Machine Learning 498_Final Write-Up

Nick Rucker

May 3, 2019

1 Preface

The target of the various machines being used is the IUCR of a crime. The IUCR is more specific than just the type of crime itself, but closely related. There is more information embedded in the IUCR which adds an extra layer of depth and understanding about the crime that was committed which is why I wanted to predict IUCR rather than simply the general type of crime.

2 About the Data

The original data set contains over 6,000,000 observations pertaining to crime in Chicago starting in 2001. Aspects such as the crime, description, IUCR, longitude, latitude, if the perpetrator was arrested, etc... are recorded in the data-set. First, some observations need to be dropped. If an observation had any missing value(s), it was dropped. It is problematic for a myriad of reasons to train a model with missing data so those observations had to be discarded before any further data manipulation. Next I decided to randomize the data. Since I chose to only use a sample of the data, collecting a random sample is important so selection bias is greatly reduced and hopefully not an issue. While the number of observations is definitely a strength of the data-set, I chose to limit the number of observations to 10,000. Lastly, I chose to map each crime to a number. This is mainly to help predict the IUCR based on the details of the crime (more on this in the next section).

Given a more powerful computer and time for training, all the valid observations would have been great to use. Reducing the size of the data-set was a decision I made which sacrificed higher accuracy and precision for the speed of the machine.

3 Machine Learning Process Overview

As previously stated, the goal of the machine is to predict the IUCR of a crime given a set of descriptive features. The parameters I used to predict the target were: 'Arrest', 'Domestic', 'District', 'Ward', 'Community', 'Latitude', 'Longitude', 'Community Areas',and 'crimeCode'. Arrest and Domestic are key to predicting the IUCR of a crime because the various positive/negative combinations of two aspects are key in what IUCR a crime gets. District, Ward, Community, and Community Areas were used to help identify patterns in a particular location with respects to crime. Maybe certain community aspects make a community particularly susceptible to a certain IUCR. Latitude and Longitude are helpful with clustering and mapping the various locations of crimes. Lastly, crimeCode adds a higher predictability to an IUCR. If crimeCode is not included the IUCR is very difficult to predict because not only does the machine have to predict the type of crime correctly, it then has to predict the correct IUCR based off the crime that was predicted. More advanced techniques are needed in order to create a viable

model to predict the IUCR without a crimeCode independent variable, so I opted to include crimeCode. I did not use variables in the data-set like zip-code because I believe the coordinates are a better option and including them will create a bias, Census Tract was not included because it is closely correlated with 'Community', other variables lacked relevance to the model.

3.1 Random Forest

The Random Forest machine had a total of 21 branches. To decide the number of branches I created a dictionary which had the number of branches as the key and a list of accuracy scores as the values. I gathered 500 scores for 1 to 50 branches and chose the highest average, 21. The average accuracy score was .4461 which can be interpreted as the machine matched 44.61% of the observations' IUCR in the test sample based off the training sample. This accuracy score is adequate in my opinion. In no way is this machine a definite way at predicting IUCR, but all elements considered, it is not terrible.

3.2 Decision Tree

The Decision Tree machine was similar to the Random Forest in many ways. The accuracy was adequate, but not quite as high. The model had an average after 500 samples of .397 which is slightly lower than the Random Forest, but to be

expected.

3.3 KNN Classifier

Clustering provided interesting results. To determine the number of clusters to use, I used the same technique to determine the number of branches in the Random Forest. After collecting 500 samples of 1 - 25 clusters I discovered that nine clusters provided the highest average accuracy score of .41. This places the KNN machine in between the Random Forest and Decision Tree in terms of accuracy. These results are not entirely surprising. I suspected similar crime to cluster in similar locations. KNN uses euclidean distance to determine clusters which pairs well with the longitude and latitude coordinates which explains why it has a slightly higher accuracy than a decision tree, but a random forest is more complex with how the probabilities are calculates which is why the random forest is slightly more accurate.

3.4 Linear Model

The linear model ultimately proved to be not a reliable model to predict a crime's IUCR. This is not too surprising given the type of data the model was trained off of. The r^2 only averaged to be .114 after 500 tests. Although the sample was random it seems there was an endogeneity problem which misconstrued the

results of the model. With that being said, a linear model is certainly not the best machine to use when trying to predict an outcome like IUCR.

4 Final Statement

crimeCode definitely helped in improving the accuracy, as the a erage accuracy score before adding the crimeCode independent variable was .093 for the random forest, .133 for KNN, .0843 for the decision tree, and .047 for the linear model. The low accuracy makes sense though. The probability exponentially decreases as the machine had to predict first the type of crime correctly, then predict the IUCR correctly. The likelihood of acheiving both is quite small, so although the accuracy was low, it is not too terrible. So needless to say the IUCR is very difficult to predict without knowledge of what the crime is. To further improve the accuracy of the machines, having the time of day the crime occurred and a numerical representation of where the crime occurred would improve the accuracy of the model. Ultimately though, the process was a positive experience.