CS 412: Introduction to Data Mining **Spring 2018**

Homework 1

Handed Out: February 7, 2018 *Due: February* 21, 2018 *11:59 pm*

General Instructions

- It is OK to discuss with your classmates and your TAs regarding the methods, but 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 CS Honor code (http://cs.illinois.edu/academics/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 first if you have questions about the homework. We will maintain a FAQ post in piazza for each HW, and will update it from time to time with major queries and clarifications. Please check FAQ post first for clarifications and then post a question in piazza if it's not answered in the FAQ. **Tag your question with HW1**. Also feel free to send TAs emails and come to office hours.
- The non-programming part of homework **MUST** be submitted in pdf format. Please DO NOT zip the PDF file so that graders can access your PDF directly on Compass. Handwritten answers or hand-drawn pictures are not acceptable
- The programming part of this assignment will be hosted on hackerrank (https://www.hackerrank.com/) as a programming contest. To participate in this contest, please open a hackerrank account with your illinois.edu email id. If your username in hackerrank is different from your net id, let us know by filling out your net id and username in the spreadsheet (link will be provided in Piazza). The contest framework will allow you to verify the correctness of your submission based on a set of sample test cases. We may use additional test cases to grade your submission. Please check the assignment page on course website on or after February 8th for accessing the contest.
- The homework is due at 11:59 PM on the due date. We will be using Compass for collecting the homework assignments. Please submit your files via Compass (http://compass2g.illinois.edu). Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment.We do NOT accept late homework!
- For each question, you will **NOT** get full credit if you only give out a final result. Necessary calculation steps and reasoning are required.

1 Correlation Analysis (20 points)

Recently, two cargo ships that carried chemical substances collided in the Lake Michigan at Illinois and some toxic substances entered the lake. The government asked researchers of a research institute at Chicago to find a solution to reduce the amount of these toxic substances and accordingly water pollution.

The researchers developed two chemical substances (substance A and substance B) to decrease the water pollution. However, they did not know how to decrease water pollution via these two substances. Therefore, they decided to do some experiments by adding several combinations of the two substances to samples of the lake water and measuring water pollution. Table 1 shows 40 variations of these experiments and their results on water pollution (i.e. the percentage of toxic substances in the water— 0 would be clean water and 100 would be the most possible toxic water).

Table 1: The impact of different combinations of substance A & B on water pollution

Experiment	1	2	3	4	5	6	7	8	9	10
Substance A	2.84	9.34	7.59	0.21	7.31	2.77	4.41	7.34	9.94	7.17
Substance B	78.9	52.6	76.7	39.6	58.4	98.1	4.8	83.4	16.4	86.9
Water Pollution	11.5	17.7	73.4	11.2	74.9	6	15.2	64.9	14.1	76.8

Experiment	11	12	13	14	15	16	17	18	19	20
Substance A	0.66	2.73	2.14	7.79	0.63	7.36	7.21	6.12	4.24	9.79
Substance B	61.6	67.2	85	61.4	11.3	13	88	35.4	53.3	16.8
Water Pollution	11.3	9.4	13.8	55.5	61.4	21.4	90.7	70.1	60	6.7

Experiment	21	22	23	24	25	26	27	28	29	30
Substance A	5.07	0.67	9.83	1.44	5.76	0.17	7.9	9.29	2.99	2
Substance B	87.9	9.2	55.3	32.9	94.4	41.9	15.5	13.6	20.6	93.3
Water Pollution	13	80.1	8.4	64.9	9.3	9.6	11.5	7.5	77	11.9

Experiment	31	32	33	34	35	36	37	38	39	40
Substance A	8.62	7.21	7.44	8.7	3.05	1.74	3.41	7.5	2.84	1.43
Substance B	78.9	20.6	98.8	48.3	86.3	12.7	1.6	22.7	33.6	26.2
Water Pollution	81.2	14.9	83.8	15.8	16.5	69	7.6	10.6	77.9	62.1

Now, the researchers need to decide what combination of substance A and B works best to decrease water pollution. To help them, please follow the analysis guideline below and draw a conclusion:

1. Separate the 40 experiments into two groups via a binning method where one group has low water quality (Group 1) and one group has high water quality (Group 2). Our goal here is to compare the correlation of substance A and B in these two groups to decide how to use substance A and B together to decrease water pollution. To do so, which binning method do you think is more appropriate here? Equal-width or equal-depth? Why? Apply it on the dataset and create two groups.
2. Normalize the Substance A and Substance B amounts based on the min-max normalization for each group, and report the normalized values
3. Draw scatter plots of the normalized substance A and normalized substance B for each group
4. Calculate the Pearson correlation coefficient between the normalized substance A and normalized substance B for each group
5. Based on the analysis in the above steps, draw a conclusion of how to use the two substances to decrease water pollution.

Note: You can find the data in Table 1 at the Q1-data.txt file (in the HW1-data.zip folder) as well.

2 Noisy Data (20 points)

In this question, we want reduce the noise in a dataset via different methods and compare them with each other. Consider the following dataset that is collected about the age of 20 people in a workplace:

$$34, 32, 53, 33, 43, 2, 43, 38, 41, 42, 49, 25, 41, 36, 42, 52, 32, 23, 43, 91$$

However, it turned out that the person in the HR who collected this data didnt report the ages properly, and some of the reported ages might not be exact or might be out of range. So, please follow the below steps to mitigate this issue:

1. Compute the *mean*, *Q1*, *median*, *Q3*, and *standard deviation (population)* of the age measure.
2. Conduct each of the following methods to reduce the noise of the data:
   (a) Draw the boxplot of the age measure, and detect the outliers by finding those observations that are 1.5 IQR above Q3 or below Q1. Now, build a new dataset by removing the outlier, and call the new dataset A.

(b) Use *smoothing by bin means* to smooth the above data, using a bin depth of 5. Call the new dataset B.

(c) Use *smoothing by bin boundaries* to smooth the above data, using a bin depth of 5 (if a value has the same distance from both boundaries of a range, replace it with the lower boundary). Call the new dataset C.

3. Now, for each new datasets of A, B, and C, compute the *mean*, *Q1*, *median*, *Q3*, and *standard deviation* of the age measure again and compare these values with the ones obtained in step 1 (Draw a table to compare these values side by side for each of the datasets including the original dataset).

   Interpret the differences that you observe, and discuss the pros and cons of each of the above methods in reducing noise (e.g. which method is best if we need the lowest variance, which method reduces the distinct values and etc.). This is an open-ended question, so discuss the main patterns you see.

3 A Fair Comparison (10 points)

A data mining company wants to hire a data scientist from the University of Illinois at Urbana-Champaign graduates. After doing a few rounds of interview with tens of applicants, they narrowed down the candidates to five students. Now, the company need to decide between these candidates, and one of the metrics that they want to use is the candidates' GPAs. However, these candidates belong to different classes (from 2014-2018) which makes doing a fair comparison difficult.

Table 2 shows the information of each of these candidates including their GPA, and the mean and standard deviation of GPA in their class. Please help the company to sort these five candidates based on their GPAs fairly. All the classes size were the same (100 students each year).

Table 2: Candidates' information

Student	**Class**	**GPA**	**Mean**	**STD**
A	2014	3.5	3.2	0.5
B	2015	3.7	3.4	0.4
C	2016	3.4	3.2	0.35
D	2017	3.8	3.9	0.5
E	2018	3.9	3.8	0.2

4 Programming Question (50 points)

One method to distinguish cancer versus normal patterns is analyzing mass-spectrometric data (that includes the mass-to-charge ratio of 10,000 ions in the body). During recent years, medical researchers have collected a dataset of mass-spectrometric data of hundreds of patients including patients with cancer (ovarian or prostate cancer), and healthy or control patients. Researchers now want to use this dataset to detect patients that are prone to have cancer by finding those patients who are most similar to the patients who have cancer.

In this question, we want to help these researchers by enabling them to specify a patient's information (patient P) and retrieve the patients whose mass-spectrometric data is most similar to the patient P's mass-spectrometric data. To do this process, follow these steps:

1. Calculate the distance of patient P from the patients via the following distance metrics:
   (a) Minkowski distance where h = 1 (Manhattan distance)
   (b) Minkowski distance where h = 2 (Euclidean distance)
   (c) Minkowski distance where h = infinite (Supremum distance)
   (d) Cosine similarity (negative values for cosine similarity are allowed)
2. Rank patients based on lower distance (higher similarity in the case of cosine similarity) to patient P and return the most five similar patients to patient P
3. Run PCA on the dataset

   Note: You are allowed to use existing PCA packages.
4. Transform the dataset (and patient P) to the new space via X principal components (that will be given as the input), and run steps 1-2 again on the transformed dataset

4.1 Implementation

Since there are several PCA packages that the HackerRank environment does not support, we evaluate your code in two phases–one that does not include PCA via HackerRank and one that includes PCA via analysis report:

1. A HackerRank challenge that evaluates steps 1-2 by measuring the similarity metric on the original dataset without running PCA
2. An analysis report that evaluates steps 3-4 by running PCA on a given dataset in the HW-data.zip folder and reporting the results

To do these evaluations properly, your code should follow the following formats exactly.

4.1.1 Input Format

- ***Line 1***: D (number of data dimensions that can be between 1 to 10000)
- ***Line 2***: N (number of patients-between 1 to 1000)
- ***Line 3***: the type of distance metric (1: Manhattan distance, 2: Euclidean distance, 3: Supremum distance, 4: Cosine similarity)
- ***Line 4***: X, the number of principal components in PCA to use to transform the original dataset to a new space.
  - If $X = -1$, you should **NOT** run PCA but only measure the distance metrics on the original dataset.
  - If $X \neq -1$, you first need to apply PCA on the input dataset, transform it to the new space via the first X components, and then measure the distance metrics on the transformed dataset.
- ***Line 5***: Patient P data that contains D integers
- ***Line 6 to 6+N***: The original dataset— each line contains D integers for each patient

4.1.2 Output format

- ***Line 1-5***: the index of the 5 most similar patients to patient P (corresponding index number in the dataset that would be between 1 to N) based on the input distance metric (The most similar patient would be first).
  - If there are two patients that are equally distant from patient P, you should put that patient with the lower index first.
- ***Line 6***: the cumulative amount of explained variance by the first X number of components.
  - *Note*: This line should be present in the output ONLY IF the input asked to run PCA on the dataset (i.e. Line 4 $\neq$ -1)

4.1.3 PCA Package

You are allowed to use any PCA package for this question as long as its answers are correct. However, given that some PCA implementations might use approximation algorithms, we STRONGLY suggest to use the sklearn package in Python that is a standardized package, if you know python:

http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

4.2 HackerRank Challenge (Steps 1,2 — Without PCA)

Your code will be evaluated on HackerRank for the steps 1 & 2 that involve measuring the similarity metrics on the original dataset without running PCA (i.e."line 4 = -1" in all the input test cases on HackerRank).

- The HackerRank link will be posted on the wiki page.
- Check the HackerRank page for sample inputs and outputs and updated instructions.
- The libraries that you use for this part (Steps 1,2) should be supported by the HackerRank environment.
- If the PCA package you are using is not supported by HackerRank (you will get compile errors, if so), simply comment the PCA parts, and then run your code on HackerRank. We will only evaluate your code for steps 1 & 2 (i.e. the input line 4 = -1) on HackerRank. The evaluation for steps 3 & 4 (PCA part) is done by the analysis part.

4.3 Analysis (Steps 3,4 — With PCA)

In this section, your code will be evaluated for steps 3 & 4 that involve running PCA on the original dataset, transforming the original dataset to a new dataset, measuring the similarity metrics on the transformed dataset, and then report the results. Follow the below steps:

- Use the the file named "Q4-analysis-input.in" (in the HW1-data.zip folder) as the input to run your code (this file follows the input format described in the Implementation section and includes a dataset of 100 patients with 10,000 dimensions)
- Vary line 3 to run the code with all four types of the similarity metrics
- Vary line 4 to set the number of PCA components to $X \subset \{1000, 100, 10, 2, 1\}$ for each similarity metric. Also, run the code when $X = -1$ (i.e. no PCA)

Now, report the following information in your pdf file:

1. A chart that shows the cumulative amount of explained variance by the first X number of components. Discuss how reducing the number of components used to transform data to a new space impacts the cumulative explained variance of the original dataset.

   Note: The explained variance might differ between PCA packages a bit. As long as your chart contains an approximation of explained variance, your answer is acceptable.

2. Use Table 3 to report the index of the five most similar patients to patient P in the original dataset and in the transformed dataset after running PCA for the aforementioned variations of X (i.e. number of components) for each similarity metric. Each cell of the table should include 5 numbers.

Table 3: The five most similar patients to patient P

	Manhattan	**Euclidean**	**Supremum**	**Cosine**
Original Dataset (X=-1)				
X=1000				
X=100				
X=10				
X=2				
X=1				

3. Compare the cells of Table 3 with each other for each distance metric and discuss how reducing the number of components for data transformation impacts PCA's effectiveness in finding the top five similar patients to patient P.

    - If you need to reduce the number of dimensions of the original dataset (i.e. 10,000) to gain a dataset that is much smaller in volume, yet closely maintains the integrity of the original data and provides the top similar patients to a patient with a fairly high accuracy, how many dimensions (1000,1000,10,2 or 1) would you suggest to use for transformation?

4.4 Code Submission

- In addition to submitting your code on HackerRank to evaluate steps 1-2, you need to submit your code, along with the pdf report, to Compass that we can test the PCA part as well. Name your code as NetId-HW1-question4.* (your code should be one single file).

- You will receive NO marks for the analysis section if you do not submit a code. We will also run a plagiarism detection software on all the codes.