IS580 Knowledge, Discovery and Mining – Assignment 3

1.2.3.

All these parts are completed in the R code. All the related code is provided in the “2093656_Assignment3.R” file.

4.

I. *kNN Models*

In the kNN models, k is a critical parameter. In a general manner, the optimum k value depends on the data characteristic. But we can easily say that,

- Smaller k values will make the system too sensitive to the noise. Also, smaller k values are more likely to result in overfitting.

- On the other hand, larger k values reduce the effect of the noise in classification process as our system is not that sensitive to the noise. Also, large k values make boundaries between classes less distinct.

II. *SVM Models*

In SVM models, there are two main purposes to achieve,

- Getting a hyperplane with largest possible minimum margin.
- Making hyperplane such that it separates the instances as correct as possible.

When the cost value is decreased, the first purpose is achieved, i.e. a larger margin is achieved whereas the second purpose cannot be satisfied perfectly. There is a sharp tradeoff between them.

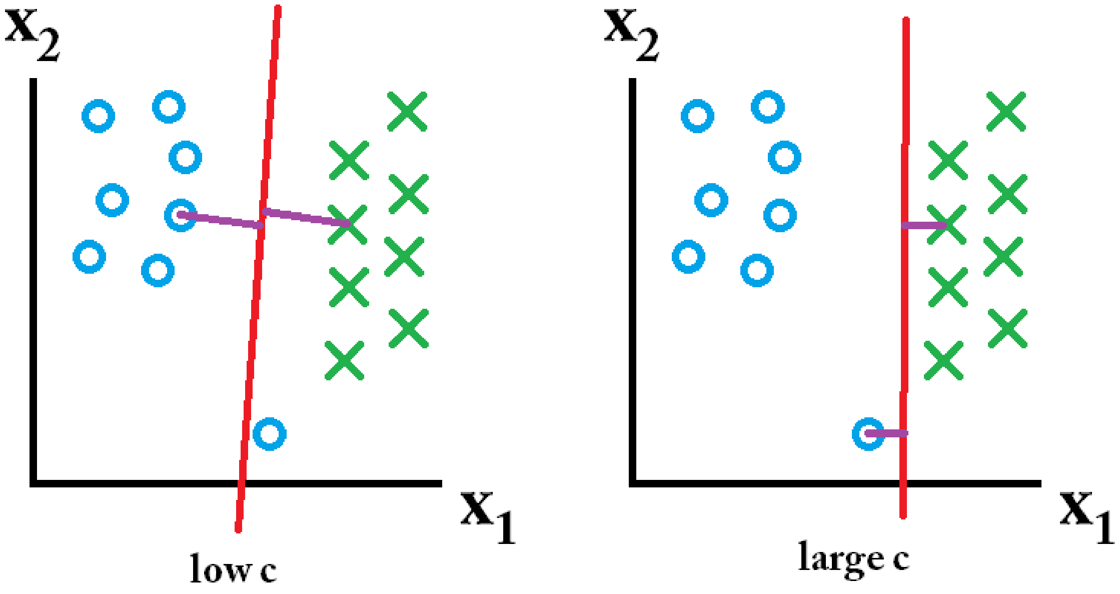

Figure 1. Effect of Cost Parameter on SVM.

As it can be seen from Figure 1, when c is small, the minimum margin is high but there are some mistaken placed blue instances. So that the hyperplane is not located perfectly.

III. *Random Forest Models*

In RandomForest models, number of trees effects the accuracy of the system. If it is set to a very small number, then there might not be enough number of prediction for each input. But after some point, increasing the nTree will not have a significant change in the accuracy.

5.

I. kNN Models

To compare the results of each model, I investigated Confusion Matrix of each model and calculated their accuracy values.

For kNN models, the accuracy values were like given in Table 1 below.

k Value	Accuracy
1	0.951
5	0.902
10	0.837
100	0.677

Table 1. Accuracy values of kNN Models with Different k Values

The best accuracy value seems to be when k=1. But as explained in Part 4, this result might be fallacious because of the overfitting issue. k=5 model gives a reasonably good result but after that point the accuracy value decreases.

So, among these models k=5 kNN model can be a good choice.

II. SVM Models

The accuracy values of SVM models with three different cost values can be seen in Table 2 below.

c Value	Accuracy
0.05	0.694
1	0.85
5	0.918

Table 2. Accuracy values of SVM Models with Different c Values

For SVM models, c=0.05 case gives the lowest accuracy result. In this case, c is pretty small and we can get a hyperplane with largest possible minimum margin. But this hyperplane is not an optimal one and cannot separate instances successfully. In c=1 and c=5 cases, we are getting better and better accuracy levels as our hyperplane can separate data in a better way.

Among three SVM models, c=5 model is the best option to apply classification task.

III. Random Forest Models

For random forest models with nTree=100 and nTree=1000 parameters, the achieved accuracy values can be seen in Table 3 below.

nTree Parameter	Accuracy
100	0.964
1000	0.964

Table 3. Accuracy values of Random Forest Models with Different nTree Values

Random Forest models give much better accuracy results with respect to the other classification methods discussed in previous parts.

But it is noticeable that when nTree is increased from 100 to 1000, the accuracy result does not change at all, but spends more memory. This occurred because of the size of the dataset. At nTree=100, a sufficient number of prediction is made for all input so increasing it to 1000 does not change anything on the accuracy value. This increase might be useful on another large dataset.

Between these two models, Random Forest with nTree=100 is the best model to apply classification task on this dataset.

6.

For all 9 models, the Confusion Matrix was created and investigated. And the related accuracy values were given in the tables in Part 5.

For kNN models, the one with k=1 is leading to overfitting so, it is unpractical.
k=10 and k=100 have low accuracy performances.

For SVM models, the one with c=0.05 cannot have a hyperplane that splits data successfully.

For Random Forest Models, both nTree=100 and nTree=1000 models perform a good classification.

7.

When the performance of all these models are investigated, the most successful one is Random Forest Model. As the performance does not change with nTree=100 and nTree=1000 and creating 1000 tree will hold more memory, it is logical to choose the Random Forest model with nTree=100 to have the best classification performance.