Econ498 Midterm Summary

Yan Hao

Feb 28, 2021

1 Program Design

This program is designed for predicting the criminal type. There are five steps from data cleaning to model selection.

1.1 Resize dataset

Since the raw dataset is quite large, I would like to only focus on criminal records from 2018 to 2020. The sample data are saved to sample_data.csv.

1.2 Draw subset from sample dataset

There are more than 700,000 observations in sample dataset. There would be 111 columns after I form the dummy variables. Hence, it is necessary to draw a subset from the sample dataset to simplify the problem. I have drawn 10 percent of observations to form the subsample, and saved to sub_sample.csv.

1.3 Plot criminal type

I would like to find the pattern of the occurrence for different type of crime. So I specify each type of crime, and plot their locations with latitude on the vertical axis and longitude on the horizontal axis. Figure for each criminal type can be found in figures directory.

1.4 Data cleaning

The program would first drop observations with missing values, extract useful variables, i.e., Date, Primary Type, Arrest, Domestic, etc. Since Arrest and Domestic columns are in Bool values, the program converts True and False to 1 and 0. I would like to say that the occurrence of criminal case might happen seasonally. Hence, I decompose Date to Month and Hour_Slot. I would try to find if particular type of crime would happen during a specific time period. Moreover, the program convert District, Community Area, and Year to dummy variables.

And you can choose to form crime_count columns by setting count_crime = True when you call CleanData. Since each observation belongs to a district, I would like to see if the sum of total number of each type of criminal case can improve the accuracy of the prediction. An example is shown below.

Obs	**District**	**BATTERY**	**CRIMINAL TRESPASS**		**OTHERS**
obs 1	10.0	810	76	...	0
obs 2	18.0	483	115	...	0
obs 3	10.0	810	76	...	0

Observations would have same values in these columns if they belong to same district, because I assume if the frequency of one type of crime is higher than others, it would be more likely that the Primary Type of this observation is this type of crime.

If you set count_crime = True, the clean dataset will be save to CleanData_with_crime_count.csv, otherwise, the dataset would be saved to CleanData_without_crime_count.csv.

1.5 Model selection

Since Primary Type is categorical, RandomForest (RF), Logistic, and SVM can be used to do the prediction. Same as before, it takes much longer time for Logistic and SVM. Hence I do a grid search for RF model. The accuracy rate for each split is shown in Figure 1 and 2. The accuracy rate decreases while the maximum depth increases. Also, it seems that the machine works slightly better with dataset without crime_count.

2 Figures

Figure 1: Accuracy rate with crime count

Figure 2: Accuracy rate without crime count