Nearest Neighbor Estimator

kNN for Regression and Classification

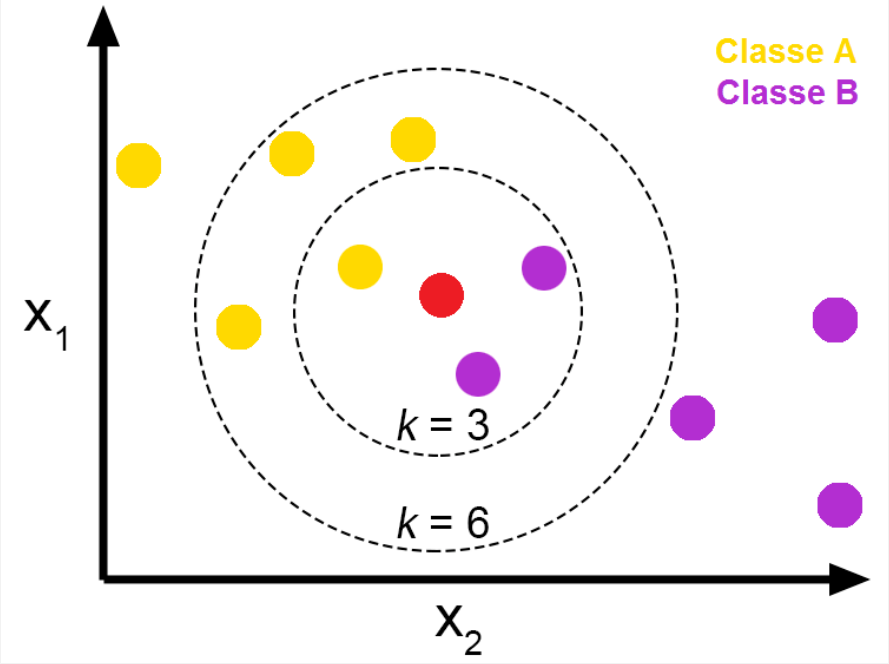
by Dr. Tumpa Banerjee (Dept of MCA)

kNN Estimator

- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Nearest Neighbor Estimator

- a non-parametric estimator is the K nearest neighbor (KNN) estimator.
- This simply “looks at” the k points in the training set that are nearest to the test input x.
- Decision is taken based on the labels of its nearest points.
- counts how many members of each class are in this set, and returns that empirical fraction as the estimate



Nearest Neighbor Estimator

- Let $Tr = (x_i, y_i), i = 1,2,3, \ldots, N$ bet the training set with data points x_i and the price y_i
- Let x be the new observation whose y value has to be predicted.
- Calculate Euclidean distance of x with the data points in the training set. Arrange the data points in ascending order.
- Select the first k members from the above set as the k-nearest neighbors $N(x) = x_1, x_2, \ldots, x_N$ of x. The y value corresponding to x can be estimated from the neighbors.

K nearest Neighbor Classifiers

- kNN algorithm classifies the new instance as follows:
- $p(y|D, x, k) = \frac{1}{k}\sum_{i \in N_k(x,D)} \mathbb{I}(y_i = c)$
- Where $\mathrm{N_k(x, D)}$ denotes k nearest neighbors of the point x in the dataset D
- $\mathbb{I}$ is the indicator variable.
- $\mathbb{I}(e) = \begin{cases} 1, if\ e\ is\ true \\ 0,\ if\ e\ is\ false \end{cases}$

K-nearest Neighbor Regression

- kNN algorithm classifies the new instance as follows:
- $p(y|D, x, k) = \frac{1}{k} \sum_{i \in N_k(x,D)} y_i$

Weighted kNN

- Prioritize the features of the data based on the importance of individual variables.
- No specific heuristic exists for determining the value of k.
- A small value of k may miss some significant neighbors.
- Large value of k may include unwanted observations in the neighbor list that may lead difficulty.