Virginia Tech. **CS 5525/STAT 5525 – Data Analytics I**

Computer Science **Spring 2018, Prakash**

Homework 1

(due Feb 8, 2018, 11:00am, submit PDF on Canvas)

Reminders:

a. Out of 100 points. Contains 5 pages.
b. Please type your answers. **No handwritten (or scanned)** assignments will be accepted (or graded). Latex is not required, but encouraged.
c. The completed assignment must be submitted on Canvas as a PDF by 11:00AM on February 8, 2018. Include your name and PID in your assignment PDF.
d. 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.
e. There could be more than one correct answer. We shall accept them all.
f. Whenever you are making an assumption, please state it clearly.
g. All programming should be done in MATLAB. **No exceptions.** For those in the CS department, you can find MATLAB here. Others not having departmental version of the software can purchase it from Software Distribution.

Q1. Similarity and Distance [10 points]

Solve the following two problems:

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

(b) (6 points) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object vector has an L2 length (magnitude) of 1. NOTE: your final answer should be independent of the original vectors.

Q2. News Articles [8 points]

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 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 $A_D = \{A \mid A \in A_e \; \forall e \in D\}$. Suggest two measures *domain weight* $C_{W_i,D}$ and *event weight* $E_{W_i,e}$ for a word w_i. $C_{W_i,D}$ quantifies the ability of word w_i in representing targeted domain *D*. $E_{W_i,e}$ 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 points for each measure.

Hint: It should be a product term.

Q3. Jogging around [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.

Q4. Geodesics and Graphs [15 points]

Solve the following problems based on the given dataset (Named 'data.mat'), in which there are 200 data points in 3-dimensional feature space.

1. (3 points) Calculate the Euclidean distances between each pair of the data points (xi, xj) (a 200×200 distance matrix), and report the distances among the first 8 data points (a 8×8 distance matrix).

   Now construct *neighborhood graphs* by using the following two different criteria respectively.

   (a) (3 points) Connect points xi and xj if xi is one of the 5 nearest neighbors of xj.

   (b) (3 points) Connect points xi and xj if their distance is less than 6.

2. (6 points) 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.

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

Figure 1: The first 8 points.

Q5. Boxing Boards [7 points]

Load the data from the file thick.csv. It contains the thickness of 2x6 SPF boards from a sawmill. It is measured with a laser and the units of measurement are mils.

(1) (5 points) Plot a boxplot of the first 100 rows of data.

(2) (2 points) Explain why the thick centerline in the box plot is not symmetrical with the outer edges of the box.

Q6. Projecting Pastries [50 points]

This data set is from a food manufacturer making a pastry product. As any Michelin star pastry chef will tell you, any batch of pastry can be evaluated on the following 5 quality attributes (each row in the data is one batch):

1. Percentage of oil/butter in the pastry
2. The product's density (higher the number, denser the product)
3. A crispiness measurement, on a scale from 7 to 15, with 15 being crispier.
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.

And if you have ever made a croissant, you know that many of these attributes are correlated. Your task in this problem is to explore projecting this dataset on a smaller dimensional space, using what you learnt in class.

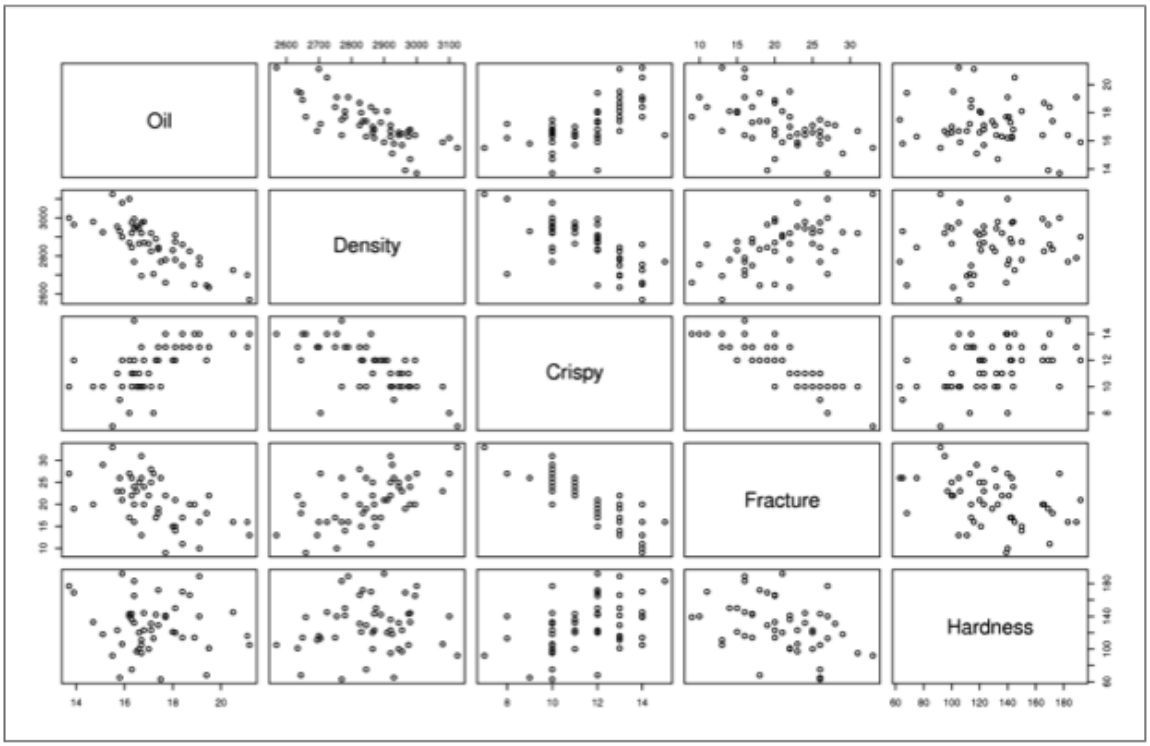

Figure 2: A scatter plot matrix of these 5 features for the N observations.

1. (2 points) **Import:** Import the data file food.csv in Matlab.

2. (3+5 points) **Processing:** Centering removes any bias terms from the data. Scaling removes the fact that the raw data could be in diverse units. You are to center the data to ensure that it results in zero mean and then scale it to have unit variance. Give the formula for performing these two operations. Perform these two tasks in Matlab.

3. (2 points) How does these two operations alter the overall interpretation of the data?

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

5. (2 points) Calculate the eigenvectors and eigenvalues of this square matrix. (Make use of built-in Matlab functions like *eigs* for these).

6. (2+2+3 points) 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*: Do the scree plot of the eigenvalues energies.

7. (2+4+1 points) 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 Matlab for a 2−D projection. You carried out the steps of which algorithm?

8. (3 points) Give the scatter plot of the first two components obtained above (use the *scatter* function in Matlab).

9. (1+2 points) So far we have used an algorithm based on eigenvalues and eigenvectors. Is it recommended? Explain in 1-2 lines.

10. (1+3 points) 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 Matlab.

11. (2+3 points) 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.

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

Note 1: Write your code in the form of a single MATLAB script file (not a function file). Try make use of commands like figure(), subplot(), plot(), scatter(), etc. Paste the code after you have written your answers.

Note 2: Your code should run on a machine with just Matlab installed; no other dependencies. We will copy the code from this file and paste it in a script file (contained in the folder with the food.csv file) and execute it. Failing to run the code will result in **no credits** for programming problems.