School of Computer Science and Electronic Engineering

CE802 Machine Learning and Data Mining

Assignment. Part 1

Tutor: Dr. Luca Citi

Student: Lyu Yaowei

Student Number: 1802697

Words: 697

1.BACKGROUND

According to the question, there is a chain restaurant manager want to use the knowledge of machine learning and data mining to predict whether the new restaurant benefit or not. The manager provide the previous restaurant data which show the success or not. We need to use these these data and find out the rules between the profitable and the location.

2.Main Body

2.1 The Type of Data We Need

First of all, as a large restaurant chain, we need to consider about the location of the restaurant. More particularly, the location is how far from the downtown, is there any working section near the restaurant and so on. Secondly we also need to know the decoration of the restaurant such as the indoor temperature and the style of the decoration. What's more, we need to know the style of the dishes such as the Thailand food、burger or the Western-style food. In my opinion, this is all types of data that I need, if we find out that the model is too simple and hard to predict the right answer we need to adopt the new categories of data

2.2 The Essential Predict Task

If we collect the data of the previous restaurant, what we need to do first is called the unsupervised learning which is find the pattern of the data and cluster the data. For instance, we collect the distance of the restaurant to the nearest office. What we need to do first is unsupervised learning. Find out the pattern of the data and clustering the data. We can use the restaurant profitable graph as the standard of the patterns. Then we need to do classification and regression which called supervised learning. In the first step, we already know the label of the data like within what distance the restaurant always profit. We can know from the table and the graph that consisted by the data. After that, we should do is regression, we need to choose the training set and the test set to fit the model. I would like to use the cross-validation to finish this step.

2.3 The learning procedures

In this case, I think that the K-NN, svm and the Decision tree is more fit than other models. First of all, the K-NN select the nearest neighbor. We can gain the data from the previous canteen and predict the output by computing the distance of the data and the previous data. For instance, we collect the data from the previous canteen and make a graph. Then we test a new data by calculate the distance between its neighbor, if the 'K' is three so we select the nearest three neighbor. The new data belong to the three neighbor's class. Second, we discuss the SVM, actually the SVM is finding the support vector of the data and the margin in order to classify the data properly. The advantage of the SVM is we don not need all the data set. What we need is the "margin data" and we can find the margin so that we can predict the data. Finally I want to talk about the Decision Tree which is a basic model in machine learning and data mining. The decision tree is going to compute the information gain by using the existing data and compare with the other categories. Then chose the larger information gain and the new data belong to its class.

2.4 How can we evaluate the performance

First of all, we need to use the cross-validation to test the model by using the existing data. We can select the data into 10 folders and use 9 of them as the training folders and the other on as test folder. Then we can use the Confusion Matrices to compute the recall and the precision. We can compare the two numbers and select the best performance.

3. Summary

In conclusion, we can adapt the SVM, K-NN and Decision Tree to predict whether the newly open restaurant profitable or not. What we need is historical data of the previous restaurant and use the machine learning knowledge to implement the task.