CE802 Machine Learning and Data Mining

Assignment.Part 2

Tutor: Dr. Luca Citi

Student:Lyu Yaowei

Student number: 1802697

Words:1017

1. Background

According to the part 1, we adapt the SVM, K Nearest Neighbor (K-NN) and the Decision Tree as the learning procedure to implement the prediction. In this part, I am going to use the scikit-learing module in python to solve this task. At the beginning, I import the scikit-learn package which include the SVM, Decision Tree and the K-NN. I am going to explain the reason and performance of each learning procedure at the last report.

2. Learning Procedure

2.1 Decision Tree

First I want to talk about the Decision Tree. We collect the data from the previous restaurant and we can get the model by using the data. We gain the data from the .csv file that there are 14 features and 1 class. So we can use the 14 features as the training data and the other one class as the target which classify the data. But how can we create the model? There is function in the scikit-learning called ‘DecisionTreeClassifier(criterion='entropy')’. Criterion means the principle while choosing the feature and the default is gini and the entropy mean that we use the information gain to choose the features. There is another question that we need to consider about the performance of Decision Tree. There is a module in sklearn called cross-validation and the main function in this module is ‘sklearn.corss_val_score(clf,raw data, raw target,cv = n). The clf is the procedure and the raw data is the training set, the raw target is the class. Finally the cv means how many cross-validation time do we need and the default is 5.

We can get the accuracy of Decision Tree from above function. The accuracy is about 0.7. From this number we can not judge the module is good or not, the reason that why decision tree can have an accuracy above 50% is we use the information gain to predict the data. The decision tree is not like the other procedure, K-NN need to know how many neighbor the solution need. It is hard to find the proper k so that the accuracy might be lower than decision tree. There is another reason that the decision tree need the whole data set to fit the model and make the prediction.

2.2 K Nearest Neighbor

Secondly I am going to discuss about the K-NN. As the define of K-NN said we need to find the optimal k for this model and this is the disadvantage of this model. After get the data, we use the data to train the model and predict. There is also a function in the

sklearn called ‘KNeghborClassifer’. In this function, we can choose the number of neighbor to implement the procedure. But how can we select the best number of neighbor, I use a loop from 1 to 20 to find the best accuracy and the best neighbor. Cross-validation also be used in this function, so we are going to choose the k-nn classifier as the clf which is one of the condition of the cross-validation. So after that we can get the accuracy of the classifier.

By using the cross-validation we can compute the accuracy of the k-nn. It is about 0.6, we can easily find out that it is smaller than decision tree and the number of neighbor is 18. We can not say that the k-nn is the worst but the accuracy is less than decision tree, so we can commit that decision tree is better than the k-nn. The reason that the k-nn got smaller accuracy is that we need to find the number of neighbor and this is hard to find the optimal number. k-nn also need the whole data to predict so in some case the k-nn obtain the higher accuracy but in this case we abandon the k-nn.

2.3 Support Vector Machine

Finally I am going to talk about the support vector machine. We do not need the whole data set of previous restaurant. What we need to do is to find the ‘margin’ of the data set. Margin mean the border between the clusters. Above the margin belong to one class and the other belong to the other cluster. So we gonna to train the model in sklearn module and there is a function in this module called SVC. There are two parameter in the function, one is gamma and the other is C. gamma is a parameter come with the function,it determines the distribution of data mapped to the new feature space. The larger the gamma,the less the support vector. The C is Penalty parameter, in other word, it shows the procedure tolerance to error term. If the procedure have a big C, the procedure is easily going to underfitting.

We are gonna to implement the function in sklearn by using the function SVC. I choose gamma and C by using the Grid-search. The range of gamma is from 10^-1 to 10^3 and pick 9 numbers and in the same way the C’s range is from 10^-7 to 0 and pick 8 numbers. We also sue the cross-validation to compute the accuracy. We can find out that the SVM’s accuracy is much larger than two other module, it is about 0.9.

So what about the reason of the SVM has the largest accuracy. The SVM don not classify the data by the liner. In some case, there are non-liner SVM, it can show us the complexly relationship between the data. So the SVM got the largest accuracy. But there is a disadvantage of SVM, it always take a lot of time. So we need to consider about the question.

3. Summary

From above we can find the SVM have the largest accuracy but it take much more longer time than the other two procedures. Although it take more time to predict, I am going to use SVM as my learning procedure to solve this problem. Because the size of training data set is not too large to handle. We want to have a accurate result so I prefer using the SVM.