CS 591 Data Analytics: Theory and Applications
Class Project

Last updated: 2/10/2017

Instructions The first part of the project consists of the first 6 questions. The second part is Question 7, where you choose your own project. The due date for the **milestone** is *March 13rd, beginning of the class.* You should deliver Problems 1, 2, and 7.1. The **due date for the rest** is *April 30th, beginning of the class.*

All answers should be written in LaTeX. Do *not* attach your code to the writeup. Instead, email your implementation to ctsourak@bu.edu with the title CS591-Project. Each file should be named by using your BU id as its prefix. Refer to the webpage for policies regarding collaboration, due dates, and extensions.

1 Probability, Information Theory [30 pt]

1.1 Isaac Newton Helps Samuel Pepys [3 pt]

Samuel Pepys wrote Isaac Newton a long letter asking him to determine the probabilities for a set of dice rolls related to a wager he planned to make. Pepys asked which was more likely:

- At least one six when six dice are rolled.
- At least two sixes when 12 dice are rolled.
- At least three sixes when 18 dice are rolled.

Your tasks are the following:

1. [1 pt] Try to answer Pepy's question using your intuition first. Now verify if your intuition was right.
2. [2 pt] Solve the general version of the problem:
   What is the probability of obtaining at least n 6 when $6n$ dice are thrown?

1.2 Sensitive questions [4 pt]

My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability $\frac{1}{2}$. The other possibilitiestwo boys or two girlshave probabilities $\frac{1}{4}$ and $\frac{1}{4}$.

1. [2 pt] Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?
2. [2 pt] Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

1.3 Bayes' rule [4 pt]

On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other two doors, and he will do so in such a way as not to reveal the prize. For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially, the prize is equally likely to be behind any of the 3 doors.

1.4 Reading Boston Globe [8 pt]

Read the related article at Boston Globe.

1. [2 pt] Translate the article into math. Specifically, define appropriate events, and state the probabilities mentioned implicitly in the article.
2. [6 pt] Please explain why the deduction about the causation of breast cancer is very problematic.

1.5 Entropy [8 pt]

1. [1 pt] Prove that the mutual information can be expressed in terms of entropies as follows

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

2. [3 pt] Prove that the KL divergence is non-negative. When is it equal to 0?
3. [4 pt] Let $p_{\text{emp}}(x)$ be the empirical distribution, and let $q(x|\theta)$ be some model. Show that $\arg\min KL(p_{\text{emp}}||q)$, is obtained by $q(x) = q(x;\bar{\theta})$, where $\bar{\theta}$ is the MLE.

1.6 Mixture of Gaussians [3 pt]

Consider a mixture of K Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k).$$

Show that if $x \sim p(x)$ then,

1. [1 pt] $\mathbb{E}[x] = \sum_k \pi_k \mu_k$, and
2. [2 pt] $\mathbb{C}ov[x] = \sum_k \pi_k(\Sigma_k + \mu_k \mu_k^T) - \mathbb{E}[x]\,\mathbb{E}[x]^T$.

2 Hashing [40 pt.]

2.1 2-wise independent hash function [5 pt]

Prove that if p is prime, a, b are chosen uniformly at random from $[p] = \{0, \ldots, p-1\}$, then the following hash function is 2-wise independent.

$$h(x) = ax + b \mod p.$$

2.2 Searching for patterns in fMRI [10 pt]

Functional magnetic resonance imaging (fMRI) is a technique widely used to measure brain activity by detecting changes associated with blood flow. Neuroscientist Dr. X has asked for your help on a problem she faces. Each fMRI dataset is modeled as a three dimensional array $M^{n \times n \times k}$., whose each entry is a real number. You can think of array M as a set of k slices, each being an $n \times n$ matrix $\{A_1, \ldots, A_k\}$.

Dr. X wants to search in each fMRI slice $A_i, i \in [k]$ dataset for a specific pattern related to a brain disease. This pattern an array $P^{r \times r}$ where $r \ll n$.

1. [1 pt] Describe a naive algorithm that searches for pattern P in each $r \times r$ subcube of a slice A_i. Discuss its pre-processing time, its matching time, and its space complexity. Your algorithm should return the number of times that P appears within M.

2. [6 pt] Describe an algorithm that uses an appropriately chosen hash function whose average case performance beats the naive algorithm. Your algorithm should return the number of times that P appears within M. Provide an analysis of your algorithm (preprocessing time, matching time, and space complexity).

3. [3 pt] Provide a working implementation of (i),(ii). Report the run times, and the number of occurences of pattern

$$P = \begin{bmatrix} H & H \\ H & L \end{bmatrix}$$

in the fMRI slice provided here.

2.3 Hashing strings [10 pt]

Download the following dataset of all english words from here. You will use the following functions to hash all strings, after you convert each one of them to all lowercase.

- djb2
- sdbm
- lose lose
- Murmur Hash 3

Report for each function, report

1. [5 pt] The number of resulting collisions, and the run time.

2. [5 pt] Submit your code as a single file.

2.4 Document Similarity [15 pt]

Download the following documents 1.txt,2.txt,3.txt,4.txt.

1. [3 pt] Create 1-grams, and 2-grams. Recall, each k-gram appears once, duplicates are ignored.
2. [4 pt] Compute *exactly* Jaccard similarity between all pairs for each type of of k-gram. Report all values.
3. [4 pt] Compute *approximately* Jaccard similarity between all pairs for each type of of k-gram. Use a varying number of hash functions, (5,10,20). Report all values.
4. Submit your code as a single file.

3 Social Networks [15 pt +10 extra pt]

Download the Epinions network.

1. [2 pt] Plot the degree distribution of the graph.
2. [3 pt] Again, plot the degree distribution but in log-log scale. What do you observe?
3. [10 pt] For each node v, compute the Closeness Centrality

$$C(v) = \frac{1}{\sum_u dist(u,v)}.$$

   Provide a scatter plot (x-axis node id, y-axis closeness centrality). Report the run time of your code.
4. Submit your code as a single file.
5. [Extra 10pt.] Download the All distance sketch implementation, and use it to get an approximate value $\bar{C}(v)$ for the closeness centrality. Provide a plot that plots the difference $C(v) - \bar{C}(v)$ versus node id. Report average error over n nodes, and the run time.

4 Machine Learning [65 pt]

4.1 Optimization [20 pt]

1. [2 pt.] Generate $n = 2\,000$ points uniformly at random in the two-dimensional unit square. Which point do you expect the centroid to be?
2. [4 pt.] What objective does the centroid of the points optimize?
3. [5 pt.] Apply gradient descent to find the centroid.
4. [9 pt.] Apply stochastic gradient descent to find the centroid. Can you say in simple words, what the algorithm is doing?

4.2 Regression [20 pt]

Download the Boston housing dataset. This dataset has 506 examples, with 13 attributes variables, and 1 continuous response variable, representing the median price of a house in Boston, in 1 000s of dollars. You can find the detailed description here.

1. [15 pt] Perform linear regression on a 10-fold cross validation. Report for each split the coefficient of determination R^2. What can you deduce?

4.3 Eigenfaces [15 pt]

Download file faces.cvs. It contains 360 lines, each containing data for a 112× 92 face image. If you are using MATLAB, you can load the images using

```
data = csvread(Xtrain.csv);
colormap(gray);
imagesc(reshape(data(i,:),112,92));
```

1. [3 pt] Compute the centroid of the 360 faces in Xtrain.cvs. Subtract it from each training point. Submit a visualization of the centroid.
2. [6 pt.] Compute the covariance matrix for the 360 faces.
3. [6 pt.] Display the eigenvectors corresponding to the 10 largest eigenvalues as 112× 92 images.

4.4 Handwritten Digits [10 pt]

Download the MNIST dataset.

1. [2 pt] Run mnist1NNdemo. What is the test error on the first 1 000 test cases?
2. [3 pt] Permute the features (columns of training and test design matrices) as in shuffledDigitsDemo. Did the error rate change?
3. [5 pt] Use the Matlab/C++ code FLANN to perform approximate nearest neighbor search, and combine it with mnist1NNdemo to classify the MNIST data. How much speedup do you get, and what is the drop (if any) in accuracy.

5 Philosophy and Data Analytics [5 extra points]

"And all knowledge, when separated from justice and virtue, is seen to be cunning and not wisdom."
Plato's Republic

Write a short essay (at most 500 words) discussing the importance of philosophy in data analytics.

6 Sed, Spark, Hadoop, Tensorflow [15 extra points]

1. [1 pt.] Get yourselves familiar with regular expressions and sed.
2. [5 pt.] Install tensorflow, and complete the codelab for Tensorflow.
3. [2 pt.] Install Spark, and run the Spark Word count example
4. [2 pt.] Install Hadoop, and run the following example Hadoop word count.
5. [] Summarize what you learnt.

7 Individual Project [110 pt]

7.1 Milestone

Please submit a report, stating your learning goals, the problem you will focus on, methods and datasets you will explore, and preliminary experiments. In case there are two or more people, describe which tasks will each person perform.

7.2 Report, Code, and Project Presentation