Safe Path Recommender: Using Crime Statistics

Bavin Ondieki and Chaitanya Patchava
Stanford University

Abstract

The purpose of this application is to use crime statistics in local areas and try to decipher that data to monitor for patterns. This application will monitor the "safety" factor of a certain area by reading the data that is pertinent to that particular area. Given the past history of different locations and their features, we used a classifier to classify if the path is safe for a safe walk or drive.

Introduction

In order to make suggested path routes useful for all users, we got a hold of real world data containing crime statistics, education levels, property value, and unemployment levels. The training data is the labeled data containing crime statistics for the area that we are to measure the safety of. This data has list of different features such as homicides, robberies, muggings, car theft, etc. we took into account real estate values of the areas that correspond to those crime statistics so as to make correlational assessments. Along with the crime, we got each crime and mapped it to a specific longitude and latitude. The training set involves data collected from a sizeable area, which effectively trains our Naïve Bayes classifier to be able to predict with a good degree of accuracy the safety of danger of a place. This is a proof of concept, but we intend to extend this so that it can serve more people over a greater area as opposed to the smaller region that was under consideration.

Once we did some feature extraction and consequently ran the classifier over the data points in the training set, we was able, with a good degree of accuracy, able to determine the how safe a path, from origin A to destination B, is. We will ask the user to enter where they are and where they wish to go, once they do this we will use our training data to avoid areas with high crime density and maneuver around in such a way that the user can get to their end destination in a safe manner.

Points with street crime figures above an arbitrarily chosen threshold are generally labeled as safe for a traveller. There is an interesting correlation between the features that we selected to use for the training and prediction, and in the event that we don't have information/ features of a particular point, the program does find the nearest points and approximate the features to be those of the region within half a mile radius of that particular latitude and longitude.

MOTIVATION:

Using the data that we collected, we will be able to use our training values to determine how "safe" a neighborhood is with respect to the neighborhood that the current homeowners are currently in. All they would need to do is put in their current address and their desired new home address. The algorithm will assign a "safe" rating to a neighborhood based on both its crime rates and other features with respect to the areas around it.

The application can be used to find safe paths to different things such as:

- Finding a good place to live in
- Finding best place to set up a learning institution or driving school
- Researching safe neighborhoods so you can invest in accessible neighborhoods

SAMPLE PLOT OF CRIME STATISTICS (ALAMEDA COUNTY) BY LATITUDE AND LONGITUDE

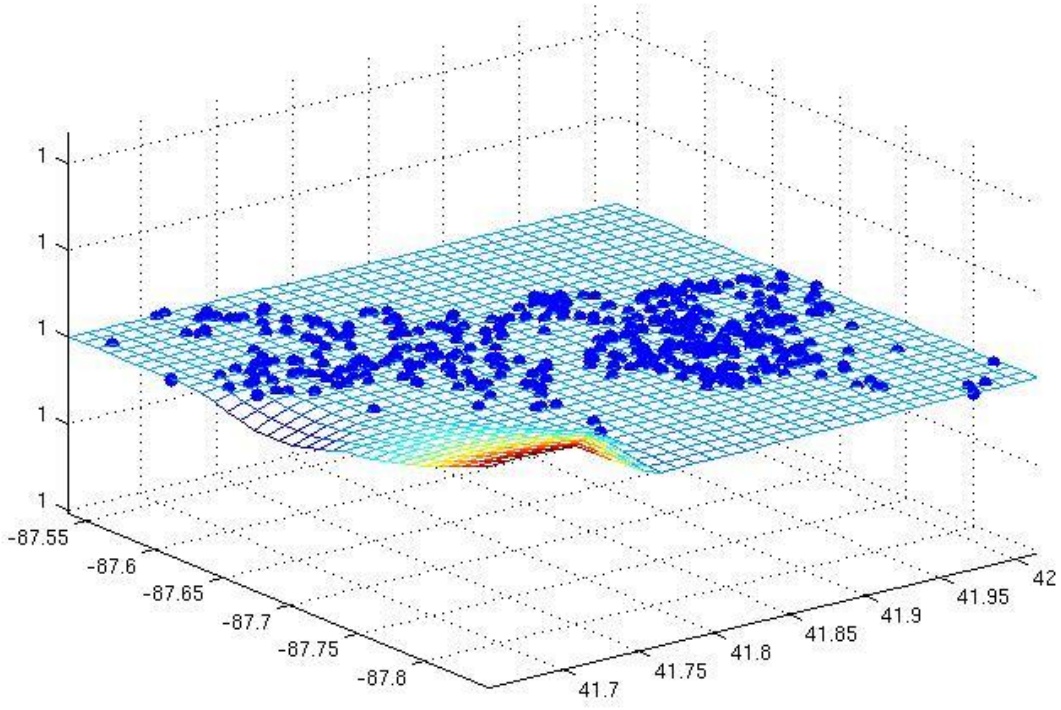


Figure 1: plotted longitude and latitude corresponding to crimes.

The data figures in the table above, as well as the figure only represent a small portion of the total set of data that we have collected through police databases. So far we have taken pieces of the data that are within a 5-mile radius of each other and grouped them onto plots like this.

Data

Primary Type	Location Description	Latitude	Longitude	Unemployed	Property Value	Education %	Street Crime
CRIMINAL DAMAGE	RESTAURANT	41.79424727	-87.62277244	1000	620500	95	0
THEFT	BAR OR TAVERN	41.94606093	-87.65577098	1200	575000	84	0
BATTERY	RESIDENCE	41.96051644	-87.71542316	1100	590000	57	0
CRIMINAL DAMAGE	STREET	41.89624808	-87.62534896	5000	55000	45	1
THEFT	STREET	41.79020103	-87.66463506	4100	88000	67	1
CRIMINAL DAMAGE	STREET	41.80024874	-87.60649936	6200	75000	65	1
CRIMINAL DAMAGE	STREET	41.92702258	-87.76746175	7800	53000	41	1
CRIMINAL DAMAGE	APARTMENT	41.9314564	-87.64794339	1000	366000	39	0
BATTERY	APARTMENT	41.86469418	-87.72743008	1900	402000	55	0
CRIMINAL DAMAGE	STREET	41.77906049	-87.74246219	4500	37500	66	1
PUBLIC PEACE VIOLATION	AIRPORT EXTERIOR - NON-SECURE AREA	41.78668044	-87.74602232	800	570000	46	0
BATTERY	RESTAURANT	41.70688511	-87.62011332	400	483500	93	0
THEFT	SIDEWALK	41.93086729	-87.73446662	290	1375810	88	0
BATTERY	RESIDENCE	41.7469	-87.59066622	850	635000	58	0
ASSAULT	SMALL RETAIL STORE	41.74992092	-87.68340589	740	455000	65	0
ASSAULT	STREET	41.79228377	-87.64646507	3746	82000	25	1
MOTOR VEHICLE THEFT	STREET	41.92246636	-87.7672991	4500	60500	68	1
CRIMINAL DAMAGE	RESIDENCE	41.7632585	-87.61172601	2500	410000	89	0
CRIMINAL DAMAGE	STREET	41.71686468	-87.65429027	2500	355000	95	0
THEFT	STREET	41.86943023	-87.68803807	4500	416500	77	1
CRIMINAL DAMAGE	STREET	41.74700259	-87.65102287	5450	394250	74	1
CRIMINAL DAMAGE	RESIDENCE	41.68853717	-87.61800948	4871	124089	59	1
THEFT	SMALL RETAIL STORE	41.93859389	-87.6497488	2100	505000	86	0
THEFT	RESIDENTIAL YARD (FRONT/BACK)	41.91198942	-87.73021237	1000	717500	87	0

THE MODEL

FEATURE EXTRACTION:

The main challenge behind this was the fact that we were unable to find lots of information regarding the crime statistics for each point under consideration. Of course, this would translate to each possible latitude and longitude location, which does not make much sense. Therefore, there was the need to round off some latitude and longitude values so as to have information regarding locations within a certain radius. This was accomplished by rounding off the latitude and longitude to the 3rd decimal place, which roughly translates to clustering all points within half a mile radius, and assigning those crime statistics as being part of that cluster.

This would be useful later on, so that, given a particular point, we would be able to query my map for features of a particular point, and if we don't' have the features of the particular point, then we would have the features of the points geographically close to it.

FEATURE SELECTION:

Unemployment Levels - {Low, Medium, High}

Street Crime Levels --{Low, Medium, High}

Property Value -- {Low, Medium, High}

Education Levels --{Low, Medium, High}

BUILDING THE NAÏVE BAYES CLASSIFIER and PREDICTION:

Using the above features, we were able to build a Naïve Bayes classifier that would then classify points as being safe of being unsafe. Since is practically infeasible to collect the data for each latitude and longitude combination, an approximation is used, thus yielding better results.

First of all, the Naïve Bayes model is built using the labeled training set, so that the user is able to query the model. User input is an origin and destination, with the origin being the start location and the destination being the final place in the path. The API (PyGeocoder) converts the physical address into latitude, longitude values that are manipulated to find the features for that location. The API also ensures that the address is well formed and correct.

After that, using a Python Google Maps API, the program sends a request so as to find a path from point A to Point B.

Given the concern that the path may not be safe, we classify each of the intermediate regions to find the relative safety of these regions. The result for each of the prediction is output alongside the directions from the Google Maps API.

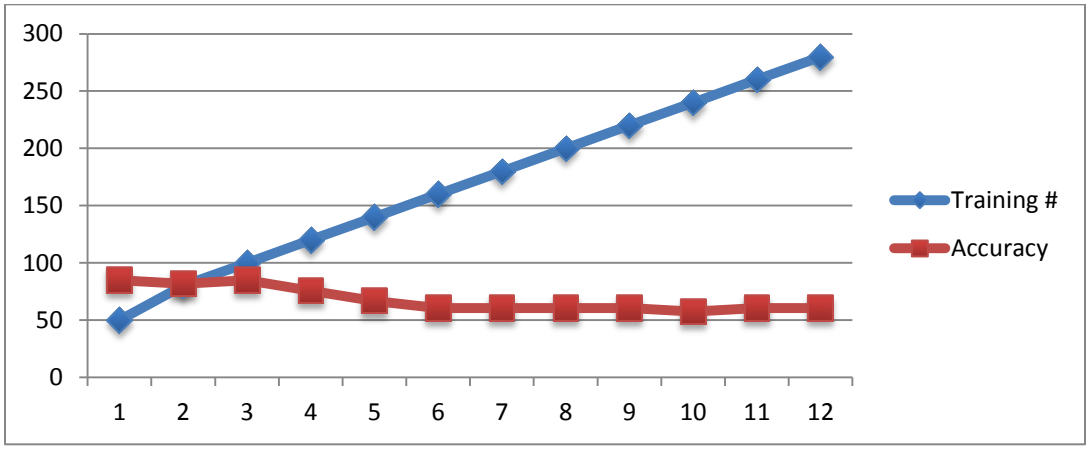
LAPLACE SMOOTHING

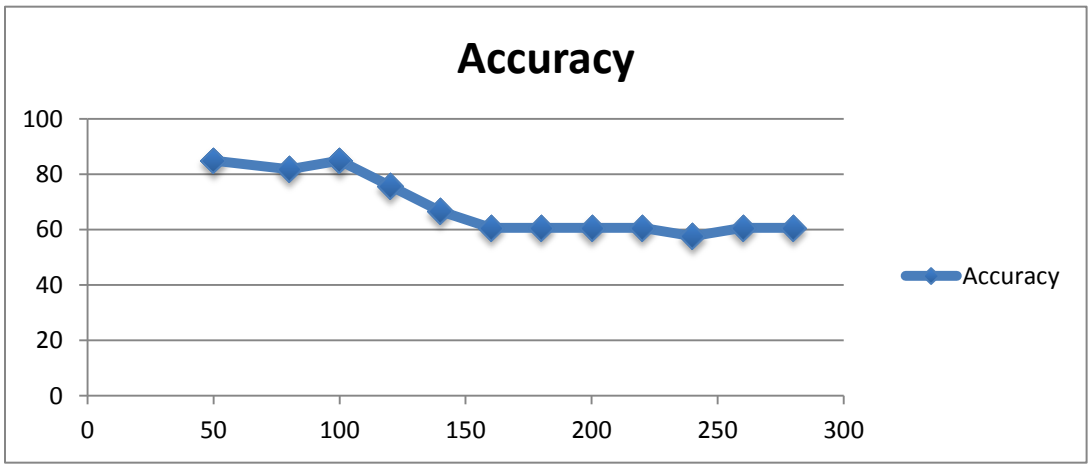
The classifier makes use of Laplace smoothing in the event that we have no information regarding a place. This is because, even though a place has never had any crime, it is possibly, but with minimal chances, that the place may experience some crime at some point.

ANOMALY DETECTION:

The crime statistics may vary quite a lot. For example, in the case that we have a large scale catastrophe, then it means that even though it's a single crime, the figures may indicate numerous incidents. This case may potentially result in classification as being unsafe, whereas that location is relatively safe. In this implementation, it is not considered as a special case as yet, since the only pertinent features include the recent crime statistics of different regions and not the crime history.

RESULTS AND ERROR ANALYSIS:





CHALLENGES:

The main challenge was an up to date source of ready to go data for all the different zip codes. Another problem that we faced was the utilization of the google maps API. It was tough to integrate the output of our learning algorithms into google maps like we had planned, so instead we just were able to label safe and unsafe paths and establish something close to what we wanted with the usage of waypoints.

Conclusion and Future Work:

The safe path recommender system yielded results that could be applicable for many things. Hopefully, it will end up as a useful companion to safe commuting, finding safe, accessible properties, as well as for personal uses such as scheduling which involves some travelling.