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

Business Data Science — Fall 2022

HOMEWORK ONE

Dimakis Due: Tuesday 9/6/22, Midnight.

Comments/Remarks: Homeworks done in groups of three. Every group member please submit the homework report. Include names of all group members in the report.

Each submission should contain: a PDF report with names of all group members, along with all code files either in `.py` or `.ipynb` format. The PDF report should include all requested deliverables: **written answers, plots, code, code output and any discussion**, as applicable. You can submit code as .ipynb format or .py format. For `.py` files, name them in the format `problemX.py` or if needed, `problemXa.py`, `problemXb.py`, and so on.

Please keep track of the contributions of each group member in the homeworks. You will be later asked to peer evaluate your group members in terms of their contributions.

Programming Questions

Note: make sure to include the code snippet relevant to each question in the PDF report, even if you're submitting separate code files.

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 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?
   Deliverables: three histograms (two for each Gaussian, one for the sum), written response, code

   (b) Estimate the mean and the variance of the sum.
   Deliverables: written response, code

2. 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).
   Deliverables: mean, standard deviation, code

3. 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).
Deliverables: mean, covariance matrix, code

4. (Introduction to Data exploration) 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.
   Deliverables: written response, code (if any)

Written Questions

1. (Linear Algebra refresh): 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).
   Deliverables: written response, code (if any)

2. (Extra credit +10pts) 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.
   Deliverables: written response and/or code used

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`