CS412: An Introduction to Data Warehousing and Data Mining **Fall 2014**

Assignment 1

Handed In: 09/19/2014

- 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 your solution yourself. Please 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.
- 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 answers and proofs 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!**
- The homework can be submitted in pdf format. You are required to submit the source code, and use the file names to identify the corresponding questions. For instance 'Question1.netid.py' refer to the python source code for Question 1, replace netid with your netid. Compress all the files (pdf and source code files) into one file. Submit the compressed file **ONLY**.
- For each question, you will **NOT** get full credit if you only give out a final result. Necessary calculation steps are required.
- For Question 4, you can use any programming language you are comfortable with. Please add legends on the figure.
- All the data can be download from compass2g or the link.

1. Given a dataset, (file data.online.scores, contained in the file data.zip) which includes the records of students' exam scores (sample from the population) for the past few years of an online course. The first column students' id, the second column is the mid-term scores, and the third column is the final scores, and data are splitted by tab. Based on the dataset, give out the following statistical description of data. If the result is not integer, then round it to 3 decimal places.
   Give out the basic statiscal description about mid-term scores.

   a. (3') Max, min
   b. (3') First quartile Q1, median, third quartile Q3.
   c. (3') The mean score.

d. (3') The mode score.

e. (3') Empirical Variance.

2. Given the inventories of two supermarkets King Kullen (KK) and J Sainsbury (JS), compare the similarity between this two supermarkets by using the different proximity measures. if the result is not integer, then round it to 3 decimal places.

   a. (3') Given 200 items, the following table summarizes how many items are supplied by corresponding supermarket in Table 1. In Table 1, for KK = 0, JS = 0, it corresponds the number of items among the 200 items that are served neither by KK nor JS. For KK = 1, JS = 0, it corresponds the number of items among the 200 items that are served by KK but not JS. So on and so forth. Based on this table, calculate the Jaccard coefficent of J Sainsbury and King Kullen.

		J Sainsbury	
		0	1
King Kullen	0	41	58
	1	22	79

Table 1: Item supplement summary

   b. (9') Based on all items (treat the counts of the 100 items as a feature vector of the two supermarkets), (file data.supermarkets.inventories, contained in the file data.zip), calculate the minkowski distance of the two vectors with regard to different h values:

      1. h = 1.
      2. h = 2
      3. h = ∞.

   c. (3') The Cosine similarity between J Sainsbury and King Kullen with regard to the feature vector. (file data.supermarkets.inventories, contained in the file data.zip).

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

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

   b. (5') Given original score of 90, what is the corresponding score after normalization?

4. In the data folder, you are given the Iris dataset (iris.data, contained in the file data.zip) and description (iris.names, contained in the file data.zip). More information about this data set can be found in the description file. In this dataset, there are 150 data

points, and each one has 4 features. Run PCA to project the data down to a two dimensional subspace *spanned by the first two principle components*.

a. (5') produce a scatter plot of the data after projection. Make sure you plot each of the three classes differently (using color or different markers), and add legend. Report the scatter plot. (Hint: you may want to play with the data via different pre-processing methods first.)

b. (5') Compare the results under different choices of two dimensional spaces.

Note: You DO NOT need to implement PCA. Using existing packages is allowed.

5. Consider 3 data points in the 2-d space: (-1,1) (0,0) (1,-1).

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

b. (5') If we project the original data points into the 1-d subspace by the principal component you choose, what are their coordinates in the 1-d subspace? And what is the variance of the projected data?

c. (5') For the projected data you just obtained above, now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, what is the reconstruction error in L2?