Programming Assignment 1
(Advanced Data Analytics DS-503)

Instructions: This assignment can be done in pairs (group of 2 students). Each student in the group must understand the code and techniques used and be able to answer each question (don't divide the problems amongst yourselves). It is fine to use web-resources and help, provided you reference them and understand the code. Please don't copy blindly.

Due Date: Aug 31, 2021 (End of day)

Late Days: You have total of 5 late days for the semester for all homework and assignments, which you can use as you need them

Submission: Submit your solution for each question as a separate Jupyter or Colab notebook. You can also submit a word or .pdf file that explains the solutions. Please use the headings and comments appropriately so we can understand your solution. Use of Matlab or other programming languages like C, C++, Java, R is also permitted, provided the code is well documented. We may ask you for a demo. Good solutions will be posted on the Course Git-hub.

Zip all your code and send to Aman along with your names and roll number in one zip file. You can also send him a Google drive link with appropriate permissions.

(30 marks) Problem 1: *Image Dataset Exploration*

IMAGE DATASET SELECTION

You can choose any one of the datasets, according to your interest and comfort. If you want to explore some other dataset (must have more than 50,000 images and multiple classes), please send the information to Aman and seek approval.

Option 1: The Street View House Numbers (SVHN) Dataset

http://ufldl.stanford.edu/housenumbers/

Training data: train_32x32.mat and test_32x32.mat

An MNIST like dataset, but from real life images from Google Streetview.

Pre-processing help: https://github.com/aditya9211/SVHN-CNN/blob/master/data_preprocess.ipynb

Option 2: Hindi character images

You can explore the Hindi Characters dataset available at the UCI ML repository.

Each image is 28x28 with padding of 2 pixels added on all sides. There are 46 classes in this dataset.

https://archive-beta.ics.uci.edu/ml/datasets/389

https://archive.ics.uci.edu/ml/datasets/Devanagari+Handwritten+Character+Dataset

Pre-processing help: Check the sample code

https://nbviewer.jupyter.org/github/rishianand9/devanagari-character-recognition/blob/master/DCRS.ipynb

Questions:

(5 marks) Compute the Data Covariance matrix for each class (e.g., digit, character) separately and then entire dataset.

(2.5 marks) Which class has the minimum and maximum total variance (spread)? Let's call them class A and B respectively. Provide some example images from class A and B.

(2.5 marks) Perform normalization on the data from each class. Which method is more appropriate here?

(10 marks) Use PCA and related techniques (https://scikit-learn.org/stable/modules/decomposition.html#decompositions) to visualize the data in class A and B individually and together. Are the two classes distinguishable?

(5 marks) Take 10 random test images from the test set of classes A and B each. Assuming that the distribution of each class is a multi-variate normal (with parameters same as that of the sample covariance matrix for that class), compute the probability of each image belonging to class A and B. How accurate is your classifier?

(5 marks) Repeat the above exercise with a 5-nearest neighbor approach in lower dimensions using your favorite dimensionality reduction technique.

Extra Credits for good visualizations and explanations.

Note: If you want to skip this question, you can instead solve the spring-mass system question on PCA from http://databookuw.com/page-17/

(20 marks) Problem 2: *Linear Regression using matrix methods*

In this exercise, we will be using the matrix methods to compute the "least squares" solution to the following problem.

Dataset: https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant# contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, the other three of the ambient variables effect the GT performance.

(5 marks) Fit a linear regression model to predict/estimate the EP of the plant based on T, AP, RH, V and EP. Write the equation in terms of the variables, including the bias term. You must use the normal equations method for this part.

(5 marks) Compute the least squares error of your solution. Find 5 data points with highest error. Are they outliers in the dataset? Compare the values of their attributes with the range of the attribute, (i. e. comment if they lie near the surface of the data hyper-space)?

(5 marks) Repeat the analysis with Pseudo-inverse (using Randomized SVD method) and QR factorization method.

https://scikit-learn.org/stable/modules/generated/sklearn.utils.extmath.randomized_svd.html

(5 marks) Identify any differences in the weight coefficients and or the least squares error obtained by the 3 methods. Comment about the numerical stability of the matrices and time taken to obtain the solutions.

You can use the linear algebra libraries in numpy. (https://numpy.org/doc/stable/reference/routines.linalg.html)

Note: For large scale problems, the gradient descent approach is used to solve optimization problems. Don't use it for this assignment. We will study about it later in the course.

Extra credits for visualizations of the dataset and good interpretation of the results.

(30 marks) Problem 3: *Mining Original and Classic Papers in NIPS Conference using LSH*

Dataset Preparation

Obtain the papers published in the NIPS conference between 1987 to 2016 from Kaggle (https://www.kaggle.com/benhamner/nips-papers)

(5 marks) **Data Collection:** Scrape the titles and abstracts from https://papers.nips.cc/ for the year 2017 and 2020. You can use BeautifulSoup or similar tools to do so.

(2.5 marks) **Pre-processing:** Create tokens/shingles from the paper (use the text of Title and abstract). You can also work with unigrams (individual words) or phrases (sequence of words).

(2.5 marks) **Build LSH**: Build a LSH data structure to store each paper from 1987-2016.

(10 marks) **Original Papers in 2017:** Our goal is to find top 5 unique or original papers in 2017, i. e. papers which had least similarity to papers in previous years (1987-2016). Design a strategy to determine originality. Justify your approach.

(10 marks) **Classic Papers:** Our goal now is to find the top 5 oldest, classic papers in NIPS proceedings whose topics were relevant even in 2020. Design your scoring function and justify its choice. You can use MinHash LSH Forest for this, which answers a top-K query. (http://ekzhu.com/datasketch/lshforest.html).

Extra credits for tuning thresholds, and controlling false-positives and false-negatives based on the requirement of the analysis.