Mining Uber Dataset

Abhilash Mysore Somashekar
Raviraj Prakash Wani
Satya Akhil Chowdary Kuchipudi
Sanil Jain

1 April 2017

Project Github Link: https://github.com/msabhi/mining-uber-dataset

Objective:

The dataset collected by FiveThirtyEight contains data on over 4.5 million Uber pickups in New York City from April to September 2014. The data set is divided across six files i.e. one file per month

Exploratory Analysis Goals:

- Preprocessing of data (rounding of the Latitude and Longitudes)
- To find out whether 4.5 million data points over 6 months are continuous or not
- To plot counts in histograms as well as plot Uber pickups on actual map and look at thecoverage of NYC
- Analyze Uber pickup data over various months
- Finding the hotspot locations in the data. These are the locations where there are pickups more than a specified threshold

Kernel

$$K_{nm} = 1/\lambda \left(\phi\phi^T\right)$$

As per Linear Rigression ::

$$\theta = (XX^T)^{-1}X^TY$$

Hence for Kernel Ridge Regression

$$E\left[w/y\right] = A^{-1}\phi^TY$$

$$A = (\phi^T\phi + \lambda I)$$

$$\phi = \phi\left(X\right)$$

$$K = 1/\lambda \left(\phi\phi^T\right)$$

$$A^{-1}\phi^T = \phi^T \left(K + \lambda I^{-1}\right)$$

$$E\left[f\left(X\star\right)/y\right] = \phi\left(x\star\right)^T E\left[w/y\right]$$