K-NN Machine Learning Alogorithm

Mannem Charan AI21BTECH11019

Abstract

This report consists of my basic understanding of one of the popular Ml methods ”K-NN”.

1 K-NN

K-NN,K-Nearest Neighbour algorithm is one of the simplest machine learning technique and it comes under supervised learning.It will not assume any relationship between dependent and independent variables so it is a non-parametric machine learning algorithm.It is also known as ”Lazy-learning” algorithm beause it doesn’t have any training phase, it will just store the training data and uses it at the time of classification.

2 Understanding K-NN

The K-NN algorithm **assumes** similarity between the new data and the available data and then it will put the new data by comparing similarity between the available categories.The name K-Nearest Neighbours signifies that we try to compare the new data point with ”k” nearest neighbours. And we will define neighbour by measuring the euclidian distance between the data points.So K-NN uses every input instances of the data to classify the new data. In case of Regression we try to take mean or median of labels of the ”k” nearest neighbours.So our job is to store the complete training data and take a new data point find the ”k” nearest neighbours(by euclidian distance method) and if it is classifcation take the category/label of majority neighbours or else take the mean/median of continuos values of ”k” nearest neighbours.So the important thing here is, choosing the value of ”k”.

3 Choosing the k

As you know that, ”k” here represent the number closest data points we are taking for comparing the similarities with the new data point.So choosing ”k” affects the performance of the model, we can study the changes in model with ”k” using error curves shown in Fig 3.1. The red line reperesnts the test error and blue line represents the train error.So for $K = 1$,the training error is 0 this can be understood since for every data point the nearest point is the point itself.But the test error is high for $k = 1$ due to the noise in the data which leads to overfitting.In general, as k increases the the test error will decrease with increase in train error.So there is no exact approach to choose k some data scientists prefer to choose a odd number if there are two classes and another simple way is take $k = \sqrt{N}$ where N are the total data samples in training dataset.

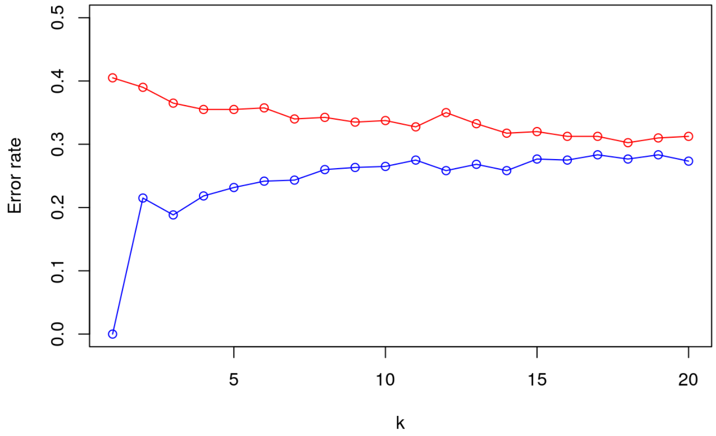

Fig. 3.1: error rates w.r.t *k*

4 Advantages of K-NN

1) It is easy to understand and implement
2) It is not sensitive to the outliers(by choosing correct value of *k*)
3) It is more efficient for large datasets.
4) It takes no time to train for obvious reasons.

5 Disadvantages of K-NN

1) K-NN gives equal importance to all features, as it is only takes distance between two data points.
2) Uses lot of computational power.

6 Data Preparation for K-NN

1) **Feature Scaling:**Proper feature scaling is required, since we are taking distance between the points as the measure to say closeness,it would be better if all the features are in the same scale.
2) **Dimensionality Reduction :** If there are more classes to classify, then it is not a good judgement to take class of majority neighbours.So using dimensionality reducion techniques like feature selection helps k-NN to work more accurately.
3) **Missing Features :** We can’t use k-NN if we don’t have the data of all features for all data samples(as we can’t find the norm),so either we have to delete the samples not having all features,or we have to predict the missing features(imputation).

7 Questions

1) Why k-NN is known as lazy-learning algorithm?
2) What is ”k” in k-NN algorithm?
3) How the model changes with the ”k” value?
4) What is the measure we are using in k-NN to find the closeness between the data points?
5) Mention the data preprocessing steps used in k-NN?

8 Solutions

1) Because in k-NN we only storing the training data there is no training phase.
2) The ”k” represents the the number of closest neighbours we are taking in the algorithm.
3) For the lower values of ”k”, there is a high chance of overfitting and as we increase k the generalization error decreases.
4) In general we use euclidean distance as measure to find the closeness between the data points.
5) The important data pre-processing steps that must be done in k-NN are,
   - Feature Selection
   - Feature Scaling
   - Imputation/deleting the missing features.