The University of Texas at Austin
Department of Electrical and Computer Engineering

460J: Data Science Lab — Fall 2019

Lab One

Caramanis/Dimakis Due: Tuesday 9/10/19, 3:00 pm.

Comments/Remarks: Submit one report for each lab group of three people. The report should include a pdf of your code and your results (plots, or output, as applicable), and also any discussion, again as applicable. Also submit all your code. For this lab, you can submit in either .ipynb format or .py format. If you choose to submit .py files, submit them in the format problemX.py or if you need, problemXa.py, problemXb.py, and so on.

Programming Questions

1. Create 1000 samples from a Gaussian distribution with mean -10 and standard deviation 5. Create another 1000 samples from another independent Gaussian with mean 10 and standard deviation 5.

   (a) Take the sum of 2 these Gaussians by adding the two sets of 1000 points, point by point, and plot the histogram of the resulting 1000 points. What do you observe?

   (b) Estimate the mean and the variance of the sum.

2. **Central Limit Theorem**. Let X_i be an iid Bernoulli random variable with value {-1,1}. Look at the random variable $Z_n = \frac{1}{\sqrt{n}} \sum X_i$. By taking 1000 draws from Z_n, plot its histogram. Check that for small n (say, 5-10) Z_n does not look that much like a Gaussian, but when n is bigger (already by the time $n = 30$ or 50) it looks much more like a Gaussian. Check also for much bigger n: $n = 250$, to see that at this point, one can really see the bell curve.

3. Estimate the mean and standard deviation from 1 dimensional data: generate 25,000 samples from a Gaussian distribution with mean 0 and standard deviation 5. Then estimate the mean and standard deviation of this gaussian using elementary numpy commands, i.e., addition, multiplication, division (do not use a command that takes data and returns the mean or standard deviation).

4. Estimate the mean and covariance matrix for multi-dimensional data: generate 10,000 samples of 2 dimensional data from the Gaussian distribution

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} -5 \\ 5 \end{pmatrix}, \begin{pmatrix} 20 & .8 \\ .8 & 30 \end{pmatrix} \right). \tag{1}$$

   Then, estimate the mean and covariance matrix for this multi-dimensional data using elementary numpy commands, i.e., addition, multiplication, division (do not use a command that takes data and returns the mean or standard deviation).

5. Download from Canvas/Files the dataset `PatientData.csv`.

   Each row is a patient and the last column is the condition that the patient has. Do data exploration using Pandas and other visualization tools to understand what you can about the dataset. For example:

   (a) How many patients and how many features are there?
   (b) What is the meaning of the first 4 features? See if you can understand what they mean.
   (c) Are there missing values? Replace them with the average of the corresponding feature column
   (d) How could you test which features strongly influence the patient condition and which do not?

   List what you think are the three most important features.

Written Questions

1. Consider two random variables X,Y that are not independent. Their probabilities of are given by the following table:

	X=0	X=1
Y=0	1/4	1/4
Y=1	1/6	1/3

   (a) What is the probability that $X = 1$?
   (b) What is the probability that $X = 1$ conditioned on $Y = 1$?
   (c) What is the variance of the random variable X?
   (d) What is the variance of the random variable X conditioned that $Y = 1$?
   (e) What is $E[X^3 + X^2 + 3\,Y^7 | Y = 1|]$?

2. Consider the vectors $\mathbf{v}_1 = [1, 1, 1]$ and $\mathbf{v}_2 = [1, 0, 0]$. These two vectors define a 2-dimensional subspace of $\mathbb{R}^3$. Project the points $P1 = [3, 3, 3], P2 = [1, 2, 3], P3 = [0, 0, 1]$ on this subspace. Write down the coordinates of the three projected points. (You can use numpy or a calculator to do arithmetic if you want).

3. Consider a coin such that probability of heads is 2/3. Suppose you toss the coin 100 times. Estimate the probability of getting 50 or fewer heads. You can do this in a variety of ways. One way is to use the Central Limit Theorem. Be explicit in your calculations and tell us what tools you are using in these.

For help: read this introduction to Pandas `http://pandas.pydata.org/pandas-docs/stable/10min.html` and this workflow of exploring features (for a different dataset) `https://www.kaggle.com/cast42/exploring-features`