Raj Ramrup

Data: NYPD Complaint Data 2006- October 2017

Goal: To find the risk of being at a certain place and time.

Machine learning problem:

If a crime is to occur at a certain time and place, what type of crime will it be (assault, robbery, etc.)

Given: Date, Time, Location

Predict: Crime Type (Classification)

Limitations: False complaints, crimes that go unreported

Cleaning Features

- Take the features from the dataset
- Divide them into parts
  - Date → weekday, year, month, day, is_month_start, is_month_end, is_quarter_start, is_quarter_end, is_year_start, is_year_end
  - Time → hour, minute, second, *is_business_hours*
  - Latitude, Longitude →

    x = np.cos(df['Latitude']) * np.cos(df['Longitude']))

    y = np.cos(df['Latitude']) * np.sin(df['Longitude']))

    z = np.sin(df['Latitude']))
  - Typos:
    - Things like ‘1017’ as a year

Data

- Data for 11 years: 5 million rows
- Data for this year: 300,000 rows

- SVM classification
  - ~20 Features
  - ~80 Classes
  - 50/50 split between training and test

Results?

- Inconclusive…

KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)

- Currently run on 20% of only this years data. Bad accuracy.
- Attempted to run on whole dataset (with bad feature) but that took over 14 hours and never finished
- If run on the whole dataset:
  - Monthly trends
  - Seasonal trends
  - More data on area trends

Uses

- Take the locations of subway terminals and various times

- → see which train lines are more dangerous
- → find the least dangerous route home at 3 am