Prediction of Crimes' Category in San Francisco

Fernando M. Wittmann, 180577

***Abstract*—This report presents the prediction of crimes in San Francisco from 2003 to 2015. Three solutions are proposed. The first uses Log. Regression having 'Descript' as the only feature vector. The model predicted 94% of test samples. When using a dictionary, the prediction was 99.13%. Second solution uses Neural Networks and ignores the features 'Descript' and 'Resolution'. It predicted 20.01% of the test set. The third solution uses Log. Regression and engineered features. It predicted 22.4% of the test set.**

I. Introduction

Section II presents the required activities. Section III a data analysis. Section IV the solutions and Section V the conclusions.

II. Activities

This list has all the activities and where the answer can be found in this report:

1) Engineer your features.
   See Section III.A for discussion and IV.C for solution
2) Propose classification techniques to solve the problem.
   a) Log. Regression: **Sec. IV.A and IV.C.**
   b) Neural Networks: **Sec. IV.B.**
3) Are all of the features important to predict the kind of crime?
   Sec. III.B.
4) Is it possible to predict the kind of crime based on DayOfWeek, PdDistrict, Longitude and Latitude (location and time features)?
   Sec. III.C for discussion and IV.B for solution.
5) Report all results comparing to the real answers from the testing set.
   Sec. IV.
6) Questions
   a) What are the most common crimes?
      Sec. III.D.
   b) Which areas are more dangerous?
      Sec. III.E.
   c) Is there a time of day in which there are more crimes?
      Sec. III.F.
   d) What can we learn about the city from the Top Crimes Map?
      Sec. III.G.

III. Data Analysis

This section presents a data analysis in order to answer some questions and propose solutions.

Contact: fernando.wittmann@gmail.com

A. (Activity 1) Engineered Features

New features were created and evaluated. They are: Year, Month, Hour, Time of Day (Morning, Afternoon, Night and Dawn). These new features were evaluated with the existing features. After simulations on, the features that gave better solutions were District, Day of Week, Hour and Address.

B. (Act. 3) Features Importance

The most important feature is Descript. It is possible to predict the category with a very good precision just by using this feature. There are 866 uniques descripts and almost all of them are associated to an unique category (there are 39 categories). The less helpful feature is the 'Resolution' because it is too ambiguous to too many categories. The resolution 'NONE' corresponds to 60% of the train set and is present in almost all categories. The features X, Y are ambiguous with Address. The features can be classified in order of importance as: X, Y, Address, Hour (engineered feature), Year (engineered feature), PdDistrict and DayOfWeek. The Day of Week when used alone is not very important, however, using with the hour feature brings good insights (some crimes are very concentrated on Friday and Saturday night).

C. (Act. 4) Analisys of DayOfWeek, PdDistrict, X and Y

This subsection discusses the question: is it possible to predict the kind of crime only based on DayOfWeek, PdDistrict, Longitude, Latitude? The solution is provided on IV.B.

- DayOfWeek: The day of the week feature is not very important when used alone. **Fig. 1** in the Appendix has an analysis for each category. The difference between the more and the less frequent days with crimes was always less than 3% (avg. of 1.71%). However, if we combine the DayOfWeek with Hour then some interesting relationships can be observed.

- PdDistrict: The Police District feature is helpful because there are some districts that are more dangerous than others (see III.E).

- X and Y: These are the best features, after the Descript, because some of the crimes are highly concentrated in determined areas. Fig. 2 in the Appendix shows X, Y sets with very high frequency of crimes. It is observed, for example, that about 3% of all crimes (or aprox. 21000) happens in the coordinates -122.4034, 37.7754 which corresponds to 800 Block of Bryant St.

D. (Act. 6.a) Most Common Crimes

Fig. 3 shows the top ten crimes. They represents about 80% of all crimes. Larcerny (or Theft) is the most common crime, representing 21%. The second and third most frequent categories - Other Offenses and Non-Criminal - represents sets of

incidents. They can be seen in details on Fig. 4. Other offenses (14%) represents small infraction and are mostly associated to transit. Non-Criminal (11%) has lost and found property as the most frequent occurrence. The next most frequent categories are Assault (9%), Drugs (6%), and Vandalism (5%).

E. (Act. 6.b) Most Dangerous Areas[1]

Southern district is the most dangerous area, with 18% of crimes incidents. Mission and Northern were also high crime locations, reporting 13% and 12% incidents respectively. Park and Richmond are the safest locations with 6% and 5% of incidents only. Fig. 5 presents the frequency of crimes for each district. Larceny and Assaults were more common in the Southern district. Drugs were twice more common in Tenderlion than any other region. Vehicle Thefts are more frequent in Ingleside. 800 Block of Bryant Street is the most dangerous address in San Francisco, with 2.9% of all reported crimes.

F. (Act. 6.c) Hour Relationship

There are some range of hours where the crimes are more predominant. Fig. 6 shows an hour analysis for all crimes, larceny, assault and drugs. From 1AM to 7AM the crimes are less frequent. The peak is almost always in the rush time (18 PM), except the assault, that has peak at 0AM. Times when people are going to work, to lunch or going back home are also highly featured with incidents.

G. (Act. 6.d) Top Crimes Map analysis

Fig. 7 plots the crimes with more than 150 incidents in the same location. It is possible to observe that they are very clustered in the downtown of San Francisco.

IV. Solutions

The reported accuracies in these solutions are using the score function of skylearn (Python), which by default is normalized.

A. (Act. 2 and 5) With Description Feature

The first solution used Log. Regression descript as the only feature training vector. When comparing to real ground-truth of the test set, **94%** was correctly classified (normalized score). A more powerful and simple solution, using a dictionary, presented a prediction of **99.13%**. This dictionary was created by linking each descript to one class. In rare cases where the same descript was present in two classes, the class with more frequency was selected.
A bag of words was also used, however it gave slighly lower prediction values. This is due to the profile of this dataset that there is a high uniqueness of each pair of descript and category, making solutions that uses the whole descript better.

[1]The categories 'Other Offenses' and 'Non-Criminal' were omitted in this analysis (they represent small and not crimes).

B. (Act. 2, 4 and 5) With Location and Time Features

The second solution uses Neural Networks and ignores the features 'Descript' and 'Resolution'. Only 20% of the data was correctly predicted from the test set. The predicted model outputs only three categories: 'Larceny/Theft', 'Other Offenses' or 'Drug/Narcotic'. The category 'Larceny/Theft' had 50% of falses positives and 26% of false negatives. Which means that half of the set classified as Larceny was actually not. It also means that it failed in detecting a quarter of the Larcenies. The category 'Other Offenses' had 24% of false positive and 60% of false negatives. The same interpretation from Larceny can be given to it. Finally 'Drug/Narcotic' has 6% of false positives, which is a good value. However, it has 72% of false negatives, which means that it a very high number of classes as Drugs related.

C. (Act. 1, 2 and 5) With Engineered Features

This third solution evaluated the combination of different features using log. regression. The set of features that gave better results were District, DayOfWeek, Hour and Address. The normalized accuracy was 22.43%.

V. Conclusions

This work presented some strategies to classify the kind of crime that happened in San Francisco. Three solutions were presented. The first solution was very efficient, however in a real case, the Descript feature would not be available. The second result used Neural Networks and had only 20% when not using the 'Description' and 'Resolution' features. The third solution addressed to increase this prediction by engineering some new features.

References

[1] "San Francisco Crime Classification", kaggle.com/c/sf-crime
[2] "sckit-learn: Machine Learn in Python", scikit-learn.org
[3] "San Francisco Crime Visualization", goo.gl/Dg4gpI
[4] "Machine Learning to Predict San Francisco Crime", efavdb.com/predicting-san-francisco-crimes
[5] MO 444 Class Notes