ML Problem Set 1——Weihan Chu——wcm350

1.Which attributes appear to have outliers?

chlorides, free sulfur dioxide, total sulfur dioxide, citric acid, volatile acidity, PH

2.What is the *accuracy* - the percentage of correctly classified instances - achieved by *ZeroR* when you run it on the training set? Explain this number. How is the accuracy of *ZeroR* a helpful baseline for interpreting the performance of other classifiers?

In this case, this means ZeroR predicts all the class value are bad. But actually, there are 1179 bad and 711 good. So the correctly classified instances is 1179 and the rate is 1179/1890, which is 62.381%. Generally, the labels on the test set are supposed to be the actual correct classification. Performance is computed by asking the classifier to give its best guess about the classification for each instance in the test. Then the predicted classifications are compared to the actual classifications to determine accuracy.

For the second question, A zeroR classifier simply assigns every value to the most common class. This means if your data is 51% A and 49% B, then ZeroR will get 51% right. If your data is 33%A, 32%B and 35%C, then ZeroR get 35% right. So ZeroR is a very basic and simple classifier you get. Given a certain data set, you can use ZeroR to find out what minimum performance you may expect.

3.Using a decision tree Weka learned over the training set, what is the most informative single feature for this task, and what is its influence on wine quality?

The most informative single feature is alcohol, because decision tree use greedy algorithms. It choose the attribute that minimize the remaining information needed every time. So if we want to know which is the most informative single feature, we can just find the first attribute that the decision tree chose to split. This is alcohol.

Totally, the higher the alcohol, the more likely the wine's quality is good.

4.What is 10-fold cross-validation? What is the *main* reason for the difference between the percentage of Correctly Classified Instances when you measured accuracy on the training set itself, versus when you ran 10-fold cross-validation over the training set? Why is cross-validation important?

Using training set:	Correctly Classified Instances	1812	95.873 %
10-fold cross-validation:	Correctly Classified Instances	1625	85.9788 %

10-fold cross-validation works in this way: first break data into 10 sets of size n/10, there train on 9 datasets and test on 1. finally repeat 10 times and take a mean accuracy.

Using training set: The classifier is evaluated on how well it predicts the class of the instances it was trained on, which means that you use the training set for testing. This option usually gives overly optimistic estimates of the classifier's future performance.

Using 10 cross-validation: The method is described above. So this is the reason for the difference between two test methods.

Cross-validation is important because the data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased.

5.What is the "command-line" for the model you are submitting? For example, "*J48 -C 0.25 -M 2*". What is the reported accuracy for your model using 10-fold cross-validation?

Native Bayes

78.8889%

6.In a few sentences, describe how you chose the model you are submitting. Be sure to mention your validation strategy and whether you tried varying any of the model parameters.

I use Naive Bayes model and I use the “Use training set ” test option. for the parameter’s of this model, there are four main parameters in settings: debug, displaymodelinoldformat, use kernel estimator, usesuperviseddiscretization. There are all initially false. When I change some form false to true, the attribute in output will become different and the correctly classified instances will become different.

7.A Wired article from several years ago on the 'Peta Age' suggests that increasingly huge data sets, coupled with machine learning techniques, makes model building obsolete. In particular it says: This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology… In a short paragraph (about four sentences), state whether you agree with this statement, and why or why not.

I disagree with this statement, I think machine leaning model are still very useful. ML techniques are the algorithms we use within the program. we use those algorithms to create a learner then learns a model from the data set. So because there are many data sets, we have many different models. But the model is very useful for us to solve a problem. model can help us predict the new data which we don't know the target.

8.Briefly explain what strategy you used to obtain the Classifiers A and B that performed well on one of the car or wine data sets, and not the other.

a. cars data sets:

J48 -C 0.25 -M 2 (decision tree) with default settings accuracy: 89.8319%
DecisionStump with default settings accuracy: 70.5042%

b.wins data sets:
J48 -C 0.25 -M 2 (decision tree) with default settings accuracy: 85.9788%
DecisionStump with default settings accuracy:80.8466%

9.What is the key difference about the output space for the car task, as compared to the wine task?

Then have different number of classes. The car task has four possible output while the wine task only has two.