Asg1_Q5

August 15, 2020

1. Explain why you had to split the dataset into train and test sets?

The dataset we use training is to fit the model. The test set is used to give an evaluation of the final model which fits good on the training sets. The process is to use the training set to find a suitable model, and then use the testing set to give the accuracy of this model on datasets.

2. Explain why when finding the best parameters for KNN you didn't evaluate directly on the test set and had to use a validation test.

If the best parameters is evaluated from the test set, the KNN model with that specific set of parameters might only work well for that specific train set and the test set. Then, we have no evidence about whether the model can work well on other datasets. If the model is evaluated on the validation set, we can tune the parameters based on the feedbacks from the validation set. As the model never sees the data in test set, we can evaluate the model on test set to get an evidence about its generalization on other datasets.

3. What was the effect of changing k for KNN. Was the accuracy always affected the same way with an increase of k? why do you think this happened?

By changing the k value for the KNN classifier, the clustering result and the accuracy is changed. As the value of k increased, the accuracy might not behave in the same way (the accuracy may reduce when k increase). As KNN is an unsupervised learning method, it discovers the natural groups of the datasets. If the value of k is small, data points in different classes may be involved in the same groups; however, if the value of k is large, the data in the same group would be separated to different groups.

4. What was the relative effect of changing the max depths for decision tree, random forests, and gradient tree boosting? Explain the reason for this.

With the increasing of tree depths, the accuracy will increase to a critical value, this will be the max accuracy. Then if tree depths increase, the accuracy will go down and it might happen overfitting on decision tree, random tree and gradient tree boosting.

5. What was the relative effect of changing the number of tree depths for random forests, and gradient boosting? Explain the reason for this.

Increasing the number of trees for random forests and gradient boosting tree will reduce the error (error means bias and variance). The error will converge to a particular value no matter how increasing the number of trees.

6. What does the parameter C define in the SVM classifier? What effect did you observe and why do you think this happened?

The parameter C is regularisation parameter, which controls the trade-off between achieving a low error on the training data and minimising the norm of the weights. High C implies we are allowing fewer outliers. The optimization will choose a smaller-margin hyperplane if that hyperplane has a better result on getting training points classified correctly. Conversely, low C implies we are allowing more outliers.