CSE 40647/60647: Data Science	Fall 2017
Homework 1	
Handed Out: August 29, 2017	*Due: September 12, 2017 11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Sakai (https://sakailogin.nd.edu/portal/site/FA17-CSE-40647-CX-01) for collecting this assignment. Contact TA if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!

- The homework MUST be submitted in pdf format. Handwritten answers are not acceptable. Name your pdf file as YourNetid-HW1.pdf

- You need to explain the logic of your answer/result for every question. A result/answer without any explanation will not receive any point.

- It is OK to discuss the problems with the TA and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the Honor code on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.

- Please use Piazza if you have questions about the homework. Also feel free to send TA emails and come to office hours.

2 Question 1 (10 points)

Suppose we sample mid-term and final scores of 9 students from 1,000 students:

- Student_A: 71 (mid-term), 73 (final)

- Student_B: 85, 87

- Student_C: 83, 83

- Student_D: 98, 97

- Student_E: 76, 87

- Student_F: 81, 83

- Student_G: 76, 83

- Student_H: 82, 84

- Student_I: 95, 97

1. (1′) Calculate *mean, median, mode* for the mid-term exam scores.

2. (1′) Draw box plot for the mid-term exam scores: calculate *min, max, Q1, Q3*. [Skip "box plot"]

3. (1′) Calculate *mean, median, mode* for the final exam scores.

4. (1′) Draw box plot for the final exam scores: calculate *min, max, Q1, Q3*. [Skip "box plot"]

5. (1′) Min-max normalization on the mid-term exam scores.

6. (1′) Min-max normalization on the final exam scores.

7. (1′) Calculate *variance, standard deviation* for the mid-term exam scores.

8. (1′) Calculate *variance, standard deviation* for the final exam scores.

9. (1′) Z-score normalization on the mid-term exam scores.

10. (1′) Z-score normalization on the final exam scores.

3   Question 2 (10 points)

Visualize the above sample data:

1. (5′) Draw the Q-Q plot of the above sample data. Compare the two variables, mid-term exam score and final exam score: Which exam is easier?

2. (5′) Take each student as a data object, and take two exam scores as two attributes. Draw a scatter plot of the students' performance.

4   Question 3 (10 points)

Suppose the sample dataset has n values: $x_1, ..., x_n$.

1. (2′) Write the formula of the mean μ and variance v of the dataset.

2. (8′) Suppose we have a new value x_{n+1}. Write down how to compute the new mean μ' and new variance v' based on μ, v, and the increment value x_{n+1}. Note that the size of dataset becomes $n + 1$. (Hint: You can only use μ, n, x_{n+1} to compute μ'. You can only use v, μ, n, x_{n+1} to compute v'.)

5 Question 4 (10 points)

Let's use the Z-score normalized exam scores we have in Question 1. Take mid-term exam and final exam as two data objects, and take the students as attributes. The attribute values are Z-score normalized exam scores.

1. (6') Calculate the Manhattan distance, Euclidean distance and supremum distance between the two data objects.

2. (4') Which distance is often the biggest? Which distance is often the smallest? Can you explain why?

6 Question 5 (10 points)

Suppose we have $1,000,000$ documents. The word "matrix" appears in $1,100$ of them, and the word "factorization" also appears in $1,100$ of them. And there are 100 documents that have both the word "matrix" and the word "factorization".

1. (5') We take a word's occurrence in a document as a binary variable. Are the variables of "matrix" or "factorization" symmetric or asymmetric binary variables? Why?

2. (5') What is the Jaccard similarity between the two variables?

7 Question 6 (10 points)

We use the same data assumption as Question 5. Take each word as a data object, and take the documents as attributes. Usually we use word vectors to represent documents when we calculate the cosine similarity between documents. But now we want to measure the similarity between words.

1. (4') Now the word "matrix" and the word "factorization" are represented with a $1,000,000$-length binary vector. What is the cosine similarity?

2. (6') Suppose we have n words in the set of documents. What is the range of cosine similarity between any pair of words? What is the range of Jaccard similarity between any pair of words? And what is the range of Euclidean distance?

8 Question 7 (10 points)

(10') We use the same data assumption as Question 5. We take a word's occurrence in a document as a binary variable. So we have two variables: (1) the word "matrix" in documents, and (2) the word "factorization" in documents. Please conduct a chi-square test on the two variables and give your conclusion.

9 Question 8 (10 points)

(10') We use original mid-term and final exam score data as in Question 1. Suppose the two exams are two variables. What is the covariance between them? Is it positive or negative?

10 Question 9 (10 points)

(10') Give a case that Linear Regression model generates much bigger fitting error than Log-Log Regression model.

11 Question 10 (10 points)

(10') Principal component analysis (PCA) is usually explained via an eigen-decomposition of the covariance matrix. Is PCA a linear dimensionality reduction technique or a non-linear dimensionality reduction technique?

12 Question 11 (0 point): Project follow-up

We hope you have made some progress on Task 1, 2, and 3.

1. How many papers and authors do you have after *data cleaning and data integration*? List the number of papers you have for each proceeding (conference+year).

2. How many entities (in lower case; like "support vector machines" and "svms") do you have? How did you find the entities?

3. How many typed entities do you have? List the number of entities for each type. List at least 10 entities for each type that you feel confident to say they are correctly typed (e.g., "svms: $Method").