**Introduction:**

Purpose

This report focuses on finding a better method for forecasting whether it will precipitate tomorrow by comparing some classification methods such as KNN, Random Forest, SVM and Neural Network. We implement some data preprocessing methods and mathematical transformation on the original dataset. Thus, you could apply an essential method in some cities’ daily climate dataset and draw a conclusion about the participation tomorrow.

Background

A large amount of weather data is collected by meteorologist during the day. One thing that they intend to do is tracking the atmospheric conditions change constantly and making the prediction for the weather more accurately and up-to-date. And we could evaluate and predict the weather from some dimensions like temperature, humidity, dew point, wind speed and direction and so on.

Data source

The data we collected is from NOAA(<https://www.ncdc.noaa.gov/cdo-web/>)

The National Oceanic and Atmospheric Administration is an American scientific agency within the United States Department of Commerce that focuses on the conditions of the oceans and the atmosphere.NOAA warns of dangerous weather, charts seas, guides the use and protection of ocean and coastal resources and conducts research to provide understanding and improve stewardship of the environment.

Attributes explain

WDF2 = Direction of fastest 2-minute wind (degrees)

WDF5 = Direction of fastest 5-second wind (degrees)

WSF2 = Fastest 2-minute wind speed (miles per hour or meters per second as per user preference)

WSF5 = Fastest 5-second wind speed (miles per hour or meters per second as per user preference)

TMAX = Maximum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file

TMIN = Minimum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file

PRCP = Precipitation (mm or inches as per user preference, inches to hundredths on Daily Form pdf file)

AWND = Average daily wind speed (meters per second or miles per hour as per user preference)

PGTM = Peak gust time (hours and minutes, i.e., HHMM)

TOBS = Temperature at the time of observation (Fahrenheit or Celsius as per user preference)

WT\*\* = Weather Type where \*\* has one of the following values:

01 = Fog, ice fog, or freezing fog (may include heavy fog)

02 = Heavy fog or heaving freezing fog (not always

distinguished from fog)

03 = Thunder

04 = Ice pellets, sleet, snow pellets, or small hail

05 = Hail (may include small hail)

06 = Glaze or rime

07 = Dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction

08 = Smoke or haze

09 = Blowing or drifting snow

10 = Tornado, waterspout, or funnel cloud

11 = High or damaging winds

12 = Blowing spray

13 = Mist

14 = Drizzle

15 = Freezing drizzle

16 = Rain (may include freezing rain, drizzle, and freezing drizzle)

17 = Freezing rain

18 = Snow, snow pellets, snow grains, or ice crystals

19 = Unknown source of precipitation

21 = Ground fog

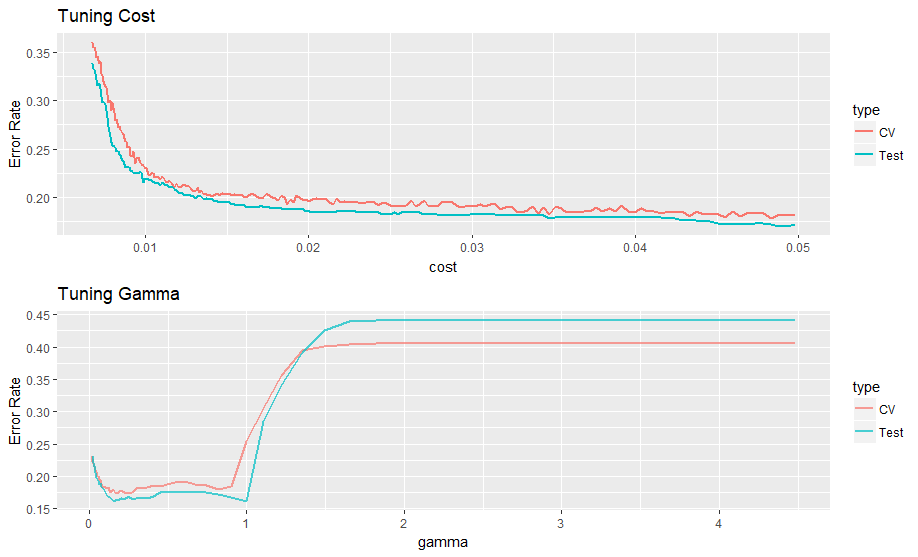
22 = Ice fog or freezing fog

**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 | 72 | 69 |
| Rain | 102 | 121 |
|  | Error rate | 0.586 | 0.363 | 0.4697802 |
|  | Type I | Type II | Overall |

Thus we can conclusion that type I error is 0.586, type II error is 0.363. 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.