Analysis of Dangerous Crimes in Boston

Jacquelyn Andrade

Yao Zhang

Abstract:

In our project we will be analyzing the types of crime that occur in given locations around Boston on specific times of day. We will perform K-means clustering on our data to get the most popular crime per district. We then plan to use regression techniques to be able to predict the amount of occurrences of a specific crime as well as the probability of the crime occurring within a district. Our findings should closely correlate to “typical” crimes associated with certain times of the day. For example, we do not expect to find a substantial amount of crimes such as DUI in the morning, whereas robberies are more likely to occur at any time of day.

Introduction:

We would like to create an interactive user application that will allow our audience to query search results from our data mining techniques. Our categorical and predictive analysis will hope to benefit the Boston Police department in granting more insight into the nature of crimes prevalent in certain areas and improving police response. Given a specific time, day, and district, we hope to be able to provide an appropriate prediction for police to account for the likelihood a certain crime will occur.

For this project we will be using the Crime Incident Reports (July 2012 - August 2015) dataset downloaded from the City of Boston open data portal. This dataset was the most current and applicable to our analysis. The main motivation of this project is to analyze potentially “dangerous” crimes e.g. Larceny, Battery, Possession, Assault.

We will be analyzing the types of crime that occur in given locations around Boston on specific times of day and categorizing them using K-means. We will also utilize regression techniques to find correlations and predictors of crime. The time of day specific for our K-means technique is classified as: Morning (6 - 11), Afternoon (12 -16), Evening (17 - 19), and Night (20 - 5). We use integer values to identify the days of the week: Sunday (1), Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6), Saturday (7).

Technique:

We chose to use the K-means algorithm to classify our districts by crime popularity. K-means uses hard assignments, such that each point belongs to one cluster or another but never belongs to more than one. Another clustering technique, Gaussian, makes soft assignments: Data points do not belong to a single cluster but have some probability of belonging to each individual cluster. This means some points may have a high probability of belonging to the same category due to their vicinity with each other and similarity of categories. We found the Gaussian mixture model to be inappropriate for this project because we wished to distinguish data points specifically for districts without having clusters mix points together. Our K-means technique vectorizes on district and offense type. We then found the top unique district and crime type per cluster on a specified time and day of the week.

The purpose of using regression techniques is to allow the user to see whether time is a good predictor for the types of crime that will occur. This is possible by finding direct predictions on the amount of crimes and the probability of each one occurring. For our linear regression technique, we will be predicting the occurrence of a crime given the district, day of week, and time. Using the logistic regression technique, we will be analyzing the probability of a certain crime occurring given the district, day of week, and time.

Datasets and Experiments:

This dataset contains field values for the crime incident type, time, day, month, and year of incident, weapon used, true false values for whether shooting was involved and if the incident was domestic, along with the coordinates, street name, reported area, and district. Pre-processing will be necessary as we need to eliminate row values that are null as well as incident types that are not relevant to our study.

We experimented on the the number of clusters used in K-means and found the value of k with the minimal error to be 12, the total number of unique districts in our dataset.

In our linear regression technique, we experimented on the types of crimes that would be used for our regression function. Our linear regression function returns the predicted amount of times a specific crime will occur. We found that in crimes less common in the dataset, the R squared value was extremely low and the coefficients were more likely to be more negative. This was expected as our function will not be able to accurately predict the amounts of crimes if there is not sufficient data to spread across all days and hours of the week. For more prevalent and popular crimes, we found our regression statistics to be more informative and accurate. In crime types such as Assault/Battery our regression prediction was more likely to give us a definitive and more accurate response.

In our logistic regression technique, we experimented on the types of crimes to be used in our regression function. Our logit function returns the probability of a specific crime occurring as opposed to all other crimes. We used our dataset on previous crimes to train our regression function in order to make these predictions. We were able to do this with a great amount of accuracy. Overall, we found Logit Regression to be more informative to the user as opposed to

OLS. For crimes that have little to no occurrence, Logit returned a probability close to zero whereas OLS would predict that the crime would occur at least once.

Overall in both OLS and Logit Regression techniques, we noticed that hour was more likely to be a predictor for crime than the day of the week.