Wen Sheng
Taylor Arnold
February 25, 2016

Problem Set 03 Data Analysis

Data Mining and Machine Learning

1. How did you tune the various parameters in your models?

Basically, I choose Random Forest Model to do the prediction and the error rate turns to be lower than 0.35. During the tuning process, there are following different parts I focus on tuning.

A. Property settings for constructing the random forest tree

For this part, at first I use the demo code: outRf <- randomForest(Xtrain, factor(ytrain), maxnodes=10. This settings turns out to be really slow (around 3 minutes) for training and the accuracy is around 0.38 - 0.41, which is both not accurate and too slow.

Then I modified the settings to: outRf <- randomForest(Xtrain, factor(ytrain), ntree=11). It drastically reduces the error rate to 0.25 - 0.35 and training process only takes several seconds (around 30 seconds).

B. Choosing and comparing variables relationships from XTrain data

At first I use all the variables in train data to predict and the error rate is around 0.39. Then I check all the variables and cluster all the data in to different type:

- un-relevant data(eg. testFlag): delete this column before training
- highly-correlated data (eg. district information with community area information)
- time-related data(month, year, hour): I tried several combinations of using this 3 columns and found using all of the data together will get the most accurate prediction.

2. How did you use hierarchical modeling (use the prediction in one model as an input to another) or stacking to combine multiple models together?

I did classification, factorization before training the random forest model.

A. Classification

Because I treat all time data as a categorical data not numerical data. Each hour, month, year represent a different level. 24 levels in hour data are too cumbersome and have more sense to classify them before train. So as the preprocessing step of using the random forest model construction, I classify the hour data into 4 different classes: **morning(5am - 12pm), afternoon(12pm - 5pm), evening(5pm - 9pm), night(9pm - 4 am).**

B. Factorization

Location data **loc**, area **data district, ward, communityArea** is categorical data and should not be treated as numerical data. So I factorize these columns before injecting all the data into random forest model.

---

3. How did you incorporate the categorical 'location' variable? Did it influence the model significantly

Random forest model cannot accept string data, so I factorize this location variable and it make the random forest model can be worked in this prediction.

---

4. How did you incorporate the categorical variables such as beat, district, ward, community area

These data are highly correlated: **Community area** is high depending on the **district code;** the **ward number** is highly depending on **combination of hour and district code**. so I factorize district code and drop the column of community area as its information is already relatively represented in district code and this also reduce the error rate around 1 - 2 %.

```
> print(sum(error)/length(error))
[1] 0.2620871
```

5. Describe the confusion matrix. What is the easiest category to differentiate. Why? What categories are hard to tell apart? Does this make sense to you?

```
OOB estimate of  error rate: 30.99%
Confusion matrix:
     1     2     3     4     5 class.error
1 39724  6674  9763  1202  6556  0.37852595
2  8095 43431  4927  4298  5342  0.34288049
```

```
3  8988  4210 39586  3737 12350  0.42521526
4   727  2378  2346 60563  1118  0.09785199
5  4441  3578 11213  1717 47498  0.30606162
```

The information shows class 1. theft over $500 has an error rate around 0.38, class 2. simple battery has an error rate around 0.34, class 3. deceptive practice has an error rate around 0.42(**highest**), class 4. narcotics has an error rate around 0.098(**lowest**) and class 5. burglary around 0.30. It also shows that class 4. narcotic has lowest frequency and class1. theft over $500 have highest frequency.

As the factors are mainly **time/location data**, it shows that crime **class 4. narcotics is highly correlated with time and location,** some district have higher rate of narcotic crime while other district remains lower rate. **Class3. deceptive practice** is not that highly correlated or not so depending on time and location information but on other factors not included in the training set.

---

6. What you expect your mis-classification rate on the test set will be?

I did experiment on training and predicting model by running iterations around 25 times and all of them the error rate for 20% of the test data turns out to be around **25 - 35%.**