Predicting Crime by Category

Muhammad Ayub & Sean Fitzgerald

Introduction

We began with a dataset from the United Nations' Food and Agricultural Organization (FAO), hoping to predict child malnutrition based on each nation's exports, imports, food balance, and other aggregate crop statistics. Our goal was to promote global food balance (i.e. solve world hunger). While altruistic, the nation-level data available from the United Nations does not provide the subtlety necessary for nuanced predictions about communities within a nation. For instance, Dr. James Holland Jones provided us with the perspective that many things, from dyadic events to popular culture, significantly effect malnutrition and starvation. The FAO datasets capture none of this. Our baseline model reflected his recommendation, achieving a woeful <8% accuracy. Thus we pivoted.

We have settled upon a dataset that we believe strikes a balance between an interesting problem and achievable results. The Chicago Crime Dataset records all reported crimes in the city of Chicago from 2001 until 2016. Our goal: provided spatial, temporal, economical, and sociological information in a Chicago neighborhood, can we determine if the area is amenable to a certain type of crime? Police can leverage this knowledge to understand the likely threats and when patrolling a neighborhood. Since the City of Chicago has classified these into 35 categories of crime, this will be a classification problem. We will supply a neighborhood's features and intend to accurately predict the crime classifications likely to occur in that neighborhood. This data has been explored by others using CNNs and RNNs. Researchers have attempted to predict the location of future crimes. Reviewing the relevant literature, we believe that ours is a fundamentally different approach. However, we plan to eventually leverage some of the convolutional techniques from past research to better incorporate location-specific data.

The Approach

Model Selection

We have used % accuracy to compare our models. However, accuracy has some serious issues. Specifically, some of the 34 class labels are very low in relative quantity. Because of this, if the model is naturally biased toward the more frequent classes, the model will not truly perform well in practice. While % accuracy might be a good first pass to prune some of the earlier models we come up with, In the future we will use the F1 score as our optimization metric. In terms of satisficing metrics, we believe that certain crimes are more important/serious than others. We would like to penalize misclassifying more important classes in the loss function, but to keep the training procedure simple, we will try to bake this into our satisficing criteria: we can establish error bounds of misclassifications related to particular classes as satisficing metrics.

Some brief comments about the loss function, right now for our fully connected neural network, we have one hot encodings for the labeled data, and our loss function is the n-label categorical cross entropy. We have a Softmax layer with 34 possible outcomes. Now, this works for the time being, but if we go on to build a CNN, we think that we could group many classes of crime happening in one "pixel"/small geographic area in a picture. This implies that we have multi-hot encodings for which the cross entropy loss function is not optimized. CNNs are valuable when learning about geographical areas with multiple classes that can will be penalized more (Spring 2018 Midterm p14-15). We would have to output 34 logistic functions.

Data Processing:

The data engineering so far has for the most part involved adding and removing, as necessary, columns from the Chicago Crime Dataset. We have successfully joined two other datasets: Chicago Liquor Stores and salary statistics from the 2008 census. During the initial stages of modeling, we achieved an 8% increase in our model accuracy

from joining the salary data. Unfortunately, joining in the nearest liquor store to every crime was intractable (computing nearest liquor vendor out of 6000 for each of the million data points) and therefore we aggregated the number of liquor stores that are present in the district each crime occurs in. After a recent meeting with our TA, we would like to focus more on the geographical occurrence of the crime before moving onto joining other datasets.

Training Baselines

We chose 2 baselines to understand the complexity of the task. Because our problem is a classification task, we decided to use an SVM as our baseline as well as a relatively shallow neural network. Both of the models gave us an initial idea of where the Bayes Error lies for simpler models in this problem domain.

The SVM was trained with a very small subset of the data chiefly because it was choking our computer. We randomly sampled 30,000 data points for training and evaluated accuracy on 2000 data points. The library used was scikit-learn and the model used was SVC. We were able to achieve an accuracy of 34% on the 2000 test data points.

We also trained a SoftMax regression on 10% of our input data and we achieved 26% accuracy after training the model (minimizing categorical cross entropy). The Softmax model had zero hidden layers but set the baseline for increasing the complexity of our neural network.

Tuning/Training

In the initial stages of exploring, we were using Keras models to rapidly iterate for feature selection as well as layer amount and number of nodes.

Just recently, we have achieved a training accuracy of 80% while the dev set accuracy was in the 30% range. While we are still determining what the Bayes Proxy is for this problem and believe that we can lower the bias, we want to start on model regularization techniques. We are thinking of trying L1/L2 norm regularization. Ideally, we want to stay away from early stopping, because of the orthogonality concern (mentioned in the Coursera videos). We could also use Inverted DropOut, but want to stay away since we might be either shifting to a CNN or RNN architecture (https://www.reddit.com/r/MachineLearning/comments/5l3f1c/d_what_happened_to_dropout/ - Ogrisel's answer) and the cost function is no longer well defined. Some of the initial training technique's we have used include strict gradient descent, and Adam optimization.

Model Selection

In terms of comparing the different models we will come up with, the optimizing metric is the F1 score. Up until now, we have been evaluating the accuracy and while this might be a good first pass to prune some of the earlier models we come up with. Accuracy has some serious issues, especially because some of the 34 class labels are very low in number. Because of this, if the model is naturally very biased towards the more frequent classes, the model actually doesn't perform as well. Soon we will therefore be transitioning to F1 score. In terms of satisficing metrics, we believe that certain crimes are more important/serious than others. We would like to either penalize misclassifying more important classes in the loss function, but to keep the training procedure the simple, we will try to bake this into our satisficing criteria: we can establish error bounds of misclassifications related to particular classes as satisficing metrics.

Some brief comments about the loss function, right now for our fully connected neural network, we have one hot encodings for the labeled data, and our loss function is the n-label cross entropy. We have a SoftMax layer with 34 possible outcomes. Now, this works for the time being, but if we go on to build a CNN, we think that we could

group many classes of crime happening in one “pixel”/small geographic area in a picture. This implies that we have multi-hot encodings for which the cross entropy loss function doesn’t do as well because geographical areas with multiple classes will be penalized more (Spring 2018 Midterm p14-15). We would have to output 34 logistic functions.

Dataset Details

To accurately describe the raw dataset, the columns present are ['ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary Type', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Location']

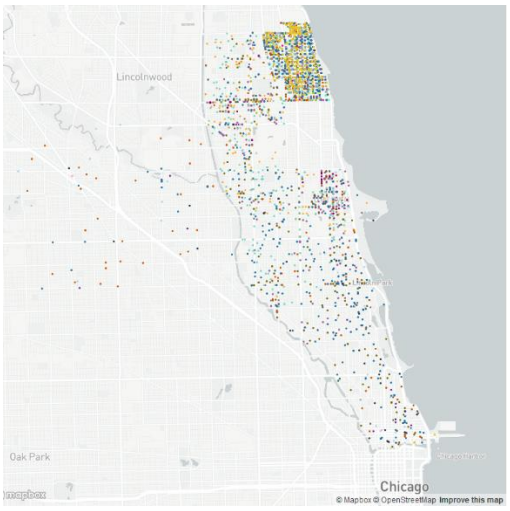
The ‘Primary Type’ is the column that consists of the 35 different types of crimes. These crimes include “BATTERY” , “NARCOTICS” , “ARSON”, “PUBLIC INDECENCY” , “HOMICIDE”, etc. . The dataset is fairly clean and requires minimal preprocessing (only normalizing the data and converting values to categories). We have also added other salary information and liquor store information from other Chicago city datasets. The links to the datasets are in the appendix.

Ethical Considerations and Next Steps/Improvements

We would like to mention that it is critical to be aware of the sociological assumptions we make when training the dataset. For example, we have included salary data because we as the modelers believe that salary/poverty has an influence on crime. Likewise, we decided that incorporating liquor store locations would also increase in prediction abilities. Sooner or later, while the assumptions we make about the causes of crime might be unbiased, certain ethnic/racial/religious/etc. groups might be stereotypically associated with some of our reasoned causes of crime. These are things to keep in mind while we train the model. If we had more time, we would more rigorously define loss function/evaluation metric strategies to address this concern.

While we think our fully connected network is capturing the latitude/longitude data, we do think that it might not be making as many of the meaning connections from the geographic data. That’s why we are thinking that converting to an image format might give us more ability to make spatial relationships in the environment.

Here is one of the data images we built with MapBox’s JS API: (maybe crop out the image) (Speak to how we would train CNN to answer our question). We think that incorporating such data can improve our models’ performance.



Appendix (not part of 3 page limit):

Datasets:

Crime data - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
Salary District data - https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2
Liquor store data - https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses-Current-Liquor-and-Public-Places/nrmj-3kcf

Github Repository:

Our Git repository is at : https://github.com/mrplants/crime_prediction .
Muhammad's Github Username is tuf22191
Sean's Github Username is mrplants