**TIM 245 Project ProposalSan Francisco Crime Classification**

Borui Wang, Han Chen

1. Basic idea: Apply data mining to an interesting data set

Our data is from Kaggle (<https://www.kaggle.com/c/sf-crime>), which are derived from SFPD Crime Incident Reporting system and provide train data and test data that 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.

We’ve already looked at the data which has 878,049 instances and 9 attributes for train set and 884,262 instances and 7 attributes for test set and there is no missing value. So what we need to do for data preprocessing is to scale data, and generate features. Also, we need to get better understanding of our data so that we can pick the right feature to classify the crime.

In the training data, it provides location of crime, so we might show the result on the maps using map data from the website (sf\_map\_copyright\_openstreetmap\_contributors.txt).

It is actually clustering problem, so we may use k-mean and some other algorithm.

Because we already have complete, large data set, so we think the most changeling aspects will be how to show the result more intuitive.

1. Time line

At the end of February, we need to finish data processing and try k-mean algorithm to see the result.

At March, we need to find different way to show the result, using different figures and get a conclusion of our project.

1. Member contribute

Both of us will work together, because we are not familiar with clustering now. Maybe in the future we will split the work load. If so, we will mention in the final report. We’ve already create the github repository, <https://github.com/boruiwang/SanFranciscoCrimeClassification>.