CSSE 519: Data Science Fundamentals
Question 6: Report
K-Nearest Neighbor model

Introduction:

K Nearest Neighbor algorithm can be used for classification and regression problems. Classification is classifying a new object to be part of one of type in a known domain. Regression is predicting a value of target variable based on a set of independent variables.

KNN regression usually calculates the average of the target variable of K nearest neighbors. It can also work by calculating the inverse distance weighted average of the K nearest neighbors. It can be Manhattan or Euclidean distance for continuous and Hamming distance for categorical values. The input is the K nearest features in the feature space and the output is the average of those values for KNN regression. In case of KNN classification, majority vote of its K nearest neighbors is taken as the result.

KNN is an also called as Lazy learning algorithm which is non-parametric. This is useful since most of the real world problems do not follow the usual theoretical assumptions since KNN model doesn't make any assumptions on the data distribution. KNN doesn't have a separate training and testing dataset. The entire dataset is used for Training which is better.

Working mechanism:

Choosing the value of K will have effect on the classifier. The best K corresponds to the lowest test error rate. The K value can be selected via many heuristics. Large value of K will reduce the effect of noise on making prediction but blur the boundaries between classes. Larger the values of K, smoother the boundary between the classes. Small values of K can lead to high variance. It is obvious that increasing K value can increase the accuracy but the computational cost will also increase with that.

Like any other algorithms, KNN algorithm can be affected by the presence of noise or irrelevant data. So it is better to remove the less relevant data and keep only those features which will contribute towards making prediction. This can be achieved by removing columns which contain more Nan value or Null value or redundant columns can be removed. This will bring down the size of the dataset as well. This is also called as Dimension reduction.

Processing the raw data to get a definite set of features is called feature extraction. Feature extraction is usually done before applying KNN algorithm. Both these steps can be combined as part of principal component analysis (PCA), linear discriminate analysis (LDA), or canonical correlation analysis (CCA).

In k-fold cross validation, the training set is divided into k groups of almost equal size. Initially one group is used for training or fitting and the remaining k−1 groups are used for testing or predicting. This process is repeated k times with different group as training set each time and observations are recorded each time and then the average is taken as the final result.

The problem at hand is a Regression problem. The sale price of houses has to be predicted with the set of features at hand. Features include information of characteristics of the houses such as room count, square feet, etc. Any suitable Regression model can be used for this such as KNN model, Random Forest Regression, Decision Tree Regression. The given dataset was first cleaned up. Almost all the columns

had Nan values and Null values. So they must be first taken care before fitting into the model. For columns with continuous data, mean of each column is found and substitute in the place of Nan values which is then used for modeling.

The entire dataset can be divided into 2 parts for testing and training. The training data is made to be fit on the model and the target variable is predicted with the testing dataset. A confusion matrix is often used as metric to validate the accuracy of KNN model. Other metrics include Mean squared error, Mean absolute error and Root Mean Square Error.

Advantages of KNN model:

- Can be used on both Continuous and Categorical data.
- Simple to implement.
- Flexibility in choosing the value of K.
- One of the fastest algorithms.
- Efficient with large training dataset.
- Performs easily with multiclass data sets as well.

Disadvantages of KNN model:

- Choosing K value can alter the results.
- Usage of different distance metric can create deviation in the result.
- Computation cost is high since distance has to be measured for each query with each of the entities in the feature space.
- Even though it has a fast testing phase, the training phase can be longer with k-fold cross validation.

Accuracy Evaluation:

KNN model performance was close to that of Random Forest model. The evaluation results of KNN model are listed below.

Accuracy score:	99.671041589741876
Mean Squared Error:	0.026391671751130233
Root Mean Squared Error:	0.162455137657
Mean Absolute Error:	0.071259356608615765
Score from Kaggle:	0.0816
Rank in Kaggle:	2227

Applications of KNN:

It is highly versatile and can work well in variety of applications. Its application includes Content retrieval based on Nearest Neighbors, Gene Expression, 3D structure prediction, Identifying and classifying an item in the picture according to its variety or type, Inter-protein interaction prediction. It is also used widely in many cases of Computer Vision to identify and classify the object viewed.

Conclusion:

KNN algorithm is one of the simplest algorithms. It can also produce highly competitive results. Working of KNN algorithm, how it is applied in classification and regression setting, its pros and cons were discussed. Out of decision tree regression, Logistic regression, Random forest regression, KNN regression performed better.