CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

Assignment 6

Due: Thursday, May 21, 2013

This assignment consists of three problems. Two are "pencil and paper" problems and one involves working with the city of Chicago crime data.

1. *Task-task* (25 points)

   A computer system carries out tasks submitted by two users. Time is divided into slots. A slot can be idle, with probability $p_I = 1/6$, and busy with probability $p_B = 5/6$. During a busy slot, there is probability $\mathbb{P}(1 \,|\, B) = 2/5$ (respectively, $\mathbb{P}(2 \,|\, B = 3/5)$) that a task from user 1 (respectively, 2) is executed. We assume that events related to different slots are independent.

   (a) Find the probability that a task from user 1 is executed for the first time during the 4th slot.

   (b) Given that exactly 5 out of the first 10 slots were idle, find the probability that the 6th idle slot is slot 12.

   (c) Find the expected number of slots up to and including the 5th task from user 1.

   (d) Find the PMF, mean, and variance of the number of tasks from user 2 until the time of the 5th task from user 1 .

2. *Let there be light* (25 points)

   Beginning at time $t = 0$ we start using light bulbs, one at a time, to illuminate a room. Bulbs are replaced immediately upon failure. Each new bulb is selected independently by an equally likely choice between a type-A bulb and a type-B bulb. The lifetime X of any particular bulb is a random variable, independent of everything else, with the following PDF:

   $$\text{for type-A bulbs:} \quad f_X(x) = \begin{cases} \exp(-x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

   $$\text{for type-B bulbs:} \quad f_X(x) = \begin{cases} 3\exp(-3x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find the expected time until the first failure.

   (b) Given that there are no failures until time t, determine the conditional probability that the first bulb used is a type-A bulb.

   (c) Find the variance of the time until the first bulb failure.

(d) Suppose the process terminates as soon as a total of exactly 12 bulb failures have occurred. Determine the expected value and variance of the total period of illumination provided by type-B bulbs while the process is in operation.

(e) Given that there are no failures until time t, find the expected value of the time until the first failure.

3. *CSI* (50 points)

For this question you will use the data for LSD project 3 as well as the raw crime data. You should first sort the grids provided to you in LSD Project 3 in terms of the total number of crimes that occurred over the 642 weeks. Select the top 10 crime centers in terms of the number of crimes committed. The following code will help you access the raw crime data.

```
#Function to calculate the week number corresponding to a time stamp
def convertToWeekNumber( timestring ):
     d = datetime.datetime.strptime(timestring,"%Y-%m-%dT%H:%M:%S")
     first_day = datetime.datetime.strptime("2001-01-01T12:00:00",
                                              "%Y-%m-%dT%H:%M:%S")
     idx = int((d - first_day).days / 7)
     return idx

 import DAL
 DAL.cleancache()
 crime = DAL.create(’crime’)
 meta = crime.metadata()
 subsets = crime.subsets()
 subsets.remove(’set0.meta.json’)
 subsets.remove(’set1.meta.json’)
 subsets.remove(’crime_counts.zip’)
 subsets.remove(’crime_list.zip’)
 subsets.remove(’region_list.zip’)

 s = subsets[2:]
 for i in crime.iter(s[0]):
       print i[25], i[26], i[10], i[13]
       # Calculating the week of a timestamp
       try :
             timestring = datum[10]
             # convert a timestring to week number
             d_idx = convertToWeekNumber( timestring )
       except :
             pass
```

Note three things: You should start from `subsets[2]` (which corresponds to set1.data1.zip) onwards as the first two entries in `subsets` do not have the data in the correct format. If that does not include any of your 10 chosen crime centers, you should add crime centers from the sorted list so that you have exactly 10 for which the data is available from `subsets[2]` onwards.

This dataset contains records from January 1, 2001 until 31st December 2012. You should use one week as the unit of time for this problem, and count the number of crimes occurring in a unit of time within each region. Ignore all entries which have an entry `None` for any of the four fields.

Note that $i[25], i[26], i[10], i[13]$ give you the latitude, longitude, time and type of the crime at a location. The code within the `try` block helps you calculate the week number of the crime where the interval Jan 1, 2001 12:00:00 to Jan 7, 2001 12:00:00 is considered week 0.

`set1.data1.zip` is more recent data in terms of time than `set1.data2.zip` and so on through `set1.data17.zip`.

We will only be looking at crimes of the types `BATTERY, BURGLARY, CRIMINAL DAMAGE, WEAPONS VIOLATION, CRIMINAL TRESPASS`. At each of the 10 top locations $\{s_i\}_{i=1}^{100}$ that you obtain, consider a ball $R_\delta(s_i)$ of radius δ centered at s_i. Start with a value of $\delta = 0.005$ degrees. Find the total number of crimes for each of the types mentioned over one week intervals (starting and ending at noon of every week) within this ball. You should calculate the Euclidean distance of each location to the centers s_i to figure out if the location lies within a ball under consideration.

(a) Let $N_{k,\delta,s}(\tau)$ be a random variable denoting the total number of crimes of type k occurring in a radius of δ around the center s over the time interval τ. Plot the empirical distribution of $N_{k,\delta,s}(\tau)$ over the different weeks for each of the centers. Does $N_{k,\delta,s}(\tau)$ satisfy a Poisson distribution? What is the rate?

(b) Do the rates of the crimes scale approximately linearly as you increase τ? Try increasing τ to be 2 or 3 weeks and look at the distribution of $N_{k,\delta,s}(\tau)$. You don't have to submit all the plots, but comment on the behavior of the distribution as you change τ. What happens if you increase or decrease δ? Try different values of δ from 0.2 degrees to 0.001 degrees. Once you have an approximate rate of occurrence of a particular crime, check if the "interarrival times" between crimes roughly follows an exponential distribution with the appropriate rate.

(c) (Optional) Suppose we merge two crime types (say `BATTERY` and `BURGLARY` or `CRIMINAL DAMAGE` and `CRIMINAL TRESPASS`. Is the number of combined crimes for the two types again an (approximate) Poisson process? Is the new rate close to the sum of the original rates? Comment on your observations from the data.