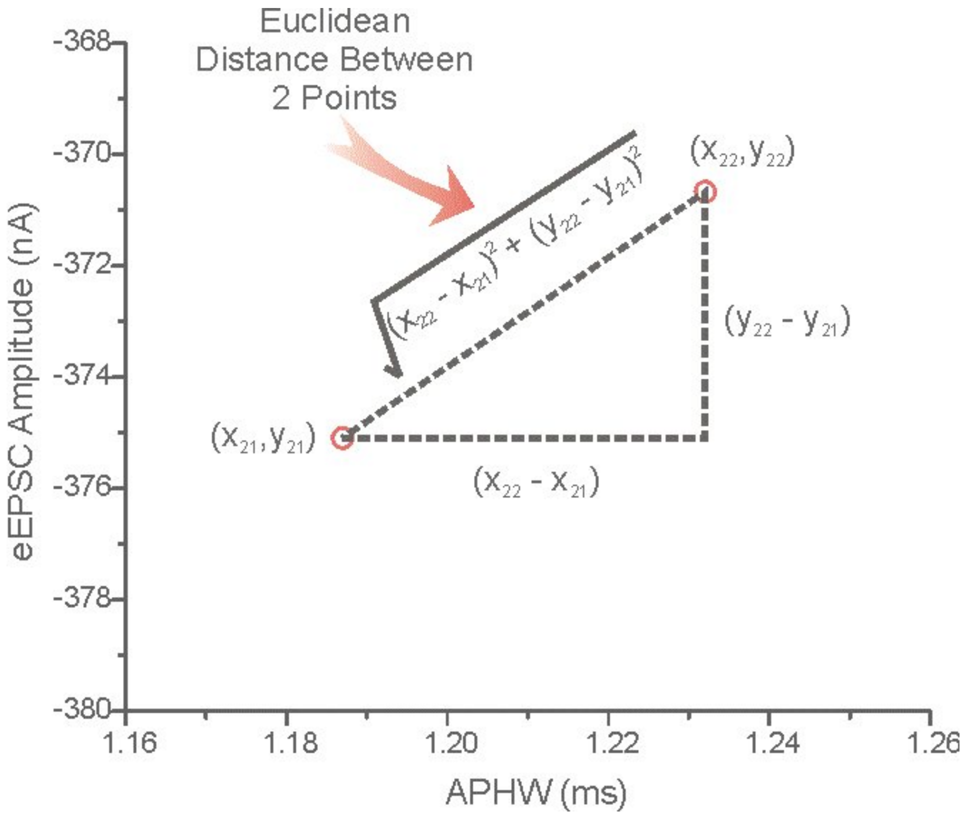
- **For Classification :** If you are using **K** and you have an **even** number of classes (e.g. 2) It is a good idea to choose a K value with an **odd** number to avoid a tie. And the inverse, use an even number for **K** when you have an odd number of classes.
- **For Regression :** When KNN is used for regression problems the prediction is based on the mean of the K-most similar instances.

KNN Working

- KNN makes predictions using the training dataset directly.
- For regression this might be the mean output variable, in classification this might be the mode (or most common) class value.
- Distance Measure used.
- For real-values input variables, the most popular distance measure in **‘Euclidean Distance’**.
- Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

  Euclidean Distance(x, xi) = sqrt(sum((xj – xij)^2))

KNN Working - Euclidean

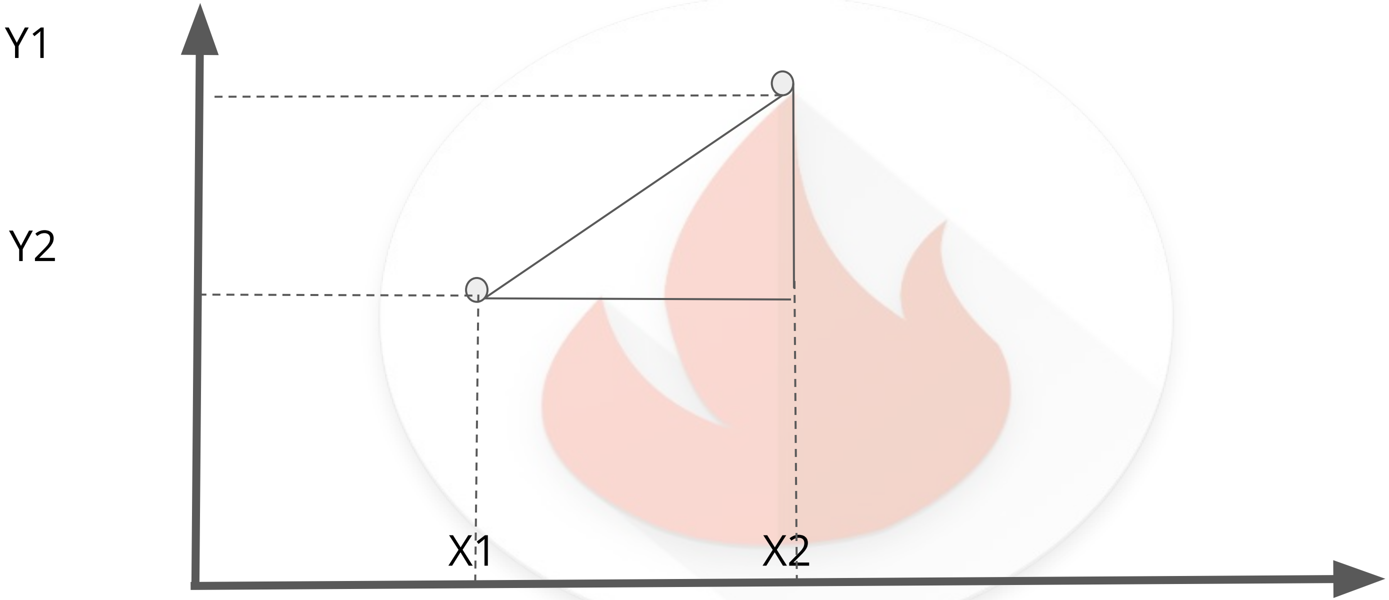


KNN Working

- Other popular distance measure is :
- **Manhattan Distance:** Calculate the distance between real vectors using the sum of their **absolute** difference. Also called as Block Distance. It is replace by a new metric in which the distance between two points is the sum of absolute difference.
- There are many other distance measure, such as Tanimoto, Jaccard, Mahalanobis, and cosine distance.

KNN Working - Manhattan

Manhattan Distance :- |X1-X2| + |Y1-Y2|



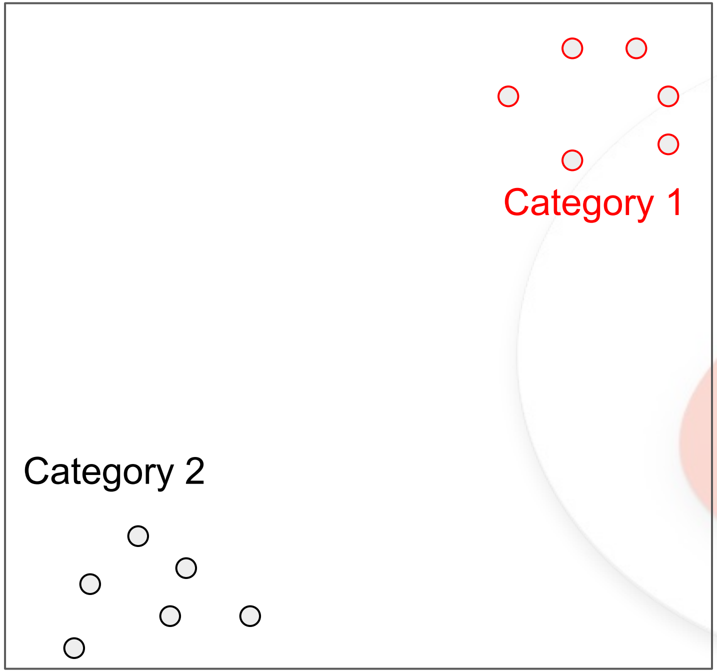
KNN Working

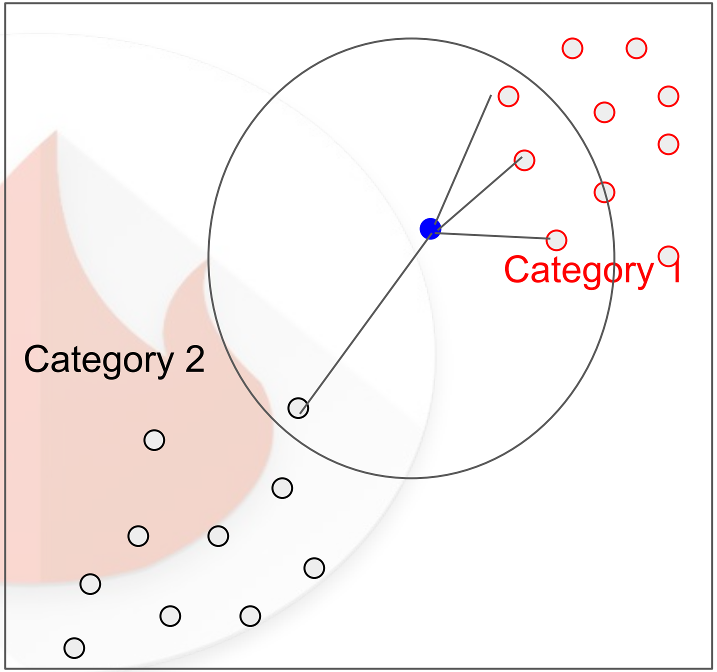
- Euclidean is a good distance measures to use if the input variables are similar in types. i.e. all measured widths and heights.
- Manhattan distance is a good measure to use if i/p variables are not similar in types i.e. age, gender, height etc.
- K values from 1 to 21.
- KNN increase when have a large size of training data.

KNN Working

- K-NN algorithm uses ‘feature similarity’ to predict the value of new data point.
- Step 1 : Load all dataset and assign the value of K. i.e. K = 3,5,7,9……
- Step 2 : For each point in the test data do the following.
    - Calculate the distance between points.
    - Now, based on the distance value, sort them in ascending order.
    - Next, it will choose the top K from sorted array.
    - Now, it will assign a class to the test point based on most frequent class.
- Step 3 : End

KNN Working





Pros and Cons of KNN

- Pros
  - Learning and implementation is extremely simple and Intuitive.
  - Flexible decision boundaries
- Cons
  - Irrelevant or correlated features have high impact and must be eliminated.
  - Typically difficult to handle high dimensionality
  - Computational costs: memory and classification time computation

Thank you