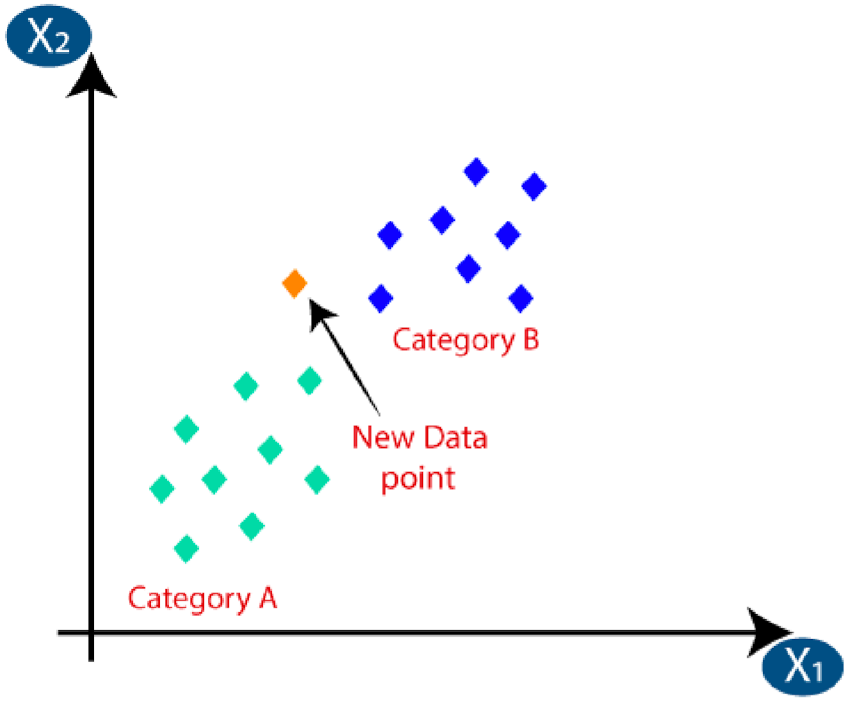
Day 14: K-Nearest Neighbors



Author: Alisha Metkari

KNN

- K-nearest neighbors is a supervised machine learning algorithm.
- It considers the similarity between data points to decide the output. First K nearest points are chosen for the decision of the output.
- The value of K is defined experimentally which best suits the given data and reduces the error.
- To calculate the similarity between data points, it uses distance functions. There are many distance functions and most used is euclidean distance.
- The distance function should be chosen depending on the size and dimensions of the data.
- It is a lazy learner algorithm. It stores all data in training phase and performs computation when new data point is provided for prediction.
- It does not make any assumptions about the underlying data.

Author: Alisha Metkari

Working of KNN

- First, we calculate the distance of new point from each point.
- Then, we choose first K nearest neighbors as per the distance calculated.
- For classification problem, we calculate count of points in each class from those K nearest neighbors and assigns the class with maximum count to the new point.
- For regression problem, the average or weighted average of the values of K neighbors is assigned as the predicted value to the new point.

Author: Alisha Metkari

Distance functions

- The most used distance functions in machine learning are as below.
- Euclidean distance: It is the length of the line segment between two points in euclidean space.

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2]}$$

- Manhattan distance: It is sum of absolute difference between coordinates of points.

$$d\left(x, y\right) = \sum_{i=1}^{n} \left|x_i - y_i\right|$$

- Minkowski distance: Euclidean and Manhattan are the special cases of Minkowski distance and it is given as below.

$$d\left(x, y\right) = \left(\sum_{i=1}^{n} \left(x_i - y_i\right)^p\right)^{\frac{1}{p}}$$

Author: Alisha Metkari

End Notes

- KNN is distance based algorithm hence normalisation of data is required to avoid dominance by high magnitude features.
- Choosing optimal value of K is very crucial. If data has more outliers then higher value of K is recommended. Choose the value of K which minimises the error.
- As the number of features increases the KNN becomes computationally expensive because the calculation of distance becomes expensive.

Author: Alisha Metkari