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*.

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*.

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.

Solution:

1. $mean_{mid-term} = \frac{71+85+83+98+76+81+76+82+95}{9} = 83$. Sorted: $71, 76, 76, 81, 82, 83, 85, 95, 98$; $median_{mid-term} = 82$. $mode_{mid-term} = 76$.

2. $min_{mid-term} = 71$. $max_{mid-term} = 98$. $Q1_{mid-term} = 76$. $Q3_{mid-term} = 85$.

3. $mean_{final} = \frac{73+87+83+97+87+83+83+84+97}{9} = 86$. Sorted: $73, 83, 83, 83, 84, 87, 87, 97, 97$; $median_{final} = 84$. $mode_{mid-term} = 83$.

4. $min_{final} = 73$. $max_{final} = 97$. $Q1_{final} = 83$. $Q3_{final} = 87$.

5. Min-max normalized mid-term exam scores: $0, 0.52, 0.44, 1, 0.19, 0.37, 0.19, 0.41, 0.89$.

6. Min-max normalized final exam scores: $0, 0.58, 0.42, 1, 0.58, 0.42, 0.42, 0.46, 1$.

7. $variance_{mid-term} = \frac{(71-83)^2+(85-83)^2+(83-83)^2+(98-83)^2+(76-83)^2+(81-83)^2+(76-83)^2+(82-83)^2+(95-83)^2}{8} = 77.5$. $standard\_deviation_{mid-term} = \sqrt{77.5} = 8.80$.

8. $variance_{final} = \frac{(73-86)^2+(87-86)^2+(83-86)^2+(97-86)^2+(87-86)^2+(83-86)^2+(83-86)^2+(84-86)^2+(97-86)^2}{8} = 55.5$. $standard\_deviation_{final} = \sqrt{55.5} = 7.45$.

9. Z-score normalized mid-term exam scores: $-1.36, 0.23, 0, 1.70, -0.80, -0.23, -0.80, -0.11, 1.36$.

10. Z-score normalized final exam scores: $-1.75, 0.13, -0.40, 1.48, 0.13, -0.40, -0.40, -0.27, 1.48$.

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, midterm 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.

Solution:

1. See HW1-DataScienceFall17-plots.pdf

2. See HW1-DataScienceFall17-plots.pdf

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'.)

Solution:

1. $\mu = \frac{x_1+...+x_n}{n}$. $v = \frac{(x_1-\mu)^2+...+(x_n-\mu)^2}{n-1}$.

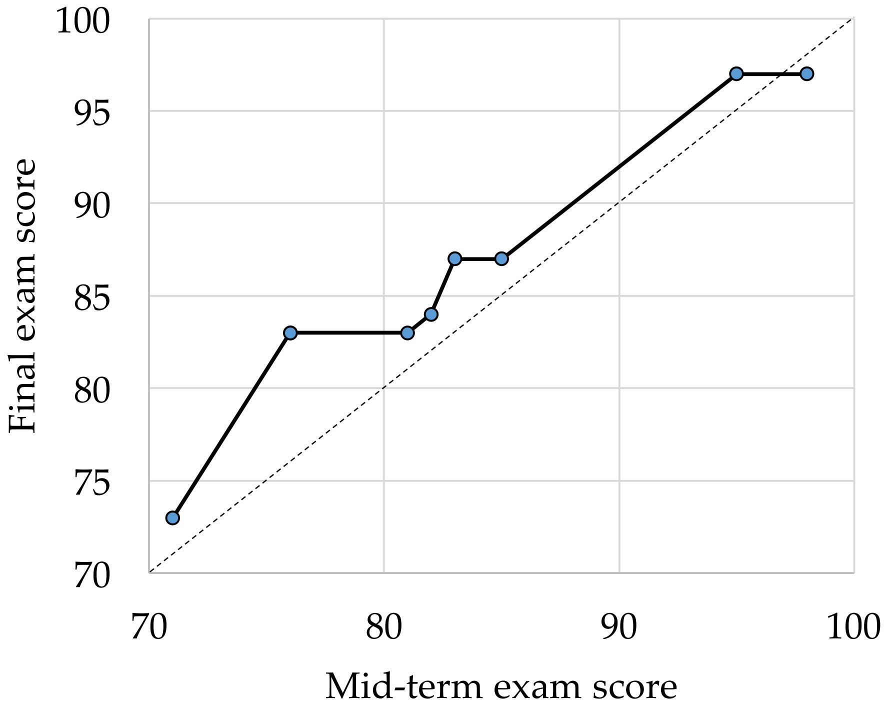
2. $\mu' = \frac{x_1+...+x_n+x_{n+1}}{n+1} = \frac{n\mu+x_{n+1}}{n+1}$. $v' = \frac{(x_1-\mu')^2+...+(x_n-\mu')^2+(x_{n+1}-\mu')^2}{n}$. We have

$$\begin{aligned}
nv' - (n-1)v &= \{(x_1-\mu')^2 - (x_1-\mu)^2\} + ... + \{(x_n-\mu')^2 - (x_n-\mu)^2\} + (x_{n+1}-\mu')^2 \\
&= (2x_1-\mu-\mu') \times (\mu-\mu') + ... + (2x_n-\mu-\mu') \times (\mu-\mu') + (x_{n+1}-\mu')^2 \\
&= \{2 \times (x_1+...+x_n) - n\mu - n\mu'\} \times (\mu-\mu') + (x_{n+1}-\mu')^2 \\
&= (2n\mu - n\mu - n\mu') \times (\mu-\mu') + (x_{n+1}-\mu')^2 \\
&= (n\mu - n\mu') \times (\mu-\mu') + (x_{n+1}-\mu')^2 \\
&= n(\mu-\mu')^2 + (x_{n+1}-\mu')^2
\end{aligned}$$

So, $v' = v + (\mu-\mu')^2 + \frac{(x_{n+1}-\mu')^2-v}{n} = \frac{n-1}{n}v + \frac{1}{n+1}(x_{n+1}-\mu)^2$.

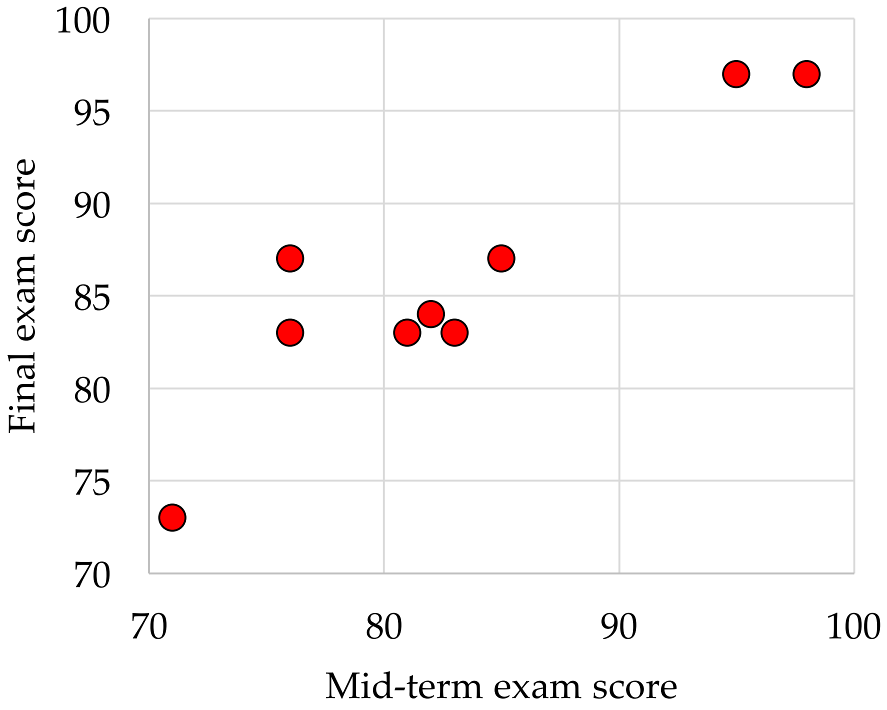
Solution 2:

1. Q-Q plot



Final exam score tends to be easier.

2. Scatter plot



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?

Solution: The feature vector of mid-term exam is
$[-1.36, 0.23, 0, 1.70, -0.80, -0.23, -0.80, -0.11, 1.36]$.
The feature vector of final exam is
$[-1.75, 0.13, -0.40, 1.48, 0.13, -0.40, -0.40, -0.27, 1.48]$.

1. Manhanttan distance is 2.87. Euclidean distance is 1.21. Supremum distance is 0.93.

2. Manhanttan distance is the biggest. Supremum distance is the smallest. Suppose we have two objects and their two features. We denote the absolute error of each feature as a and b. We know that Manhattan distance is $a+b$, Euclidean distance is $\sqrt{a^2+b^2}$, and supremum distance is $\max(a,b)$. It is easy to know $a+b \geq \sqrt{a^2+b^2} \geq \max(a,b)$. We have Manhattan ≥ Euclidean ≥ Supremum.

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?

Solution:

1. They are asymmetric binary variables. Negative is much more frequent than positive in the data. In other words, positive is more important than negative.

	Positive	Negative	Sum
Positive	100	1000	1100
Negative	1000	997900	998900
Sum	1100	998900	1000000

2. The Jaccard similarity is $\frac{100}{100+1000+1000} = \frac{100}{2100} = 0.048$.

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′) Again, we use the same data assumption as Question 5 (like we have 1 million documents). And suppose we have at least two unique 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 between any pair of words? (Hint: Can the similarities be negative in this case?)

Solution:

1. The cosine similarity is $\frac{100}{\sqrt{1100}\times\sqrt{1100}} = 0.091$.

2. The range of cosine simliarity is $[0, 1]$. The range of Jaccard similarity is also $[0, 1]$. The range of Euclidean distance is $[0, \sqrt{1000000}] = [0, 1000]$.

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.

Solution:

	Positive	Negative	Sum
Positive	100 (1.21)	1000 (1098.79)	1100
Negative	1000 (1098.79)	997900 (997801.21)	998900
Sum	1100	998900	1000000

$$\begin{aligned}\chi^2 &= \frac{(100-1.21)^2}{1.21} + 2\times\frac{(1000-1098.79)^2}{1098.79} + \frac{(997900-997801.21)^2}{997801.21}\\ &= 8083.4\end{aligned}$$

We strongly reject the null hypothesis of independence between these two words (variables). They are strongly correlated.

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?

Solution:

$$\begin{aligned}\sigma_{12} &= \frac{71 \times 73 + 85 \times 87 + 83 \times 83 + 98 \times 97 + 76 \times 87 + 81 \times 83 + 76 \times 83 + 82 \times 84 + 95 \times 97}{9} \\ &\quad -83 \times 86 \\ &= 53 > 0\end{aligned}$$

They are positively correlated.

10 Question 9 (10 points)

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

Solution: Any power-law distribution example: e.g., degree distributions of social networks.

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?

Solution: Linear dimensionality reduction. Wikipedia: *The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. In practice, the covariance (and sometimes the correlation) matrix of the data is constructed and the eigen vectors on this matrix are computed. The eigen vectors that correspond to the largest eigenvalues (the principal components) can now be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few eigen vectors can often be interpreted in terms of the large-scale physical behavior of the system. The original space (with dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors.*