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.

Note: If intermediate steps are shown and there is a calculation error, -1 for that part.

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.
   (2 points) maximum = 100
   (2 points) minimum = 37

   (b) (9 points) First quartile Q1, median, and third quartile Q3.
   (2 points) median = 77
   (2 points) Q1 = 68
   (2 points) Q3 = 87
   (1 points) Sufficiently explain the calculation of the median. First, sort the data. The median is middle value or average of middle two values if $n \ (mod\ 2) \equiv 0$
   (2 points) Sufficiently explain the calculation of the quartiles. There are multiple ways to calculate the quartiles. Please see https://en.wikipedia.org/wiki/Quartile.

   (c) (3 points) Mean.
   (2 points) $\mu = 76.715$
   (1 points) Sufficiently explain the calculation of the mean. The mean is calculated as the sum of the data points divided by the number of samples. $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

   (d) (5 points) Mode.
   (4 points) mode = $[77, 83]$ (2 point each mode)
   (1 points) Sufficiently explain the mode. The mode is the most common data points.

   (e) (5 points) Variance.
   (3 points) $s^2 = 173.279$, $\sigma^2 = 173.106$
   (2 points) Sufficiently explain the calculation of the variance. The variance is the average squared difference from the mean. The following formulas are used (for sample and population, respectively):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

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.

(1 points) midterm-original variance: $s^2 = 173.27, \sigma^2 = 173.106$ (points for same as above)
(1 points) midterm-normalized variance: $s^2 = 1.000, \sigma^2 = 1.000$
(1 points) Compare: After normalization, the variance is approximately 1. This makes sense because we divide by the standard deviation in z-score normalization, so the new standard deviation is 1 and therefore also the variance.

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

$$v' = \frac{v - \mu}{\sigma} = \frac{90 - 76.715}{13.164} = 1.009$$

$$v' = \frac{v - \mu}{s} = \frac{90 - 76.715}{13.157} = 1.010$$

(1 points) Show the calculation or explain the steps.

(c) (5 points) Compute the Pearson's correlation coefficient between midterm-original and finals-original.
(3 points) Correct answer: $r = 0.544$ (both population and sample round to the same)
(2 points) Show the calculation or explain the steps.

$$r_{A,B} = \frac{\sum_{i=1}^{N}(A - \mu_A)(B - \mu_B)}{N\sigma_A\sigma_B}$$

(d) (5 points) Compute the Pearson's correlation coefficient between midterm-normalized and finals-original.
(5 points) Correct answer: r = 0.544 (both population and sample round to the same). Having the same as 2.c. will also be counted.
(2-4 points) Partially correct answer without correct results.

(e) (5 points) Compute the covariance between midterm-original and finals-original.
(3 points) Correct answer: $cov_{pop} = 78.176, cov_{sample} = 78.254$
(2 points) Show the calculation or explain the steps. Correct steps with a wrong answer receive full credit.

$$Cov(A, B) = \frac{\sum_{i=1}^{N}(A - \mu_A)(B - \mu_B)}{N}$$

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) $h = 1$.
Let p_i and q_i be the count vector for CML and CBL, respectively.
(3 points) Correct answer: 6152
(2 points) Show the calculation or explain the steps. **e.g.** $\sum_{i=1}^{100} |p_i - q_i|$

(ii) (5 points) $h = 2$.
(3 points) Correct answer: 715.328, round up (1 point)
(2 points) Show the calculation or explain the steps. **e.g.** $\sqrt{\sum_{i=1}^{100} (p_i - q_i)^2}$

(iii) (5 points) $h = \infty$.
(3 points) Correct answer: 170
(2 points) Show the calculation or explain the steps. **e.g.** $\max_{i=1}^{100} |p_i - q_i|$

(b) (8 points) Compute the cosine similarity between the feature vectors for CML and CBL.
(5 points) Correct answer: 0.841, round up (1 point)
(3 points) Show the calculation or explain the steps. **e.g.** $\frac{\sum_{i=1}^{100} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{100} p_i^2} \cdot \sqrt{\sum_{i=1}^{100} q_i^2}}$

(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.
(5 points) Correct answer: 0.207, round up (1 point)
(3 points) Show the calculation or explain the steps. **e.g.** $\mathrm{KL}(\mathrm{CML} \| CBL) = \sum_{i=1}^{100} p_i \log \frac{p_i}{q_i}$

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.
(3 points) Correct answer: 0.016
(1 points) Show the calculation or explain the steps.

$$distance = \frac{r + s}{q + r + s + t}$$

(b) (4 points) Calculate the Jaccard coefficient between Buy Beer and Buy Diaper.
(3 points) Correct answer: 0.732
(1 points) Show the calculation or explain the steps.

$$JC = \frac{q}{q + r + s}$$

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

(2 points) Correct answer for expected value of each item: Table 2.

(1 points) Show the calculation or explain the steps.

$$e_{ij} = \frac{count(A = a_i) * count(B = b_i)}{n}$$

(3 points) Correct answer for χ^2 statistic: Any number in [2450, 2470] is acceptable

(1 points) Show the calculation or explain the steps.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

	Buy Diaper	Do Not Buy Diaper
Buy Beer	8.94	181.05
Do Not Buy Beer	156.05	3158.94

Table 2: Expected value for each item.

(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.

(3 points) Correct answer: Reject

(2 points) Correct answer: 1 degree of freedom

(2 points) Explain the reason: we obtain $\chi^2 = 3.841$ by checking chi-square distribution table under the condition of $\alpha = 0.05$ and $dof = 1$. Because the computed χ^2 in (c) is much larger than 3.841, we reject the hypothesis