[UDACITY MACHINE LEARNING NANODEGREE]
SAN FRANCISCO CRIME CLASSIFICATION

July 28, 2018

1. Project Background

The project is mainly concentrated on ***supervised learning*** technique. As it's a very interesting and trending field, it's an opportunity to make use of algorithms and techniques learned through the course, to solve such a problem. The problem to be solved is about **Crime Classification**, which can be applied to spread more safety and security to towns, cities, regions and even countries. As performed before, ***supervised learning*** can be used for such a classification problem like what is done in *Charity donors* project.

2. Problem Statement

As it's a competition on Kaggle, so it's better to copy the problem statement here as it is. It's as follows: "From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. From Sunset to SOMA, and Marina to Excelsior, this competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, you must predict the category of crime that occurred.". You can find the competition here. As can be observed from the description, it's a *classification* problem, where the input is about the climate of the incident, as: the district, the crime category, the date of the crime... etc. The target variable is the crime category in the testing set.

3. Datasets and Inputs:

This dataset contains incidents derived from **SFPD** Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. In details, the figure below explains the data fields:

Data fields

- **Dates** - timestamp of the crime incident
- **Category** - category of the crime incident (only in train.csv). **This is the target variable you are going to predict.**
- **Descript** - detailed description of the crime incident (only in train.csv)
- **DayOfWeek** - the day of the week
- **PdDistrict** - name of the Police Department District
- **Resolution** - how the crime incident was resolved (only in train.csv)
- **Address** - the approximate street address of the crime incident
- **X** - Longitude
- **Y** - Latitude

Figure 1: Data fields

As shown, the target variable is the category of the crime.

The training dataset provides 878,000 data points with 9 attributes, while the testing one provide 848,000 ones with only 7 attributes. The training data must be split into training set and validation set. The validation set can be 20-25% of the whole training dataset size. In addition, it mentioned before that the problem provides the testing set explicitly.

4. Solution Statement

As mentioned earlier, the project is an application for ***supervised learning***. It means that the best way to make the most use of it is to apply many algorithms and techniques learned in that field, as: *Linear Regression*, *Decision Trees*, *Neural Networks*, *SVM (Support Vector Machine)*, *Ensemble B&B* … etc. Some or all those techniques may be implemented in the project and compare the results for reaching the best accuracy.

5. Benchmark Model

The easiest way to compare the solution, or measure the model selected is to show where it stands among those scores in the competition leaderboard on Kaggle. For more resources, we can consider a random version as a model to compare to. The solution must be more accurate than the random version with a reasonable rate to be regarded as a valid solution.

6. Evaluation Metrics

As it's a competition, it's reasonable to consider the same evaluation metrics of the competition itself to apply on the implementation of the project. It's evaluated using multi-class logarithmic loss. The formula is shown in the figure below.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

Figure 2: Multi-class logarithmic loss formula

7. Project Design

1. Data exploration: It includes reading, display and visualization.
2. Data preprocessing: We must deal with missing or invalid data (if exists). For example, assume that the missing field will take the value of the most common value for other data points. Also, we should deal with skewed features as we did in a project before when we consider the logarithm instead of the actual value to regularize the data, apply normalization and one-hot encoding, and split the data into: training and cross-validation sets.
3. Implementation: It mainly contains the application of the selected techniques and the comparison between their results. For instance, I can use *Linear Regression*, *Decision Trees*, *Neural Networks*, *SVM (Support Vector Machine)* and *Ensemble B&B*.