CS 5525 Assignment 1

Due Date:

Due 11:59 pm, 24th September 2020

Reminders:

1. Out of 100 points. Contains 4 pages.
2. Please type your answers. No handwritten (or scanned) assignments will be accepted (or graded). Latex is not mandatory, but encouraged.
3. The completed assignment must be submitted on Canvas as a zip file that contains both a PDF and a Jupyter Notebook before the due date. Include your name and PID in both files. Files should be named HW1_<name>_<pid>.<ext>, where <ext> is pdf, .ipynb or .zip.
4. Each solution must include all details and an explanation of why the given solution is correct. In particular, write clear sentences and give proper illustrations that support your claim. A correct answer without an explanation is worth no credit.
5. There could be more than one correct answer. We shall accept them all.
6. Whenever you are making an assumption, please state it clearly.

Coding Instructions:

1. All the programing should be done in **Python 3**. You may use any of the popular python packages like numpy, scipy, sklearn, matplotlib.
2. Use a single **Jupyter Notebook** file to submit all your code.
3. Please comment your code properly. Make sure to also include the corresponding problem number for each code block.
4. The data should be read from the same path as the code. Do NOT use absolute file path.
5. We will run each code block in the Jupyter Notebook you submit. Failure in running or incorrect output might cause credit deduction.

Problem 1 **10 points**

look at this example in the form of a column vector https://corporatefinanceinstitute.com/resources/knowledge/finance/covariance/

Solve the following problems.

(a) Two vectors x and y have zero mean. What is the relationship of the cosine measure and correlation between them? [4]

(b) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object vector has an L2 length (magnitude) of 1. [6]

(NOTE: your final answer should be independent of the original vectors).

Problem 2 **8 points**

individual articles A belong to A_e where e is an event, and events belong to a domain D.

Consider this hypothetical problem. You have a set of news articles $A = \{A\}$ given to you, where A indicates one article. These news articles span across different domains like civil unrest, earthquakes, sport, etc. Assume that $\mathbf{A}_e$ is the set of news articles related to event e. An event e can fall under any domain D. On aggregating the news articles of event e of particular domain D, we get the domain-set $\mathbf{A}_D = \{A \mid A \in \mathbf{A}_e \forall e \in D\}$. Suggest two measures *domain weight* $\mathsf{C}_{w_i,D}$ and *event weight* $\mathsf{E}_{w_i,e}$ for a word w_i. $\mathsf{C}_{w_i,D}$ quantifies the ability of word w_i in representing targeted domain D. $\mathsf{E}_{w_i,e}$ quantifies the ability of word w_i in distinguishing event e from other events in the same domain. Make use of this function $f(w, A)$ which gives the frequency of word w in article set A. You are free to make assumptions like merging all the articles related to one event into a single article. [4 + 4]

(*Hint*: It should be a product term.)

Problem 3 **10 points**

Mike completes jogging one round on a (circular) athletic track of radius 1 mile. John is waiting for him at the center of the track. Compute the minimum and maximum possible values for the following distance measures between Mike and John while Mike is jogging: Manhattan, Euclidean and Chebyshev distance. For full credit give the proper mathematical notations.

Problem 4 **15 points**

Solve the following problems based on the given dataset (Named 'data.mat', you can read it with the python library scipy), in which there are 200 data points in 3-dimensional feature space.

1) Calculate the Euclidean distances between each pair of the data points (x_i, x_j) (a 200×200 distance matrix), and report the distances among the first 8 data points (an 8×8 distance matrix). [3]

   Construct the neighborhood graphs respectively by using the following two different criteria.

   (a) Connect points x_i and x_j if x_i is one of the 5 nearest neighbors of x_j. [3]

   (b) Connect points x_i and x_j if their distance is less than 6. [3]

   Write a function `constructGraph()` to implement it and provide the graph you constructed along with the code.

2) For the first 8 points (as shown in the table), compute the Geodesic distance between each pair of these points using Dijkstra's shortest path algorithm. Write a function `geodesic()` to implement and provide the 8×8 distance matrix along with the code. [6]

1	2	3	4	5	6	7	8
-7.8167	11.6325	4.9895	-3.3580	7.9544	-5.4562	11.3690	-2.3936
-6.4150	-3.8339	-2.8779	7.5597	-8.6345	-8.9078	-4.4145	7.7927
15.5175	12.9623	27.8192	19.8803	12.8265	7.1760	9.2878	33.0461

Fig. 1. The first 8 rows of the data.

transposed data

Problem 5 Analyze the data. **7 points**

Load the data from the file [1]`thick.csv`. [5]

1) Plot a boxplot of the first 100 rows of data.

2) Explain why the thick center line in the box plot is not symmetrical with the outer edges of the box. [2]

[1]Thickness of 2x6 SPF boards from a saw mill. It is measured with a laser and the units of measurement are mils.

Problem 6 **50 points**

This data set is from a food manufacturer making a pastry product. Each sample (row) in the data set is taken from a batch of product where 5 quality attributes are measured:

1) Percentage oil in the pastry

2) The product's density (the higher the number, the more dense the product)

3) A crispiness measurement, on a scale from 7 to 15, with 15 being more crispy.

4) The product's fracturability: the angle, in degrees, through which the pasty can be slowly bent before it fractures.

5) Hardness: the amount of force required before breakage occurs.

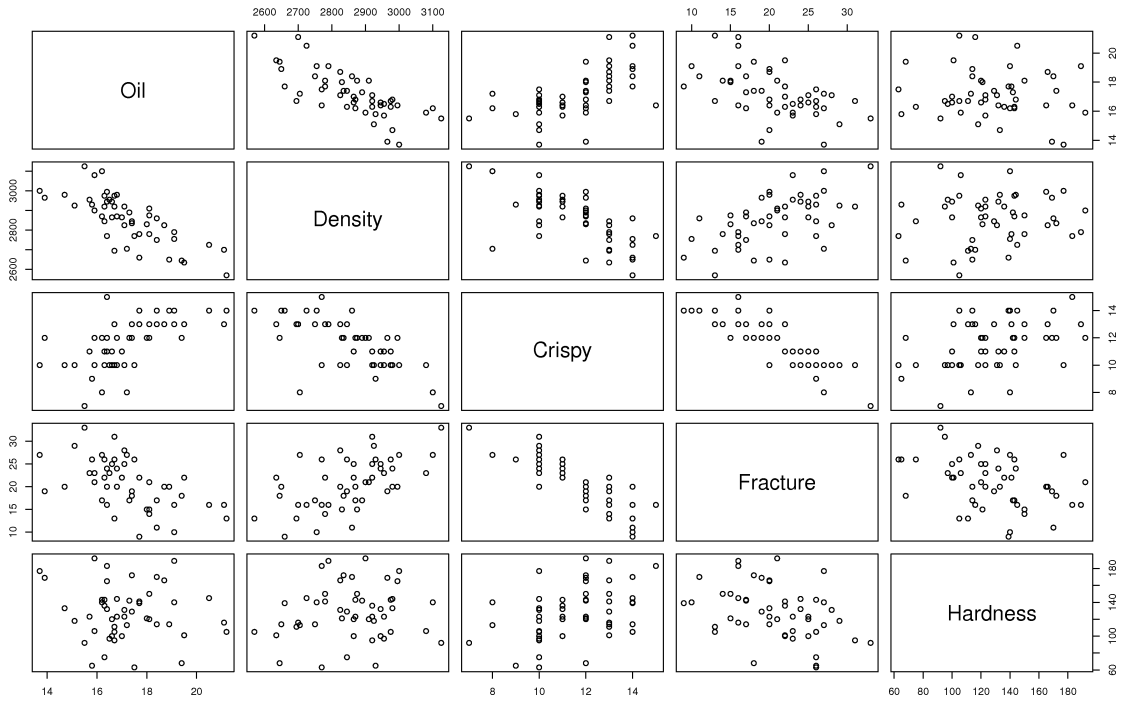

Fig. 2. A scatter plot matrix of these 5 features is shown for the *N* observations.

Do the following.

1) **Import:** Import the data file **food.csv** in python. [2]

2) **Preprocessing**: Centering removes any bias terms from the data. Scaling re
raw data could be in diverse units You are to center the data to ensure that it re
then scale it to have unit variance. Give the formula for performing these tw
these two tasks in python. [3+5]

https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#:~:text=A%20Summary%20of%20the%20PCA%20Approach,-Standardize%20the%20data&text=Sort%20eigenvalues%20in%20descending%20order,from%20the%20selected%20k%20eigenvectors.

3)How does these two operations alter the overall interpretation of the data? [2]

4)Now you have the pre-processed data matrix **X**. What is the formula for correlation matrix? Implement it in python. [2+2]

5)Calculate the eigenvectors and eigenvalues of this square matrix. [2]

6)Sort the eigenvalues from largest to smallest. Accordingly update the order of the eigenvectors in matrix. Plot the percentage of variance captured by the individual components in decreasing order (**Hint**: Scree plot of the eigenvalues). [2+2+3]

7)If you were to project the matrix **X** on the eigenvectors to obtain the principal components, how many components would you use and why? Implement it in python for $2-D$ projection. You carried out the steps of which algorithm? (Hint: Not PCA.). [2+4+1]

8)Give the scatter plot of the first two components obtained above. [3]

9)So far we have used an algorithm based on eigenvalues and eigenvectors. Is it recommended? Explain. [1+2]

https://stats.stackexchange.com/questions/448200/is-pca-always-recommended#:~:text=PCA%20is%20a%20linear%20model,nonzero%20effect%20on%20each%20PC.

10) Suppose you have an alternative approach that factorizes **X** into product of two orthonormal matrices **A** and **B**, and one diagonal matrix **D**. Which method comes to your mind? Implement it in python. [1+3]

11) Summarize (project) the data matrix **X** in 2-D space making use of the matrices **A**, **B** and **D** obtained above. Give the $2-D$ scatter plot. [2+3]

plotting

https://hadrienj.github.io/posts/Deep-Learning-Book-Series-2.8-Singular-Value-Decomposition/

12) Compare the plots obtained in part 8 and 11. Do you notice any similarity? Explain with proper mathematical equations to justify your answers. [1+2]