CS589 homework2

Yifu liu

March 2018

1 Question 1

- SVM: kernel, gamma
  Since SVM has the better performance for two classifications, and in this project, that asks us to distinguish if those samples are ads or not ads.
- Random forest: n_estimators, min_samples_leaf, min_samples_split
  Fast, low bias, it works will with multiple numerical and categorical features, which has been showed in the result below.
- Logistic Regression: penalty, C
  Logistic regression deals with this problem by using a logarithmic transformation on the outcome variable which allow us to model a nonlinear association in a linear way

2 Question 2

Honestly, it's hard for me to find the relationships between those three classifier, since SVM is better for two classes but there should be a wide boundary between them, but Random forest and Logistic Regression don't have. Random forest, it works will with multiple numerical and categorical features, which has been showed in the result below. For the Logistic Regression, it has the similar speed and the storage with Random Forest, but SVM uses huge amount of space while it's running.

3 Question 3

a. Graph(Next page)

b. Overall, the Random forest is the best, since the accuracy is the highest. Logistic Regression is the fastest classifier with acceptable accuracy. In my opinion, I would like to use Random forest, because the accuracy the highest one and the time spent for Random forest is only slightly slower than Logistic Regression.

classifier	training time	predict time	accuracy
SVC	113.28s	20.40s	49.79%
Random forest	0.73s	0.011s	88.15%
Logistic Regression	2.623s	0.0032s	83.53%

Table 1: Table for Question 3-1

4 Question 4

a. graph

Classification	n_estimators	min_samples_leaf	min_samples_leaf
Random Forest	range(300, 600, 100)	[1]	[5]

Table 2: Table for 4-a

n_estimators	min_samples	min_samples	mean_train_	mean_test_score
300	1	5	0.99862963	0.90266667
400	1	5	0.99866667	0.90325926
500	1	5	0.99862963	0.90296296
600	1	5	0.99866667	0.90348148

Figure 1: Data for 4-a

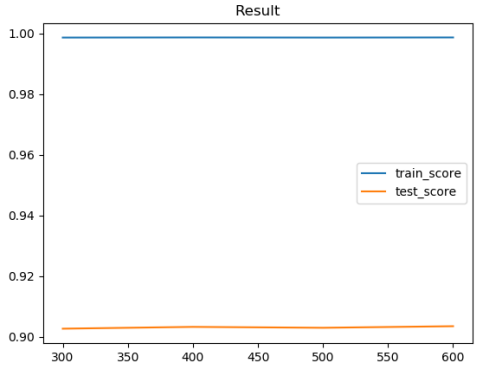

Figure 2: Graph for 4-a

b. We should use training_set to train our model, and then use test_set to predict our model. Even though the score of training_set is higher than the test_test, but if we use training_set to train and predict, that does not make any sense, because we training_set is as same as itself. However, if we use test_set to predict, that gives more accurate information, which tells us if the performance of training_set.

c. Using the parameters that generated by our-self is better than the default parameters based on the accuracy.