CS412: Introduction to Data Mining **Fall 2015**

Assignment 1

Due: 09/21/2015 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.

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)

The dataset `data.online.scores` (in `data.zip`) provides the exam score records for students who take online courses. These records are sampled from a general population. Data in each row are separated by tabs. The first column shows students' IDs. The second column is students' midterm scores and the third column is students' final scores. Please give the following statistical descriptions of the **final scores**. If the result is not integer, then round it to 3 decimal places.

Purpose

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

Requirements

- For sub-questions (a), (b) and (c), 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. You are required to submit your source code for sub-questions (a), (b) and (c).
- For sub-question (d), you are required to answer the question in the PDF file you will submit.

a. (6') First quantile Q_1, the median, and the third quantile Q_3.

b. (3') Mean.

c. (3') Mode.

d. (3') For the distribution of students' final scores, is the data positively skewed or negatively skewed? Explain why you could get your conclusion.

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) 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), 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).

a. (3') Given two objects *Obj1* and *Obj2*, each of them has 200 binary attributes. Table 1 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 28 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*.

		Obj 2	
		1	0
Obj 1	1	21	28
	0	39	112

Table 1: Contingency Table for *Obj1* and *Obj 2*

b. (6') Given two points in the 3-D space, $A = (3, 1, 2)$ and $B = (-1, 0, 8)$. Please calculate the following distances between these two points.

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

c. (2') Suppose we have two random points A and B in space, explain why the *Euclidean* distance between A and B is always shorter than (or equal to) the *Manhattan* distance?

d. (4') Given the dataset `vectors.txt`, you will find two vectors (A and B). Each vector has 100 attributes (separated by tabs). Calculate the following distance between these two vectors:

   1. *Minkowski* distance where $h = 2$.
   2. *Minkowski* distance where $h = 3$.

Question 3 (10 points)

Based on the data of students' scores (file `data.online.scores`, contained in the file `data.zip`), normalize the mid-term scores using z-score normalization (use **empirical standard deviation** for standard deviation).

Purpose

- Understand the intuition and usage of z-score normalization.

Requirements

- Write a script to normalize the data using z-score normalization, in any language of your choice. You need to include the script file in your submission.

a. (5') Compare the mean and empirical variance before and after normalization.

b. (5') For original score of 90, what is the corresponding score 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 a toy dataset
- Give explanations based on your understanding of the algorithms
- Use Matlab, MS Excel, or similar software applications for calculation and visualization.

Consider 10 data points in 2-D space as specified in the table below.

X	0.69	-1.31	0.39	0.05	1.29	0.49	0.19	-0.81	-0.31	0.71
Y	0.89	-1.11	0.59	0.45	1.19	0.69	0.25	-0.71	-0.21	0.71

a. (5′) 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. (3′) 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. (6′) What is the covariance matrix for the data set above? Show your calculation. *Hint: It is easy to miss a few steps. Follow the steps described in the lecture slides.*

d. (6′) How many principal components does the dataset have? What are they? What is the first principal component, i.e., the most important one? Show your calculation. *Hint: You can use Matlab or similar software applications to find eigenvectors.*

e. (5′) Scatterplot all the data points and draw the lines showing the directions of all the principal components. *Hint: You may use Matlab or Excel to draw.*

f. (5′) Suppose we only use the first principal component, i.e., the most important component, as the basis for the new space. Project the data points $A = (0.05, 0.45)$ and $B = (0.49, 0.69)$ to the new space. Show your calculation. Draw the projections on the figure in sub-question (e).

Mini Machine Problem 1 (15 points)

This MP borrows a quite considerable amount of material from a certain source. We will publish the source after the submission's due date because it contains answers for a few questions. Don't try to find the existing answers because the MP is not hard, and it is really fun and useful. To finish the MP, please read this document carefully.
In this MP, we use the Matlab built-in data set **carsmall**, a data set containing information for 100 cars in 1970, 1976 and 1982. For this data set, we focus on 5 attributes: `Acceleration` (the rate of change of velocity of a car), `MPG` (Miles Per Gallon, fuel efficiency), `Displacement` (the volume of the cylinder), `Horsepower` and `Weight`. We use the `Cylinders` attribute (the number of cylinders) to group our observation. You will be required to run some code provided to you in this PDF file. **However, do not copy the code from this file to Matlab directly since the encoding mechanism for some special symbol in PDF is not supported by Matlab. You should type the code into Matlab.**

Purpose

- Learn the basic techniques for data visualization using Matlab.

Requirements

- This MP requires Matlab. *Please do not use other softwares since that will make the assignment harder for some questions.* The software is free for UIUC students in UIUC Webstore. And it is also available in EWS machines on campus. If you are not able to access both sources, please let us know ASAP. Please also note that it may take you only 1-2 hours to finish the MP, so if you don't often use the heavy Matlab software, you may want to use one of the EWS machines on campus.

- You should write all your answers (code, graphs and texts) in the PDF file you will submit. For code and graphs, you could paste them to the file.

1. Load the data **carsmall** in Matlab using the following code.

```
load carsmall
X = [MPG,Acceleration,Displacement,Weight,Horsepower];
varNames = {'MPG'; 'Acceleration'; 'Displacement'; 'Weight'; 'Horsepower'};
```

2. (2') Comet graph is an animated graph. To trace the data points on the screen for the `Displacement` attribute, we use the following code to visualize the `Displacement` attribute. Show the **final comet graph** in the PDF file you will submit by running the following code on Matlab.

```
comet(Displacement)
xlabel('Index of Car')
ylabel('Displacement')
```

3. (5') Drawing boxlpot is a popular way to visualize a distribution. The two whiskers show the Min observation and the Max observation. The central line shows the median. The edges of the box are the first quantile and the third quantile.

   a. (1') Run the following code on your Matlab to draw a boxplot for the `Acceleration` attribute. Show the **boxplot** in the PDF file you will submit.

```
boxplot(Acceleration)
ylabel('Acceleration')
```

   b. (4') Write code to visualize the `Acceleration` attribute using the boxplot for cars with different number of cylinders. In this graph, you group cars using the `Cylinders` attribute (the number of cylinders). For each group of cars, you draw a box to show the five-number summaries on `Acceleration`. All the boxes should be drawn on the same graph. In your graph, x-axis represents the number of cylinders and the y-axis shows the `Acceleration`. You should also add the label for x-axis (`Cylinders`) and y-axis (`Acceleration`). (*Hint: only several lines of codes are needed to finish this task. Try to use the boxplot(X,G) function where X is the attribute to be visualized and G is the grouping attribute.*) Show your **code and grouping boxplot** in the PDF file you will submit.

4. (4') 3-D scatter plots are popularly used to visualize 3 attributes at the same time.

   a. (2') Run the following code to draw a 3-D scatter plot. Show the **3-D plot** in the PDF file you will submit.

```
scatter3(Displacement,Cylinders,Horsepower,'filled','r')
xlabel('Displacement')
ylabel('Cylinders')
zlabel('Horsepower')
```

   b. (2') By observing the graph you get, could you identify a pair of correlated attributes? Could you explain why the positive or negative correlation makes sense? Give your **answer** in the PDF file you will submit. (*Hint: You could rotate the graph in Matlab when you try to find the correlation between two attributes on a 3-D graph.*)

5. (4') Interactive star plots are used to show the values of attributes for each observation. In each star (observation), the spoke length is proportional to the value of that attribute for that observation.

   a. (2') Run the following code. Show the **graph** in PDF file you will submit.

```
h = glyphplot(X(1:9,:), 'glyph','star', 'varLabels',varNames,...
'obslabels',Model(1:9,:));
set(h(:,3),'FontSize',8);
```

b. (2') In the Matlab figure dialog menu, there is a button called **data cursor** (See Figure 1. The data cursor item is in the red circle.) Based on the graph you get in 5a, if you click on the data cursor button, and then click on any star (car), you will get the value for each attribute of that car. Show the **value** of each attribute for the star (car) at the top left corner of the graph you plotted in the Question 5a in the PDF file you will submit.

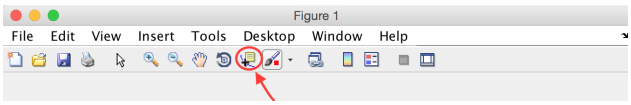

Figure 1: Matlab Figure Dialog Menu

Mini Machine Problem 2 (15 points)

This mini-MP asks you to play around with a few basic functionalities of Pentaho Kettle (Spoon) software to do data preprocessing for a customer table. In particular, you need to build a simple workflow that outputs pairs of last names that look similar. They currently belong to different customers, possibly because of mistakes of the employees when inputting the data.

Purpose

- Know how a tool specifically designed for data preprocessing may look like.
- Consider using the open-source software in your future work.

Requirements

- Do a few basic tasks with Spoon. You will have to install the software on your machine. In particular, you may want to watch Long's demonstration on the usage of Kettle Spoon in the video lecture on 09/08/2015. It shows basic things about Spoon, and you only need to combine and modify those things to finish this assignment.

You can download the software (≈ 800 MB) at http://community.pentaho.com/projects/data-integration/, and the tutorial about how to launch it at http://wiki.pentaho.com/display/EAI/02.+Spoon+Introduction.
As we will need MySQL, you need to copy `mysql-connector-java-5.1.36-bin.jar` to the `lib` folder of your kettle installation folder: http://dev.mysql.com/downloads/connector/j/
We will use `Sakila` sample database from MySQL. We are particularly interested in the `Customer` table. We uploaded it to our database, so you can use the database online, which means you do not need to install MySQL server. You do not need knowledge of MySQL to do this mini MP either.
To get started, open file `cs412_minimp1.ktr` in Spoon. You can find the file in the `data.zip` file on the assignment page of the course website. After opening it, you will see the following components:

- `ReadSource`: It downloads the table `Customers` from our online database. If you double click on the component, you will see it contains a SQL query to obtain the necessary information. The component is incomplete because it does not specify the connection. You will have to create a new one by clicking on "New...", and then enter the following information:
  - Host name: engr-cpanel-mysql.engr.illinois.edu
  - Database name: ltpham3_sakila
  - Port number: 3306
  - Username: ltpham3_cs412
  - Password: cs412kevin

  You can test the connection by clicking on "Test", or clicking on "Preview" after double clicking on the icon of the component.
- `Lkp_Lastname`: It downloads a list of last name. You also need to specify the connection you created for the component above.
- `MatchLastName`: It compares the last names from `Lkp_Lastname` with the last names from `ReadSource`. You feel free to choose one of the built-in algorithms for the comparison.

Your tasks are as followed:

1. (5') Your first task is to make `MatchLastName` works by specifying the flow of data from `ReadSource` to `Lkp_LastName`, as well as filling in necessary information in the component. We specified the flow of data from `ReadSource` to `MatchLastName` as an example. You will also notice that we specified the min value 0.8 and max value 0.99 in `MatchLastName`. Can you explain why the max value must be 0.99 rather than 1.0? You may try with max value 1.00 to see why we must do that.

2. (5') Search for "Filter Rows" component in the Design tab on the left, and drag it to the canvas. It helps you input rows from `MatchLastName` and output rows that satisfy your criteria. Specify the filter with necessary criteria so that it will output the rows containing information about the customers who have last names matching with those of someone else. Report the screenshots of the workflow and its output.

3. (5') The `score` column in the output above seems to be redundant. Search for component "Select values" in the `Design` tab on the left, and fill in necessary information to remove the `score` column. Report the screenshots of the workflow and its output.