Regression and Classification tasks with Oakland Crime Statistics(2011-2016)

Jiarong Ye, Han Shao, Yichi Qian

1. Motivation for solving the problem.

To determine the standard of living of a nation, it is important to know the crime rate. The crime rate of a country will help to also determine the general welfare quality of citizens. Considering the futility of using crime statistics in the tracking of crimes, and the implementation of measures against them, for instance, we start an experiment about the crime statistics dataset which hosted by the city of Oakland in California. Our discussion is listed as follows.

2. How will the problem be solved/addressed?

Two possible problems regarding the motivation discussed above would be:

1. Crime Rate prediction in the city of Oakland (2011-2016). WIth the crime statistics across 6 years from 2011 to 2016 collected in the dataset we would be using, we could aggregate the number of crime occurrence of each month during these years of different priority (crime seriousness, whether it's a felony or not) and prepare new datasets to do a simple crime rate tendency estimation among different crime categories, accompanied with exploratory analysis (visualization). If the tendency appears linear, then the regression task could be simply tackled by linear regression algorithm, if not, the neural network is also a possible option.

2. Crime location classification in the city of Oakland (2011-2016). With certain attributes of crime occurrence known, for instance, Create and Closed Time, Beat, Incident Type, we could be able to encode the categorical data to construct a classifier, such as Decision Tree, Random Forest, etc. to predict the probability of occurrence of certain crimes of specific features and seriousness in certain location (the label in the algorithm)

3. Datasets that will be used to evaluate the solution. Remember to argue why your data is sufficient, i.e., representative of a broad sample of data over which the solution you proposed will be applied in real life and not an anecdote.

To solve the problem of crime rate estimation and location classification based on its severity, the dataset our group would be using is the open source dataset published and maintained by the city of Oakland regarding the Oakland crime statistics from 2011 to 2016 (kaggle dataset). The dataset contains attributes such as:

1. Agency
2. Create Time
3. Location
4. Area Id
5. Beat
6. Priority
7. Incident Type Id
8. Incident Type Description
9. Event Number
10. Closed Time

Considering that we are currently attempting to solve the regression of the crime rate in Oakland from 2011 to 2016 and the classification of certain crimes happened in different districts in Oakland, the published dataset hosted and maintained by the city of Oakland in California, thus the credibility guarantees the quality of data. Also, with the attributes listed above, the feature set of the Oakland crime dataset basically covers the essential elements required for a well-feature-engineered dataset for a decent regression or classification algorithm. And it collects numbers from authorities (police force) on a daily basis of the same feature set across 6 years, summing up to approximately **1 million** data entries. To summarize, the credibility, attributes and amount of the dataset indicate that it is representative.

4. **Metrics that will be used to evaluate the efficacy of the solution.**

For Regression:

- Mean Squared Error (or Root Mean Squared Error)
- Mean Absolute Error
- Mean Squared Log Error

For Classification:

- Accuracy Score
- Precision Score
- Recall Score
- F1 Score
- AUC (ROC)

5. Experiments that you will run.

- Data preprocessing
    - Transform the date under Create and Closed Time columns into the format that's MySQL-compatible, and then create tables for each year and insert the data into the MySQL database;
- Regression Algorithm to predict crime rate
    - Linear Regression
    - Polynomial regression
    - Neural Network approach
- Classification Algorithm to estimate crime location
    - Decision Tree
    - Ensemble
        - Boosting (AdaBoost, XGBoost, etc)
        - Bagging (Random Forest, etc)

6. Progress you have made in the last month.

We collect dataset for all attributes, including agency, create time, location, area id, beat, priority, incident type id, incident type description, event number and closed time. Thus we are able to analyze the whole statistic datasets about crime rate in Oakland systematically and particularly focus on two problems we addressed.

7. Schedule listing milestones for your progress

1. Finished collecting the data
2. Come up with a basic idea that solves the problem
3. Refine the solution and modify it if necessary.
4. Test the metrics in the data to evaluate the efficiency of our solution
5. Check if there is still any flaw in our management of data and the conduction of experiment
6. Finishing the Final report.