CS412: An Introduction to Data Warehousing and Data Mining

Assignment 1

Handed In: 09/21/2017

- 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 both answers (fill in blank in compass) and details (via pdf) in 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 should 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 all the questions below, use degrees of freedom N - 1.

1. (32 points) Given a dataset, (file data.online.scores.txt) 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 statistical description about mid-term scores.

a. (8') Max, min

b. (12') First quartile Q1, median, third quartile Q3.

c. (4') The mean score.

d. (4') The mode score.

e. (4') Empirical Variance.

2. (25 points) Based on the data of students' score (file data.online.scores.txt). Please normalize the mid-term score using z-score normalization (divided by the empirical standard deviation).

a. (10') Compare the empirical variance before and after normalization.

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

c. (5') Pearson's correlation coefficient between midterm scores and final scores is:

d. (5') Covariance between midterm scores and final scores is:

3. (38 points) Given the inventories of two libraries Citadel's Maester Library (CML) and Castle Black's library(CBL), compare the similarity between this two libraries by using the different proximity measures. if the result is not integer, then round it to 3 decimal places.

a. (5') Given 200 books, the following table summarizes how many books are supplied by corresponding library in Table 1. In Table 1, for CBL = 0, CML = 0, it corresponds the number of items among the 200 items that are served neither by CBL nor CML. For CBL = 1, CML = 0, it corresponds the number of items among the 200 items that are served by CBL but not CML. So on and so forth. Based on Table 1, calculate the Jaccard coefficient of Citadel's Maester Library (CML) and Castle Black's library(CBL).

	Citadel's Maester Library (CML)		
Castle Black's library(CBL)		0	1
	0	20	120
	1	2	58

Table 1: Book supplement summary

b. (15’) For each kind of books, we have multiple copies. Based on all books (treat the counts of the 100 books as a feature vector of the two libraries), (file data.libraries.inventories.txt), calculate the minkowski distance of the two vectors with regard to different h values:

1. $h = 1$

2. $h=2$

3. $h = \infty$

c. (9’) The Cosine similarity between Citadel's Maester Library (CML) and Castle Black's with regard to the feature vector. (file data.libraries.inventories.txt).

d. (9’) Kullbac-Leibler divergence between Citadel's Maester Library (CML) and Castle Black's library(CBL) with regard to the feature vector. We denote that there are i_1 of book1 in Citadel's Maester Library (CML), and j_1 of book1 in Castle Black's library(CBL). Assume that someone will pick up a book randomly, the probability of this person to pick up book 1 in Citadel's Maester Library (CML) is i_1 / (i_1 + ... + i_100). Based on this probability distribution, calculate the Kullback–Leibler divergence of these two libraries P(CML || CBL).

4. (5 points) The Table 2 is a summary about customers’ purchase history of diapers and beer. Calculate the chi-square correlation value. If the result is not integer, then round it to 3 decimal places.

	Buy diaper	Do not buy diaper
Buy beer	150	40
Do not buy beer	15	3300

Table 2: Purchase history.