CMSC 25025 / STAT 37601

# Machine Learning and Large Scale Data Analysis

LSD PROJECT 3

Due: Tuesday, May 21, 2013

This is the third of four large scale data projects. In this project, you will work with the Chicago crime dataset, building a point-process of several different crime types.

A Poisson process can be used to model the number of occurrences of events within a given time interval. In this project, we will use the Poisson process to model crime. Based on a dataset of past crimes, the goal is to make predictions about future crime rates.

We have divided the city of Chicago into a set of non-overlapping spatial regions, $\{\Delta(s_i)\}_{i=1,\ldots,M}$, with the center of the regions being $s_i \in \mathbb{R}^2$. For each region, we count the number of $d$ different types of crimes that occur in a given week. These counts are denoted $v_{i,t}$, where $v_{i,t}$ is a $d$ dimensional vector whose $k$-th component, $v_{i,t}^{(k)}$, indicates the number of crimes of type $k$ occurring in region $\Delta(s_i)$ during week $t$. Under a Poisson process, each $v_{i,t}^{(k)}$ is a random variable which follows a Poisson distribution with crime rate intensity parameter $\lambda_{i,t}^{(k)}$.

The project has two parts. In the first part, you will implement a kernel smoothing method to estimate $\lambda_{i,t}^{(k)}$ for the week following the last week in the data. In the second part, you can adopt your own approach.

*The Crime Dataset*

The area of Chicago is partitioned into $N = 2{,}985$ square regions, indexed from 0 to $N-1$, with centers $s_i$. The centers are given in the form of key-value pairs: $\{i : (\text{lng}_i, \text{lat}_i)\}$, where $\text{lng}_i$ and $\text{lat}_i$ are the longitude and latitude. You will build models for ten different crime types:

```
BATTERY, BURGLARY, CRIMINAL DAMAGE, WEAPONS VIOLATION, CRIMINAL TRESPASS,
THEFT, NARCOTICS, PROSTITUTION, MOTOR VEHICLE THEFT, DECEPTIVE PRACTICE.
```

Each of these types may have a different geographical or temporal profile. The crime types are given in the form of key-value pairs: $\{k : \text{crime}_k\}$, where $\text{crime}_k$ is the type of crime index $k$.

The dataset contains records of crimes from January 1, 2001 until April 27, 2013. We use one week as the unit time in this project, and count number of crimes occurring in a unit of time within each region. The interval Jan 1, 2001 12:00:00 to Jan 7, 2001 12:00:00 is considered week 0. Note that we are using noon as boundary between weeks. The data in the final week is not complete.

Each of the crime records in the dataset contains information including (but not limited to) the type of the crime, the date, time and the location of a crime. For the first part of the project, you will be

given $\{v_{i,t}^{(k)}\}_{i=0,1,...,2984;\, t=0,1,...,642;\, k=0,1,...,9}$ in the form of key-value pairs:

$$\left\{(i,k) : [v_{i,0}^{(k)}, v_{i,1}^{(k)}, ..., v_{i,T-1}^{(k)}]\right\}$$

where each key is a tuple of indices, and each value is an array with length $T = 643$. The value $v_{i,t}^{(k)}$ is the number of type $k$ crimes that were committed in region $\Delta(s_i)$ during week $t$.

Access to the data looks something like this:

```
#!/usr/bin/python
import DAL
crime = DAL.create('crime')
crime_list = crime.get_crime_list()
# the following can be slow; do this once at the beginning
# of the program and use this data structure throughout
crime_counts = crime.get_crime_counts()
region_list = crime.get_region_list()
```

In each of the two parts below, you are asked to build a model to predict the number of crime events in each region during week 643 (April 29, 2013 12:00:00 to May 6, 2013 12:00) ~~and week 644 (May 6, 2013 12:00:00 to May 13, 2013 12:00:00)~~. We will evaluate your model by calculating the log-likelihood on the actual crime counts collected during those periods:

$$\ell(t') = \sum_{i,k} \log \mathbb{P}\big(V_{i,t'}^{(k)} = v_{i,t'}^{(k)}\big).$$

Note: We will make available Python/Javascript utilities to generate Google maps to show crime locations and heatmaps.

1. *Kernel Smoothing* (60 points)

   In this part you will implement kernel smoothing to estimate a Poisson rate function $\lambda_{i,T}^{(k)}$ for $T = 643$ (corresponding to the week of April 29 to May 6, 2013). To apply kernel smoothing, we first need to choose a kernel function $K_h(x, y)$. Here are the commonly used kernel functions:

   - Gaussian kernel $K_h(x,y) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{\|x-y\|^2}{2h^2}}$
   - Epanechnikov kernel $K_h(x,y) = \frac{3}{4}(1 - \|\frac{x-y}{h}\|^2)\mathbf{1}_{\{\|x-y\| \le h\}}$
   - Boxcar kernel $K_h(x,y) = \frac{1}{2h}\mathbf{1}_{\|x-y\| \le h}$

   Then, the rate $\lambda_{i,T}^{(k)}$ can be estimated as the weighted summation of $v_{i,s}^{(k)}$:

$$\widehat{\lambda}_{i,T}^{(k)} = \frac{\sum_{i',s<T} K_\gamma(s,T) K_\sigma(s_i, s_{i'}) v_{i',s}^{(k)}}{\sum_{i',s<T} K_\gamma(s,T) K_\sigma(s_i, s_{i'})}$$

where $\gamma$ and $\sigma$ are bandwidth parameters in and space, respectively. You should try different kernels and different values for $\gamma$ and $\sigma$, and select those that produce high log-likelihood on test data.

For each type of crime, report your estimated rate for week 643 in a file. Save a pair of values $i, \widehat{\lambda}_{i,T}^{(k)}$ in each line of the file. Name the files `lsd_proj3_prob1_[crime-type].txt`. Replace any space in [crime-type] with a hyphen; e.g., the estimate of crime rate of `THEFT` should be saved in `lsd_proj3_prob1_theft.txt`, while estimation of `CRIMINAL DAMAGE` should be saved in `lsd_proj3_prob1_criminal-damage.txt`.

A third of the points (20 points) will be assigned based on your predictive log-likelihood, scored relative to your classmates' models.

2. *A Better Model?* (40 points)

For this part, you may design and build your own predictive model. You are free to use any approach that you think is effective, and you are not restricted to a Poisson process.

To report your predictions for each type of crime, in each line of the file, write a list of values

$$i, F_{i,T}^{(k)}(0), F_{i,T}^{(k)}(1), \ldots, F_{i,T}^{(k)}(20)$$

In this output $i$ is the index of a region, as for Problem 1. However, now $F_{i,T}^{(k)}(v)$ is your estimate of the cumulative distribution function:

$$F_{i,T}^{(k)}(v) = \mathbb{P}\big(V_{i,T}^{(k)} \leq v\big).$$

These values will be used to compute log-likelihoods.

Name your files `lsd_proj3_prob2_[crime-type].txt`.

For this problem, half of the points (20 points) will be assigned based on your predictive log-likelihood, again scored relative to your classmates' models.