Machine Learning Engineer Nanodegree

Capstone Proposal

Omar Ahmed Kassem August 11th,2018

San Francisco Crime Classification

Introduction

As in all cities, crime is a reality San Francisco: Everyone who lives in San Francisco seems to know someone whose car window has been smashed in, or whose bicycle was stolen within the past year or two. Even Prius' car batteries are apparently considered fair game by the city's diligent thieves. The challenge we tackle today involves attempting to guess the class of a crime committed within the city, given the time and location it took place. Such studies are representative of efforts by many police forces today: Using machine learning approaches, one can get an improved understanding of which crimes occur where and when in a city — this then allows for better, dynamic allocation of police resources.

Domain Background

Supervised learning is one of the most promising field in machine learning and used to achieve many predictions used nowadays even in business goals. Using machine learning techniques to predict the crimes area and category shows the power of technology in the field of safety. Supervised learning is used in many feild similar to this problem like Predicting Crime Using Time and Location Data in this paper or Weather Forecasting using the weather data of the past two days, which include the maximum temperature, minimum temperature, mean humidity, mean atmospheric pressure, and weather classification for each day in this paper. Also This problem was a challenge on Kaggle on this link and many papers provide solutions for this kind of problem like this one.

Problem Statement

This is a Multi-Class Classification problem given the time and location of the crime, We need to predict the category of crime that occurred.

Datasets and Inputs

To aid in the SF Crime Classification, Kaggle has provided about 12 years of crime reports from all over the city — a data set that is pretty interesting to comb through. We can find the dataset in this link. The dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set. Each record contains the following information:

- Dates: a timestamp of the moment that the crime occurred. It is in the following format: Y-m-d H:i:s. E.g.: 2015-05-13 23:53:00
- Category: the category of the crime (the target, it has 39 unique class). E.g.: WARRANTS
- Descript: a short description of the crime. E.g.: WARRANT ARREST
- DayOfWeek: the day of the week in which the crime occurred. E.g.: Wednesday

- PdDistrict: the district of the city where the crime was committed. E.g.: NORTHERN
- Resolution: short description of the crime resolution. E.g.: “ARREST, BOOKED”
- Address: the address where the crime was located. E.g.: OAK ST / LAGUNA ST
- X: latitude of the crime position. E.g.: -122.425891675136
- Y: longitude of the crime position. E.g.: 37.7745985956747

Solution Statement

Different algorithms will be used to come up with a good result. Each of them will be tried and tested, and finally we will get to see which of them works best for this case. Cross-validation will be used to validate the models, so the database has to be split into test, train and validation subsets. The resulting train dataset is still too large, and running the testing programs would take too long. To speed up tests and development, we will reduce the database using a clustering algorithm. This algorithm will be K-Means. Then, having the number of elements per cluster, we will be able to decide which element has more weight inside the algorithm. Technically there is no data loss. Once the data has been treated, the following algorithms will be tried:

- K-Nearest Neighbours
- Neural Networks

Benchmark Model

There are 3 types of results that the model can be compared to:

- comparing with the best kernels solving this problem in Kaggle competition for example:
  - careyai (score = 2.42381)
  - EDA and classification (score = 2.56180)
- Test the model with it self using several algorithms and comparing them with each other.
- Kaggle scores: Comparing the results with Kaggle public scores which is a great indicator for how our model performe.

Evaluation Metrics

Models will be evaluated using the multi-class logarithmic loss. Each incident has been labeled with one true class. For each incident,the model will provide a set of predicted probabilities (one for every class). The formula is then, $$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(P_{ij})$$ where N is the number of cases in the test set, M is the number of class labels, log is the natural logarithm, y_{ij} is 1 if observation i is in class j and 0 otherwise, and P_{ij} is the predicted probability that observation i belongs to class j.

Project Design

The project will be processed in steps:

- Data Analysis
  - Data reading
  - Data visualization – using maps, heat maps and contour plots to observe the crime distribution by different features.
- Data preprocessing

    - Relevant information
    - Data transformation
    - Dataset split
    - Dataset reduction (clustering)
- Models implementation
    - Implement many algorithms to compare them
- Models evaluation
    - Evaluate each algorithm and choose the best using the evaluation metrics.
- Conclusion and result