CS412: Introduction to Data Mining **Summer 2016**

Assignment 1 (Final Version)

Due: 06/15/2016 11:59pm

General Instruction

- Errata: After the assignment is released, any further corrections of errors or clarifications will be posted at the Errata page at Piazza. Please watch it.
- Feel free to talk to other members of the class in doing the homework. We are more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down the solution yourself.
- Try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- For each question, you will **NOT** get full credit if you only give out a final result. Necessary calculation steps and reasoning are required.
- For a good balance of cognitive activities, we label each question with an activity type:
  - **L1 (Knowledge)** Definitions, propositions, basic concepts.
  - **L2 (Practice)** Repeating and practicing algorithms/procedures.
  - **L3 (Application)** Critical thinking to apply, analyze, and assess.

Assignment Submission

- Please submit your work before the due time. **We do NOT accept late homework!**
- We will be using Compass for collecting the homework assignments. Please submit your answers 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.
- The homework **MUST** be submitted in **pdf** format. **Scanned handwritten and hand-drawn pictures** inside your documents **are not acceptable**. Answers to the written part and mini-MP should be included in one `.pdf` file.
- Please **DO NOT** zip the PDF file so that graders can access your PDF directly on Compass. You can compress other files into a single zip file. In summary, you need to submit one PDF file, named as `hw1.netid.pdf`, and one `.zip` file, named as `hw1.netid.zip`.

- If scripts are used to solve problems, you are required to submit the source code, and use the file names to identify the corresponding questions or sub-questions. For instance, `question1.netid.py` refers to the python source code for Question 1; and `question1a.netid.py` refers to the python source code for sub-question 1(a); replace `netid` with your netid. You can submit separate files for sub-questions or a single file for the entire question.

Dataset

- The data set file, data.zip, can found in the course website.

Question 1 (15 points)

In the following questions, you are required to determine the various statistical descriptions for the provided data.

Purpose

- Have a better understanding of basic statistical descriptions of data.

Requirements

- For sub-questions (a), and (b) you should write important steps and the result in the PDF file you will submit. Only giving a result will not get credits.
- For sub-question (c), and (e), you should explain clearly in the PDF file.
- For sub-question (d), you should write a script to calculate. There's no restrictions on the language you use. You are required to submit your source code for sub-question (d) along with the answers in the PDF file.

0	1	2	4	5	5	7	10	10	12	13	17	39

Table 1: Data for Problem 1(a), 1(b), 1(c)

a. (3, L1) Determine the mean, variance (**population**), and standard deviation (**population**) of the data provided in table 1. (*Round your results to 3 decimal places.*)

b. (4, L2) What is the Interquartile range of the data provided in table 1? Based on this range identify the possible outliers.

c. (1, L1) What is the modality of the data in table 1?

d. (6, L2) The dataset `data.freeway.txt` (in `data.zip`) provides the speed (mph) and occupancy (%) measured by a physical sensor on the freeway *US101*. These records are sampled at every 5 minutes for a particular day starting at 00:00 hrs and ending at 23:59 hrs. Data in each row are separated by tabs. The first column shows the timestamp. The second column is speed and the third column is occupancy. Determine the Q1 (first quantile), Q3 (third quantile), median, mode, and mean of the **speed** attribute. If the result is not integer, then round it to 3 decimal places. You should write scripts to calculate statistical descriptions. There is no restrictions on the language you use. You are not allowed to calculate using calculators or by hands.

e. (1, L1) For the distribution of speed, is the data positively skewed or negatively skewed? Explain why you could get your conclusion by just looking at the measures determined in the previous part.

Question 2 (15 points)

In the following questions, you are required to evaluate the similarity/dissimilarity among data samples. If the result is not integer, then round it to 3 decimal places.

Purpose

- Have a better understanding of measuring data similarity and dissimilarity.

Requirements

- For sub-questions (a), (b), (c), (d) you should write important steps and the result in the PDF file you will submit. Only giving a result will not get credits.
- For sub-question (e), you should write a script to calculate. There's no restrictions on the language you use. You are required to submit your source code for sub-question (e).

a. (1, L1) Given two objects *Obj 1* and *Obj 2*, each of them has 200 binary attributes. Table 2 is the contingency table for these two objects. Each cell in the table shows the number of attributes where *Obj 1* and *Obj 2* have the corresponding combination of values. E,g., for cell *Obj 1* = 1 and *Obj2* = 0, there are 38 attributes with such a combination. Suppose all the attributes are **asymmetric** binary attributes, you are required to calculate the Jaccard coefficient of *Obj1* and *Obj2*. *(Round your result to 3 decimal points.)*

		Obj 2	
		1	0
Obj 1	1	42	38
	0	30	90

Table 2: Contingency table for Obj 1 and Obj 2

b. (3, L2) The records of three patients for various medical tests are indicated in table 3. For each row the first column is *name*, second column is *gender* which is a symmetric attribute, and the remaining attributes are asymmetric binary. If the dissimilarity is measured based only on asymmteric attributes then determine which two patients are most likely to have the same type of disease.

Name	**Gender**	**Fever**	**Cough**	**Test-1**	**Test-2**	**Test-3**	**Test-4**
Jack	M	Y	Y	P	P	N	N
Jim	M	Y	N	P	N	N	N
Mary	F	Y	Y	N	P	N	N

Table 3: Medical test reports for patients

c. (4, L2) For the data provided in table 4 construct the dissimilarity matrix between each pair of objects based on mixed attribute types. Test-1 consists of categorical variables and Test-2 consists of ordinal variables.

Identifier	**Test-1**	**Test-2**
	Categorical	*Ordinal*
1	type A	excellent
2	type B	fair
3	type C	excellent
4	type B	good

Table 4: Objects with mixed types of attributes

d. (3, L1) Given two points in the 3-D space, $A = (4, 4, 2)$ and $B = (-3, 2, 6)$. Please calculate the following distances between these two points.

1. *Euclidean* distance.
2. *Manhattan* distance.
3. *Minkowski* distance where $h = \infty$.

e. (4, L2) The dataset `home.txt` contained in `data.zip` consists of several locations in California with 14 different attributes such as percentage of homes that are 2/3/4 Bedrooms, etc. Each row consists of Geo-ID of the place, name of the place and the corresponding attributes (all separated by tab). You are interested in buying a home in Alto, California and would also like to see places which are very similar to Alto. Write a code to find the top 5 similar locations using the cosine similarity method described in the class. The data provided in table 5 can be used to build the feature vector of Alto.

Place	Alto, California
% of Homes Build 2000 to 2009	0
% of Homes Build 1990 to 1999	0
% of Homes Build 1980 to 1989	12
% of Homes Build 1970 to 1979	34
% of Homes Build 1960 to 1969	14
% of Homes Build 1950 to 1959	21
% of Homes Build 1940 to 1949	13
% of Homes Build 1930 to 1939 or earlier	5
% of Homes with 0 bed room	0
% of Homes with 1 bed room	30
% of Homes with 2 bed rooms	41
% of Homes with 3 bed rooms	23
% of Homes with 4 bed rooms	3
% of Homes with 5 or more bed rooms	2

Table 5: Properties of homes in Alto, California

Question 3 (10 points)

Purpose

- Understand the intuition and usage of data normalization.

Requirements

- Provide the normalized data table for (a) in the PDF file
- Write a script to normalize the data using z-score normalization for (b), in any language of your choice. You need to include the script file in your submission.

a. (5, L2) Table 6 shows a sample data from an employee database. You are required to normalize both age and salary in the range [0.0, 1.0] using the min-max normalization technique.

Identifier	**Age**	**Salary**
1	27	19,000
2	51	64,000
3	52	100,000
4	33	55,000
5	45	45,000

Table 6: Employee database

b. (5, L2) Based on the file `data.freeway.txt` contained in `data.zip`, normalize the speed(mph) using z-score normalization (use **population standard deviation**). Compare the mean and variance before and after normalization. For original speed of 55 mph, what is the corresponding speed after normalization?

Question 4 (30 points)

Purpose

- Understand the intuition and usage of Pearson correlation coefficients and Principal Component Analysis (PCA).

Requirement

- Apply the algorithms described in the lecture slides on the provided datasets.
- Give explanations based on your understanding of the algorithms

X	2.5	0.5	2.2	1.9	3.1	2.3	2	1.0	1.5	1.1
Y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

Table 7: 2D space dataset

I. Consider 10 data points (**population**) in 2-D space as specified in the table 7.

   a. (5, L2) What is the Pearson correlation coefficient between X and Y in the data set above? Show your calculations. What do you learn about the data set from the quantity?

   b. (2, L1) Based on the quantity and conclusion above, without actually applying PCA, can you guess if PCA may or may not help to reduce the data size? Explain your guess by the intuition of PCA.

   c. (3, L3) What is the covariance matrix for the data set above? Show your calculation.

II. Consider 4 data points (**population**) in the 2-d space: (1,1) (-1,0) (0,0), (0,-1)

   a. (10, L3) What is the first principal component (write down the actual vector)?

   b. (5, L3) What are the coordinates of the original data if we project in a 1-d space?

   c. (5, L3) From the projected 1-d space data if we try to reconstruct the original 2-d space data, how many points can we recover?

Machine Problems (30 points)

Requirements

- You can access the MPs for Assignment 1 using this link for Machine Problems.
- MP1 (15 points) and MP3 (15 points) are **required** to be solved by all students. The deadline for the required MPs is same as the Assignment.
- You should attach the output figures and codes (if asked to attach) for the required MPs (MP1 and MP3) in the same **PDF** file as your Assignment 1.
- MP2 and MP4 are *optional* and you will get extra credits for solving them. We have separate submission for optional MPs on Compass2G.
- The deadline for optional MPs (MP2 and MP4) is two days after the Midterm Exam (07/13/2016).
- Each MP webpage consists of instructions, input data and the expected output data.