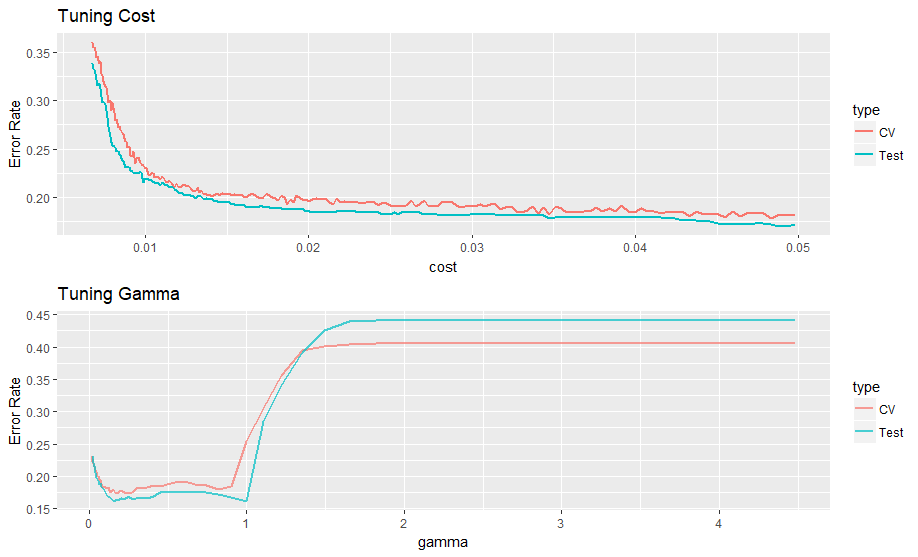
**SVM**

For model selection, I divide the original dataset with 2192 observations into 30% test dataset and 70% train dataset.

In order to acquire better prediction consequence, we need to test how sensitive the results are to some tuning parameter. As we all know, radial kernel has a more complex structure. Thus we use radial kernel for the SVM analysis which will result in classification with low error rates. Firstly we try different values for the cost.

The belowing figure shows the test error of SVM on classifying climate dataset using different cost and gamma. Furthermore, in tuning cost plot we use gamma=1/9, and then we choose the best cost with a lowest error rate for investigating the sensitivity of the gamma parameter to the error rate.The best cost is 0.05



The figures show a rationale that as the cost increase, the error rate will become decreasing which has a concave shape and as when the gamma has no distinct changing at start but increase suddenly after 1. We still choose the best gamma with a lowest error rate according to the cross validation results. Gamma is 0.13533528.

Again, we use such parameters to fit an appropriate model of SVM for such dataset to predict whether it will rain or snow tomorrow.

**Confusion Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | True condition | |  |
|  |  | Non-rain | Rain |  |
| Prediction | Non-rain | 326 | 47 |
| Rain | 54 | 231 |
|  | Error rate | 0.142 | 0.169 | 0.1535 |
|  | Type I | Type II | Overall |

Based on confusion matrix, we can evaluation the efficiency and accuracy of the results instantly. The confusion Matrix above shows some basic criterion to us. We can see that Type I error is 0.139, Type II error is 0.173 and overall error is 0.1535. Because the aim we pursue is predicting the precipitation days, so we concentrate more on the scenario that tomorrow will have precipitation. So the Type I error represents non rain days with a wrong forecasting of raining and Type II error represents rain days with a wrong forecasting of non rain.

Then we use this model for NYC climate data in 2017

**Confusion Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | True condition | |  |
|  |  | Non-rain | Rain |  |
| Prediction | Non-rain | 96 | 45 |
| Rain | 142 | 81 |
|  | Error rate | 0.596 | 0.357 | 0.5137 |
|  | Type I | Type II | Overall |

Thus we can conclusion that type I error is 0.596, type II error is 0.357. This model performs not good at predicting the tomorrow precipitation in NYC. And the results for rain days with a wrong prediction is better than non rain days with a wrong prediction.