CS 412: Fall'21
Introduction To Data Mining

Assignment 1

(Due Thursday, September 23, 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.

- The homework is due at 11:59 pm on the due date. We will be using Gradescope for collecting the homework assignments. You should have been added to the Gradescope page for our class – if not, please email us. 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!

- Please use Canvas or Slack first if you have questions about the homework. You can also send us e-mails, and/or come to our office (zoom) hours. If you are sending us emails with questions on the homework, please start subject with "CS 412 Fall'21: " and send the email to *all of us* (Arindam, Yikun, Dongqi, Zhe, and Hang) for faster response.

- The homework should be submitted in pdf format and there is no need to submit source code about your computing. .

- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary 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: $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. 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 $\chi^2 = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

- All the data can be download from Canvas (`https://canvas.illinois.edu/courses/13790`) Assignment 1

1. (24 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. Solution: Sort the data array in descending order. The first element is Maximum and last element is Minimum. Results: max:100; min:37

   (b) (9 points) First quartile Q1, median, and third quartile Q3.
   Solution: Median is defined as the middle value if odd number of values, or average of the middle two values otherwise. The first quartile (Q1) is defined as the middle number between the smallest number (minimum) and the median of the data set; The third quartile (Q3) is the middle value between the median and the highest value (maximum) of the data set.
   results: 68, 77, 87

   (c) (3 points) Mean.
   Solution: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ mid: 76.715

   (d) (4 points) Mode.
   Solution: value that occurs most frequently in the data, (77, 83, count = 37)

   (e) (4 points) Variance.
   Solution: $\sigma = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$ or $\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$. (173.106 or 173.279)

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.

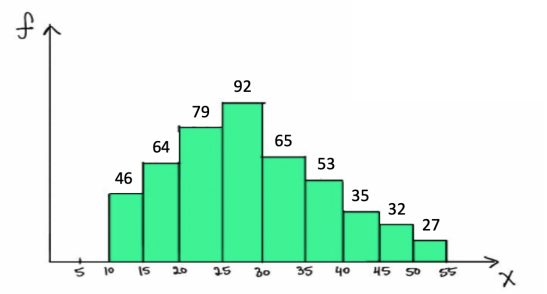

Figure 1: Histogram of hourly pay (in dollars per hour) at SkyNet.

Solution: Based on the definition of cumulative frequencies: $F_m = \sum_l^m f_l$ we have $F_1 = 46$, $F_2 = 110$, $F_3 = 189$, $F_4 = 281$, $F_5 = 346$, $F_6 = 399$, $F_7 = 434$, $F_8 = 466$, $F_9 = 493$. $n = 493$. Hence, the median falls into the 4-th bin, i.e., $[25, 30]$. The approximate median can be obtained by $median \approx L_m + \frac{n/2 - F_{m-1}}{f_m} \times (L_{m+1} - L_m)$ where L_m and L_{m+1} indicate low and high interval limits. We get $median \approx 25 + \frac{246.5 - 189}{92} \times 5 = 28.125$.

3. (18 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 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.

   Solution: [Step 1. Obtain the normalized data] For any dataset, the z-score normalization is $z = \frac{x-\mu}{\sigma}$, where x is the raw score to be normalized, μ is the mean of the data, and σ is the standard deviation. After we plug in the midterm-original data, we get the midterm-normalized data.

   [Step 2. Get the variance] For a sampled data set, the variance is computed as $\sigma^2 = \frac{1}{n-1}\sum_1^n(x_i - \bar{x})$ (or $\sigma^2 = \frac{1}{n}\sum_1^n(x_i - \bar{x})$). Then after we plug in the data, we get the variance of midterm-original: 173.106 and the variance of midterm-normalized: 1.000.

   (b) (3 points) Given an original midterm score of 90 (which is already in the dataset, e.g., student id 11), what is the corresponding score after normalization?

   Solution: [Step 1. z-score normalization equation] For any dataset, the z-score normalization is $z = \frac{x-\mu}{\sigma}$, where x is the raw score to be normalized, μ is the mean of the population, and σ is the standard deviation.

   [Step 2. Get the mean and the standard deviation] The mean and standard deviation of midterm-original are 76.715 and 13.157.

   [Step 3. Get normalized] After normalization, the raw midterm score 90 becomes: 1.010.

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

   Solution: For two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$, the correlation coefficient is expressed as $r_{\boldsymbol{A},\boldsymbol{B}} = \frac{Cov(\boldsymbol{A},\boldsymbol{B})}{\sigma_{\boldsymbol{A}}\sigma_{\boldsymbol{B}}}$.
   [Step 1. Get the covariance] $Cov(\boldsymbol{A},\boldsymbol{B})$ stands for the covariance that can be computed as $Cov(\boldsymbol{A},\boldsymbol{B}) = E((\boldsymbol{A}-\bar{\boldsymbol{A}})(\boldsymbol{B}-\bar{\boldsymbol{B}})) = E(\boldsymbol{A}\cdot\boldsymbol{B}) - \bar{\boldsymbol{A}}\bar{\boldsymbol{B}}$.

   [Step 2. Get the standard deviation] $\sigma_{\boldsymbol{A}}$ and $\sigma_{\boldsymbol{B}}$ are the standard deviation of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.

   [Step 3. Get the correlation] After we plug in midterm-original and finals-original, we get the correlation between midterm-original and finals-original: 0.544.

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

   Solution: The procedure is same as the question (c), the only difference is that we plug in midterm-normalized and finals-original to obtain the correlation between midterm-normalized and finals-original: 0.544.

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

   Solution: [Step 1. Get covariance] For two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$, the covariance $Cov(\boldsymbol{A},\boldsymbol{B})$ can be computed as $Cov(\boldsymbol{A},\boldsymbol{B}) = E((\boldsymbol{A}-\bar{\boldsymbol{A}})(\boldsymbol{B}-\bar{\boldsymbol{B}})) = E(\boldsymbol{A}\cdot\boldsymbol{B}) - \bar{\boldsymbol{A}}\bar{\boldsymbol{B}}$. After we plug in midterm-original and finals-original, we obtain the covariance between midterm-original and finals-original: 78.176.

4. (29 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) (7 points) Compute the cosine similarity between the feature vectors for CML and CBL.

   (c) (7 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.

Solution:

(a) (i) when $h = 1$, Minkowski distance is equivalent to Manhattan (or city block) distance (i.e., $d(i,j) = \sum_l |x_{i,l} - x_{j,l}|$) which is 6152.

(ii) when $h = 2$, Minkowski distance is equivalent to Euclidean distance (i.e., $d(i,j) = \sqrt{\sum_l |x_{i,l} - x_{j,l}|^2}$) which is 715.328.

(iii) when $h = \infty$, Minkowski distance is equivalent to the maximum difference between two vectors (i.e., $d(i,j) = \max_l |x_{i,l} - x_{j,l}|$) which is 170.

(b) The cosine similarity between two given vectors is defined as $cos(x_1, x_2) = \frac{x_1 \cdot x_2}{||x_1|| \times ||x_2||}$, where $\cdot$ denotes dot product and $||x||$ denotes the length of vector x. Based on it we get $cos(x_1, x_2) = 0.8414$.

(c) The discrete KL divergence is defined as $D_{KL}(p(x)||q(x)) = \sum_x p(x) \ln \frac{p(x)}{q(x)}$. In our case, the x-th entry of p and q (i.e., $p(x)$ and $q(x)$) denotes the probability of a person randomly picking up the x-th book in the corresponding library. Based on that we have $D_{KL}(CML||CBL) = 0.207$.

5. (21 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. (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	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.
Solution: Use $d(i,j) = \frac{(r+s)}{(q+r+s+t)}$ to get the result. Specifically, r and s are 40 and 15(order does not matter); $(q+r+s+t) = 3505$, which is the sum of all four numbers. The distance between the binary attributes Buy Beer and Buy Diaper is thus 0.016.

(b) (4 points) Calculate the Jaccard coefficient between Buy Beer and Buy Diaper.
Solution: Use $sim(i,j) = \frac{q}{(q+r+s)}$ to get the result. Specifically, q is 150; $(q+r+s) = 3505 - 3300 = 205$, since $t = 3300$. The Jaccard coefficient between Buy Beer and Buy Diaper is thus 0.732.

(c) (6 points) Compute the χ^2 statistic for the contingency table.
Solution: Directly apply the formula to get the result. $\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{o_{ij}}$. Within the formula, o_{ij} represents the actual value, e_{ij} represents the expected value. In this case. the (actual, expected) pairs should be (150, 8.94), (40, 181.06), (15, 156.06), (3300, 3158.94). The χ^2 statistic for the contingency table is thus 2468.183.

(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.
Solution: The value needed to reject null hypothesis at a significance level of $\alpha = 0.05$ is 3.841. This can be computed using the appropriate function of Scipy. Since 2468.183>3.841, we are able to reject the null hypothesis. The degrees of freedom used for the hypothesis test is 1 ($(2-1)^2$).