CS 412: Fall'22
Introduction To Data Mining

Assignment 1

(Due Wednesday, September 21, 11:59 pm)

- The homework is due at 11:59 pm on the due date. We will be using Gradescope for the homework assignments. You should join Gradescope using the entry code shared on Aug 3. Please do NOT email a copy of your solution. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!

- Please use Slack or Canvas first if you have questions about the homework. You can also come to our (zoom) office hours and/or send us e-mails. If you are sending us emails with questions on the homework, please start subject with "CS 412 Fall'22: " and send the email to *all of us* (Arindam, Mukesh, Chandni, Mayank, Hang, and Shiliang) for faster response.

- Please write down your solutions entirely by yourself and make sure the solutions are clear. The homework should be submitted in pdf format and there is no need to submit source code about your computing. You are expected to typeset the solutions. If your solution has any handwritten components, e.g., equations, tables, etc., please make sure they are legible—otherwise you may not get credit.

- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary steps and details, including any formulae you use, what the variables mean in the formulae, any derivation or calculation steps, and explanations as appropriate. Consider the following two examples:

Example 1 **Q:** Given a dataset $\mathcal{X} = \{3.1, 4.2, -1\}$, compute the mean.
A: For any set of n numbers $\mathcal{X} = \{x_1, \ldots, x_n\}$, the mean can be computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For the given dataset $\mathcal{X}$, the mean is $\mu = \frac{3.1+4.2-1}{3} = 2.1$

Example 2 **Q:** A coin claimed to be unbiased has been tossed 100 times, with 54 heads and 46 tails. What is the χ^2 statistic?
A: For a categorical variable taking k possible values, if the expected values are $e_i, i = 1, \ldots, k$ and the observed values are $o_i, i = 1, \ldots, k$, then the χ^2 statistic can be computed as: $S = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. For the problem, since the coin is claimed to be unbiased, the expected values are $50, 50$. Further, the observed values are $54, 46$. Then, the chi-squared statistic is given by $S = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

- All data for the assignmment can be download from Canvas (`https://canvas.illinois.edu/courses/30198`) Assignment 1

1. (25 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 final scores, and respond to associated questions. If the result is not an integer, then round it to 3 decimal places.

   (a) (5 points) Mean and Standard Deviation.

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

   (c) (4 points) Maximum and minimum.

   (d) (3 points) Mode.

   (e) (2 points) Is the median within one standard deviation of the mean? Justify your answer.

   (f) (2 points) Is (maximum - median) less than 1.5 times (Q3 - median)? Justify your answer.

2. (8 points) Consider the histogram of hourly pay (in dollars per hour) in a company called SkyNet (Figure 1). Approximately compute the median hourly pay at SkyNet using the histogram. Show the details of how you are doing the computation and clearly define any intermediate variables you use.

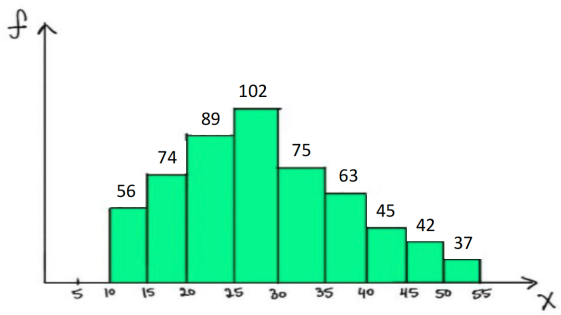

Figure 1: Histogram of hourly pay (in dollars per hour) at SkyNet.

3. (17 points) Consider the dataset of 1000 students' score (file: `data.online.scores.txt`) in a midterm exam (second column) and a final exam (third column). The first column is the student id and runs from 0 to 999. Please normalize the finals scores using z-score normalization. We will refer to the original finals scores as finals-original and the normalized finals scores as finals-normalized. We will refer to the original midterm scores as midterm-original.

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

   (b) (2 points) Given an original finals score of 90 for a student, what is the corresponding score after normalization?

   (c) (4 points) Compute the Pearson's correlation coefficient between midterm-original and finals-original.

   (d) (4 points) Compute the Pearson's correlation coefficient between midterm-original and finals-normalized.

   (e) (4 points) Compute the covariance between midterm-original and finals-original.

4. (26 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) (3 points) Let $d_h(CML, CBL)$ denote the Minkowski distance of the vectors for CML and CBL for a given $h \in \{1, 2, \infty\}$. Professor Rick Sanchez claims the following inequality to be true:

$$d_1(CML, CBL) \leq d_2(CML, CBL) \leq d_\infty(CML, CBL) \ .$$

   Do you agree with Professor Sanchez? Justify your answer.

   (c) (8 points) Compute the Kullback-Leibler (KL) divergence $D_{KL}(CML \| CBL)$ 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. (24 points) Table 1 is a summary of customers' purchase history of diapers and beer. In particular, for a total of 3500 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. (Be sure to include necessary intermediate steps, e.g., formulas, variable references, calculation results.)

	Buy Diaper	Do Not Buy Diaper
Buy Beer	150	40
Do Not Buy Beer	10	3300

Table 1: Contingency table for Beer and Diaper sales.

(a) (5 points) Calculate the distance between the binary attributes Buy Beer and Buy Diaper by assuming they are symmetric binary variables.

(b) (5 points) Calculate the Jaccard coefficient between Buy Beer and Buy Diaper.

(c) (6 points) Compute the χ^2 statistic for the contingency table.

(d) (8 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.