Analyzing Crime in Chicago Through Machine Learning

A full report consisting of all of the components mentioned in the document

JANUARY 27, 2020

BEN

Introduction where you discuss the business problem and who would be interested in this project.

The city of Chicago publishes an up-to-date list of all reported crimes. The records span from 2002 to the modern date, allowing a few days of delay to catalog the crimes and publish them. According to official FBI data, Chicago is one of the leading cities in homicides, having more than quadruple the amount of crimes in New York City, and more than double the amount of crimes in Los Angeles [3]. Being able to parse the data presented by these huge data sets is a problem central to understanding the crimes in the city of Chicago. By analyzing the data in a mathematically rigorous way, researchers may be able to glean insight into the underlying causes of crimes, and also may be able to figure out indicators of future crimes to occur. All of this categorization and analysis falls under the umbrella of Data Science, a field which analyzes large sets of data using probability and statistics, and makes useful conclusions from the analysis. In this paper, we will attempt to parse the city of Chicago's up-to-date dataset, and try to perform some crime "prediction." We will do this by utilizing techniques from machine learning: specifically, K-means clustering.

Data where you describe the data that will be used to solve the problem and the source of the data.

The data set is publicly available through the city of Chicago’s website [2]. The information presented in this data set is quite comprehensive, including information about the date and time of the crime, location of the crime, type of crime, etc. For the purposes of this paper, we will focus on the time of the crime and the type of crime. The type of crime is given a standardized set of codes called the Illinois Uniform Crime Reporting (IUCR) codes. Thus, each IUCR corresponds to a specific type of crime. The list of crime codes and corresponding crimes can also be found through the city of Chicago’s website [1].

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

The Data Set

The data set is publicly available through the city of Chicago’s website [2]. The information presented in this data set is quite comprehensive, including information about the date and time of the crime, location of the crime, type of crime, etc. For the purposes of this paper, we will focus on the time of the crime and the type of crime. The type of crime is given a standardized set of codes called the Illinois Uniform Crime Reporting (IUCR) codes. Thus, each IUCR corresponds to a specific type of crime. The list of crime codes and corresponding crimes can also be found through the city of Chicago’s website [1].

K-means Clustering

K-means clustering provides a way to group data points together in a way that minimizes differences between the data points in the same group. By applying these methods, we can take n data points and partition them into k clusters. The algorithm seeks to minimize the following function:

arg min S k ∑ i=1 ∑ x∈Si kx − µik 2

In this equation, x is a vector that corresponds to a specific crime instance in the data set. µi is the "centroid." It is is the average point (also a vector) in a given cluster. The entire algorithm works by minimizing the distance between all data points (crimes), and the corresponding centroid of the cluster that they’re in. We square the Euclidean distance between these two points to make all distances positive, a common technique in statistics. S is the set containing all cluster assignments. Thus, Si contains all the points in the ith cluster. We sum over all the elements in a given cluster, Si , and then we sum over all the different clusters, to obtain a total "distance" of all points from the centers of their corresponding clusters. We then attempt to minimize this by finding the optimal assignment of clusters, S.

Serialization of Data

In order to make K-means clustering viable to our data set, we must serialize the data to convert all crime information into a corresponding point on the graph. For simplicity, we are focusing on three specifications of a crime: the date the crime occurred, the time the crime occurred, and the type of crime that occurred.

RESULTS

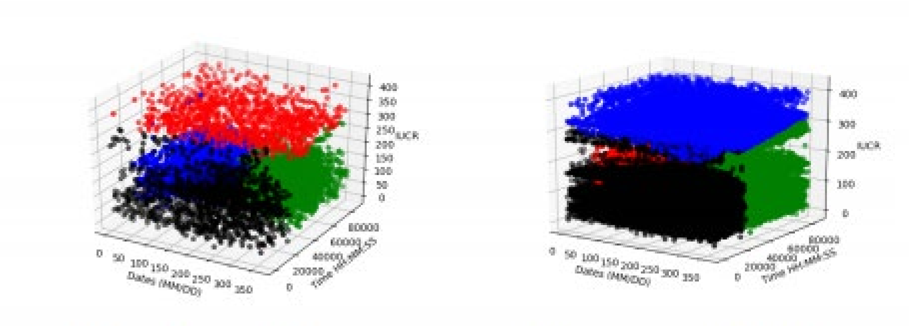
Immediately, by plotting these points we can garner insight into the underlying structure of the data that would not otherwise be visible in just a spreadsheet. We notice that crimes corresponding to IUCR codes between 200 to 250 are very spare in contrast to other crimes. When cross-referencing this with the database of IUCR codes provided by the state of Illinois, we obtain that these values of IUCR codes correspond to sexual crimes. Immediately, we see that sexual crimes such as rape or sexual assault are far less frequent than other crimes.

Also gather that gambling is far less frequent than other crimes by the same method. It’s worth noting that the database includes all reported crimes, but unreported crimes are not listed in the database (of course, that’s because no one has reported them!), so it may not be entirely accurate of what actually happens in the city of Chicago.

Results section where you discuss the results.

Now that data is serialized, we can run our Python implementation of Lloyd’s Algorithm and K-means Clustering. We have to make a choice of value of K (decide how many clusters to partition into). We partition into 4 clusters first.

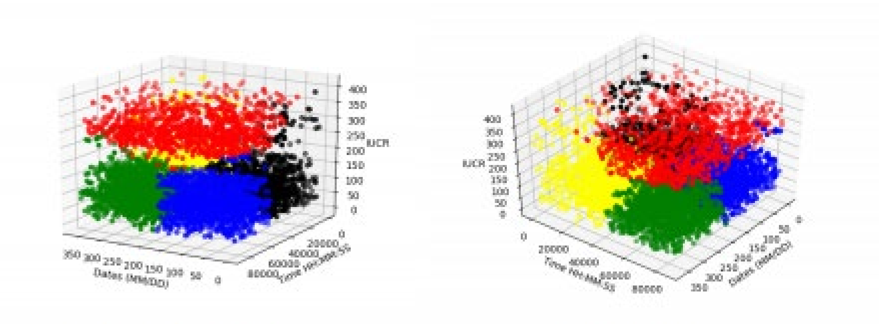
When run, we obtain the following output:



(a) *Sampling every 1000points*

(b) *Sampling every 100 points*

Figure 3: *Plots showing the result of K-means clusters for K = 4*



(a) *The default view*

(b) *A rotated view*

Figure 4: *A plot of a sampling every 1000 points, when K = 5*

Discussion

This project has largely been a proof of concept to show that clustering data sets about crime is entirely possible, and allows us to gain insight into the underlying trends of crime. i. Future Work Clustering data and analyzing it is an important step to many other machinelearning algorithms. In particular, prediction of crime points can be achieved by multiple methods: • Minimizing the distance between a crime occurrence and the centroid of a cluster • Performing regression analysis on the identified clusters and fitting crimes to the best fit line In the current implementation, we can take a partially filled out crime report and match it to the most likely cluster based on the minimization of the dimensions provided. This leads to the possibility of studying crimes on specific dates, times of the day, or specific types of crime. For example, perhaps law enforcement would be interested in knowing the most statistically likely type of crime happening on January 1 at midnight.

REFERENCES

1) *Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes | City of Chicago | Data Portal. url: https : / / data . cityofchicago . org/Public-Safety/Chicago-Police-Department-Illinois-UniformCrime-R/c7ck-438e/data.*
2) *Crimes - 2001 to present | City of Chicago | Data Portal. url: https://data. cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzpq8t2.*
3) *Eliott C. McLaughlin. With Chicago, it's all murder, murder, murder ... but why? Mar. 2017. url: http://www.cnn.com/2017/03/06/us/chicagomurder-rate-not-highest/.*

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

Conclusion section where you conclude the report.