CS 412: Spring'21
Introduction To Data Mining

Assignment 1

(Due Thursday, February 25, 11:59 pm)

- Feel free to talk to other students of the class while doing the homework. We are more concerned that you learn how to solve the problems 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 Slack first if you have questions about the homework. You can also use CampusWire, send us e-mails, and/or come to our office hours. If you are sending us emails with questions on the homework, please cc all of us (Arindam, Carl, Jialong, and Qi) for faster response.
- 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 via Compass (`http://compass2g.illinois.edu`). Please do NOT email a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!
- The homework should be submitted in pdf format. You are required to submit source code, and use proper file names to identify the corresponding questions. For instance, 'Question1.netid.py' should 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. Please show the necessary details, calculation steps, and explanations as appropriate.
- All the data can be download from Compass2g.

1. (26 points) Consider the dataset (file: `data.online.scores.txt`) which contains the records of students' exam scores (sample from the population) for the past few years of an online course. The first column is a student's id, the second column is the mid-term score, and the third column is the finals score, and data are tab delimited. Based on the dataset, compute the following statistical description of the mid-term scores. If the result is not an integer, then round it to 3 decimal places.

   (a) (4 points) Maximum and minimum.

   (b) (9 points) First quartile Q1, median, and third quartile Q3.

   (c) (3 points) Mean.

   (d) (5 points) Mode.

   (e) (5 points) Variance.

2. (21 points) Based on the dataset of students' score (file: `data.online.scores.txt`), please normalize the mid-term scores using z-score normalization. We will refer to the original mid-term scores as midterm-original and the normalized mid-term scores as midterm-normalized. We will refer to the original finals scores as finals-original.

   (a) (3 points) Compute and compare the variance of midterm-original and midterm-normalized, i.e., the midterm scores before and after normalization.

   (b) (3 points) Given an original midterm score of 90, what is the corresponding score after normalization?

   (c) (5 points) Compute the Pearson's correlation coefficient between midterm-original and finals-original.

   (d) (5 points) Compute the Pearson's correlation coefficient between midterm-normalized and finals-original.

   (e) (5 points) Compute the covariance between midterm-original and finals-original.

3. (31 points) Given the inventories of two libraries Citadel's Maester Library (CML) and Castle Black's library (CBL) (file: data.libraries.inventories.txt), we will compare the similarity between the two libraries by using different proximity measures. The data for each library is for 100 books, and contains information on how many copies of each book each library has. When computing a similarity, if the result is not an integer, then round it to 3 decimal places.

   (a) (15 points) Each library has multiple copies of each book. Based on all the books (treat the counts of the 100 books as a feature vector for each of the libraries), compute the Minkowski distance of the vectors for CML and CBL with regard to different h values:

      (i) (5 points) $h = 1$.

      (ii) (5 points) $h = 2$.

      (iii) (5 points) $h = \infty$.

   (b) (8 points) Compute the cosine similarity between the feature vectors for CML and CBL.

   (c) (8 points) Compute the Kullback-Leibler (KL) divergence between CML and CBL by constructing probability distributions for each library based on their feature vectors. With i_1 denoting the count of Book 1 in a library, the probability of a person randomly

picking up Book 1 in that library is $\frac{i_1}{i_1+\ldots+i_{100}}$. The KL divergence will be computed based on these distributions for the libraries.

4. (22 points) Table 1 is a summary of customers' purchase history of diapers and beer. In particular, for a total of 3505 customers, the table shows how many bought both Beer and Diapers, how many bought Beer but not Diapers, and so on. For the problem, we will treat both Buy Beer and Buy Diaper as binary attributes.

	Buy Diaper	Do Not Buy Diaper
Buy Beer	150	40
Do Not Buy Beer	15	3300

Table 1: Contingency table for Beer and Diaper sales.

(a) (4 points) Calculate the distance between the binary attributes Buy Beer and Buy Diaper by assuming they are symmetric binary variables.

(b) (4 points) Calculate the Jaccard coefficient between Buy Beer and Buy Diaper.

(c) (7 points) Compute the χ^2 statistic for the contingency table.

(d) (7 points) Consider a hypothesis test based on the χ^2 statistic where the null hypothesis is that Buy Beer and Buy Diaper are independent. Can you reject the null hypothesis at a significance level of $\alpha = 0.05$? Explain your answer, and also mention the degrees of freedom used for the hypothesis test.