SVM

Part I: Data Processing

Firstly we delete the ids and set the price as factors 0, 1, 2 based on the price range 3500 and 10000. Then we change the categorical variables (cut, color, clarity) into numerical variables. After processing the data, we can select the parameters to choose a best SVM model.

Part II: Select Model

Firstly, we split the data set into training set (90%) and test set (10%).

Then, we could train the model and compute its test errors by 10-fold cross validation with different parameters.

```
Cost=c(0.01,0.1,1,10,100)
Gamma=c(1,2,3,5,10)
```

Since running the source code need a lot of time (for the size of the data set and the complexity of svm multi class model), we save our variables in a RData file and thus to avoid running it again. The source codes are in the appendix.

1. All Variables on No-scaled Data

At the beginning, we use all of nine variables based on the no-scaled data to build the model with different parameters. And the best parameters for 'linear', 'polynomial' and 'radial' kernel based on the no-scaled diamond data set are shown in Table 1.

Table 1. *SVM Error*

kernel	cost	gamma	degree	TE
linear	10	0.111	NA	0.0407861
polynomial	100	0.111	3	0.0394883
radial	1	1.000	NA	0.0357805

From Table 1, we can choose the radial kernel SVM model with cost = 1, gamma = 1 as our best model for it has the smallest test error 3.578%.

2. All Variables on Scaled Data

Since SVM is a distance-based model, the scaled variables may improve the performance of the data. Thus we use the scaled training set to fit the SVM model, and the best performance for 'linear', 'polynomial' and 'radial' kernel of the scaled data are shown in Table 2.

Table 2. *SVM Error of Scaled Data*

kernel	cost	gamma	degree	TE
linear	1	0.111	NA	0.0407861
polynomial	10	0.111	3	0.0411568
radial	1	1.000	NA	0.0355951

From Table 2, we can choose radial kernel svm model with cost = 1, gamma = 1 as our best model for the scaled data. The test error rate is 3.560%, which is lower than the test error of no-scaled data set.

3. '4C' Variables on Scaled Data

Finally, as we know that the price of a diamond mainly depends on the '4C' metrics, which is the carat, cut, clarity and color of a diamond. So, if we only use these four variables to classify the price of a diamond, would the performance of the model be improved?

The best performances for SVM models with different kernels are shown in Table 3, which are build only by the scaled 4 variables.

Table 3. *SVM Error of Scaled Data by 4C*

kernel	cost	gamma	degree	TE
linear	10	0.25	NA	0.0424546
polynomial	10	0.25	3	0.0424546
radial	1	1.00	NA	0.0315165

From Table 3, we choose radial kernel svm model with cost = 1, gamma = 1 by the '4C' variables as our best model for the scaled data. The test error rate is 3.152%, which is even lower than the test error of scaled data set.

Part III: Best Model Performance

In conclusion, the best SVM model for the diamonds price prediction is of radial kernel with cost = 1, gamma = 1 built by the 4 scaled variables: carat, cut, clarity and color. It has 5477 support vectors. It has the following functions:

svm(formula = price ~ carat + cut + color + clarity, scale(data), kernel = "radial", cost = 1, gamma = 1)

Then we can use our best SVM model to get the F1 score Table for training and test set. They're shown below as Table 4 and 5. From the tables below, we can get test error = 3.152% and training error = 3.073%.

Table 4. *F1 Score of Training Set*

	0	1	2
0	28783	425	0
1	477	13847	270
2	0	320	4424

Table 5. *F1 Score of Test Set*

	0	1	2
0	3182	46	0
1	53	1548	34
2	0	37	494

Part IV: Analyze Model

Finally, we plot the misclassified points for the training set and test set below. We can see that most misclassified points are distributed near the boundary, except for some very high prices (above 10000).