Project Proposal
CSCI-P556: Applied Machine Learning
Spring 2019

Bivas Maiti (bmaiti)
Virendra Wali (vwali)
Darshan Shinde(dshinde))

February 16, 2019

1 Problem Statement

In one or two paragraphs, explain the problem you are trying to solve in this project. You can talk about why is this problem important/interesting to you. You cannot use a project/problem you are working on in another class (such as data mining). Also, briefly mention which strategies you will use to solve this problem and how? For instance, supervised learning or unsupervised learning or reinforcement learning or a combination of either of there.

India, being a very diverse place, there is a diverse group of crimes that are being recorded. During the State and Central Annual Budgets, various amount of resources are allocated to try and prevent crimes in the country. We are trying to find a pattern in the crimes with respect to time and different socio-economic features of the places. We plan to use supervise learning and some reinforcement learning methods to predict the occurrence of different classes of crimes in the Indian States and districts.

We will try and predict different classes of crime in various Indian states according to time. We will take into account economic and educational features, as well as time series data of crimes in those times and will also consider the demographic of the regions from the census data.

2 Data

Describe the dataset you will use in this project. What is the domain of the data? Give it's characteristics. For instance, tell us how big the data is? How many samples does it have? How many features does it have? Are there are a lot of missing values? Is it imbalanced? How many categorical/numeric/binary/other features? How did you acquire the data? If you plan on collecting the data yourself, by crawling or merging two or more data sources, then you do not have to give exact numbers for the above questions (estimates should work).

Our dataset will be a combination of different datasets from the data repositories of the Indian Government. The different datasets we'll use are:

- Below are the crime related datasets we plan to use:
    1. District wise crimes committed against women(from 2001 to 2012)(< 1000 rows)

2. District Wise Complaints against police(from 2001 to 2012)(< 1000 rows)
3. Education of Juveniles arrested(from 2001 to 2012)(< 1000 rows)
4. District wise crime committed against ST(from 2001 to 2012)(< 1000 rows)
5. District wise crime committed against SC(from 2001 to 2012)(< 1000 rows)
6. District wise all IPC crimes committed(from 2001 to 2012)(< 1000 rows)

- Below are the non-crime related datasets we plan to use:

1. District wise GDP (from 2001 to 2012)
2. District wise primary and secondary schooling dataset(from 2001 to 2012)
3. Census data for district wise demographic(from 2001 to 2012)

3 Questions

This is the heart of your proposal. In this section you will explain the specific technical questions you will be addressing in the project. While you can certainly use a dataset which has been used in some competition such as Kaggle, you should remember that you have to go beyond acquiring a certain dataset with a predefined target variable and learning n different models and finding which works best. You are expected to go beyond that by using the dataset as a launchpad to answer interesting questions. You are expected to answer at least 3 questions using the data. The following things are a few instances of what can count as a question:

- If you creating your own dataset by either crawling it from the web or by combination of two or more data sources or by simulation, then it will count as a question.
- Designing experiments to study competing machine learning techniques and find out what works best on the given data. For instance, if you are doing sentiment analysis of online comments from very negative, negative, neutral to very positive, you could either formulate this is a multiclass classification problem or as a regression problem? Which one would you choose? Why? You can design experiments to find out.
- Designing experiments to study the traits of the data. For instance, you can design experiments to study how to separate noisy features from others? Is it possible to extract signal from them or is it better to just ignore them?
- Formulating new target variable from the given dataset. For instance, if you have data about customers of a bank and their loan repayment details. The original dataset/question could be to build a binary classifier which predicts whether a new person applying for loan is fraudulent or genuine. You could also use the same dataset (possibly by augmenting more data if available), to build a regression model which predicts how profitable a customer is to the bank in his/her lifetime.
- If the data is imbalanced, then do standard machine learning techniques work well or aren't of much use? What would you do in such case? Two standard methods would be either to use sampling or to change the cost function. You can design experiments to find out which works best on the given data.

4 Evaluation Criteria

You should clearly mention how will you evaluate your experiments. If there is existing work on the given dataset, you should compare your results with it. You are not expected to beat the state-of-the-art. This is not a competition. Given that this is a proposal, these don't have to make it to the final report, but do want to see what ideas you have on evaluation.

5 Timeline and Roles

Give a rough timeline of the work that needs to be done in the project. This should include approximate dates. Remember that you will have to give an update on the timeline during your first milestone (progress report) which will be due in a month.

Please also report how the work is being divided between team members. If one person is working on one question, then you should list who will work on what.