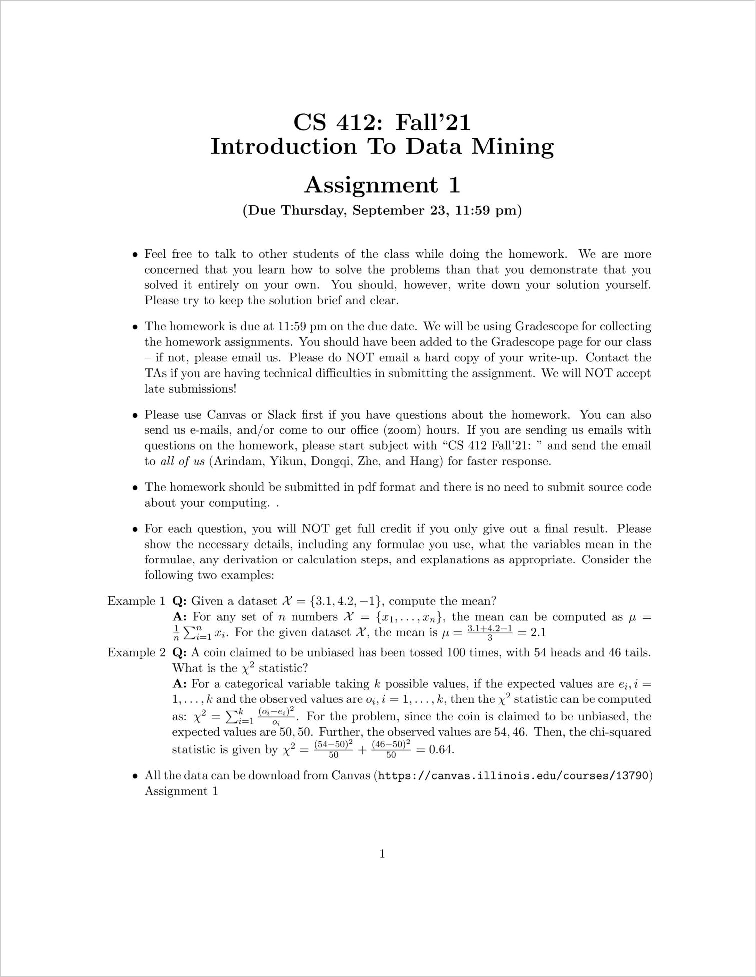
Assignment1

2021年9月11日 15:04



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.

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

   (c) (3 points) Mean.

   (d) (4 points) Mode.

   (e) (4 points) Variance.

(a) Maximum is the largest number in all mid-term scores, i.e., 100.

Minimum is the smallest one, i.e., 37.

(b) Q_n quartile is the value x such that at most $n/4$ of the data values are less than x and at most $4-n/4$ of the data values are more than x. And median is Q_2 quartile.

After traversing all mid-term scores we can obtain

$Q_1 = 68.0$, median $= Q_2 = 77.0$, $Q_3 = 87.0$

(c) For any set of n numbers $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$, the mean can be computed as $\frac{1}{n}\sum_{i=1}^{n} x_i$. For the mid-term scores, the mean is 76 715.

(d) Mode is the most frequent value in a sequence.

For the dataset, the mode is 77 and 83.

(e) The variance of N observations, $x_1, x_2, \ldots, x_N$ (when N is large), for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2$$

where $\bar{x}$ is the mean value

For this dataset, the variance is given by

$$\sigma^2 = \frac{1}{1000}\sum_{i=1}^{1000}(x_i - \bar{x})^2 \approx 173\ 279$$

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.

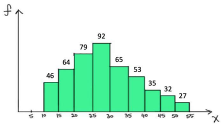


Figure 1: Histogram of hourly pay (in dollars per hour) at SkyNet.

We can easily approximate the median by the formula:

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) \times width$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval. and width is the width of the median interval

For the problem,

$$N = \sum freq = 46 + 64 + 79 + 92 + 65 + 53 + 35 + 32 + 27 = 493$$

so the median number locates in the 25 ~ 30 interval where $(\sum freq)_l = 46 + 64 + 79 = 189$, $freq_{median} = 92$, and width of this interval is 5 dollars per hour

Then we can calculate the approximate median,

$$\text{median} = 25 + \left(\frac{493/2 - 189}{92}\right) \times 5 = 28.125$$

Then we can calculgte the approximate median,

$$\text{median} = 25 + \left(\frac{493/2 - 189}{92}\right) \times 5 = 28.125$$

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.

   (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?

   (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-normalized and finals-original.

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

(a) The process of computating variance of midterm_nomalized is refer to the statement in Q1(e), then the variances of midterm_original and midterm_normalized are 173.279 and 1.

(b) In z-score normalization, the value for an attribute, A, are normalized based on the mean and standard deviation of A. A value, v_i, of A is normalized to v_i' by computing $v_i' = \frac{v_i - \bar{A}}{\sigma_A}$

in this problem, $v_i = 90$, according to question 1,

$\bar{A} = 76\ 715$. $\sigma_A = \sqrt{var\_mid\_ori} \approx 13.164$.

Then, for the given score, $V_{11}' = \frac{90 - 76.715}{13.164} \approx 1.009$

(c) The pearson's correlation coefficient is

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i, b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

where n is the number of tuples. a_i and b_i are the respective values of A and B in tuple i, $\bar{A}$ and $\bar{B}$ are the respective value of mean values, σ_A and σ_B are the respective standard deviations. $\Sigma(a_ib_i)$ is the sum of the AB cross-product.

(d) The formula is the same as which is in the (c). Therefore we can obtain the Pearson's correlation coefficient between mid-term normalized and finals-original is 0.544.

(e) The covariance between A and B is defined as

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

So we can obtain from the computation by codes, the covariance is 78 254.

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.

Let $i = (x_{i1}, x_{i2}, \cdots, x_{i100})$ and $j = (x_{j1}, x_{j2}, \cdots, x_{j100})$ be two objects denoting the inventories of books in two libraries respectively.

(a) Minkowski distance is a generalization of the Euclidean and Manhattan distances. It is defined as

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where h is a real number such that $h \geq 1$.

(i) $d(i,j)_{h=1} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots\cdots + |x_{i100} - x_{j100}|$

$= 6152$

(ii) $d(i,j)_{h=2} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots\cdots + (x_{i100} - x_{j100})^2}$

$= 715.328$

(iii) $d(i,j)_{h=\infty} = \lim_{h \to \infty} \left(\sum_{f=1}^{100} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = 170$

(b) suppose that x and y are the first two term-frequency denoting the feature vectors for CML and CBL.

we get $x \cdot y = \sum_{i=1}^{100} x_i \times y_i = 1344428$, $\|x\| = \left(\sum_{i=1}^{100} x_i^2\right)^{\frac{1}{2}} \approx 1229.637$

$\|y\| = \left(\sum_{i=1}^{100} y_i^2\right)^{1/2} = 1299\ 439$

$$sim(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{1344428}{1229.637 \times 1299.439} \approx 0\ 841$$

Therefore the cosine similarity between the feature vectors for CML and CBL is 0.841.

(c) Kullback-Leibler divergence of $q(x)$ from $p(x)$, denoted $D_{KL}(p(x), q(x))$.

Let $p(x)$ and $q(x)$ be two probability distributions of a discrete random variable x. That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0$ for any x in X.

$$D_{KL}(p(x) \| q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

$p(1) = \frac{i_1}{i_1 + i_2 + \cdots i_{100}}$

$q(1) = \frac{n_1}{n_1 + n_2 + \cdots n_{100}}$

For the problem, we can use $p(x)$ and $q(x)$ denote probability distributions for each library based on their feature vectors

To compute the $D_{KL}(P \| Q)$, we introduced a small constant $\varepsilon = 0.001$, and define a smoothed version of P and Q, P' and Q', as follows.

$P' : (a: p_1 - \frac{0.001}{100},\ b: p_2 - \frac{0.001}{100},\ \cdots\cdots)$

$Q' : (a: q_1 - \frac{0.001}{100},\ b: q_2 - \frac{0.001}{100},\ \cdots\cdots)$

We can conclude that $D_{KL}(P', Q')$ is 1508.069.

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.

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

(c) (6 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.

(a)

For symmetric binary attributes, each state is equally valuable. Dissimilarity that is based on symmetric binary attributes is called symmetric binary dissimilarity.

Suppose that i denotes "Buy Beer", j denotes "Buy Diaper", then the dissimilarity between i and j is

$$d(i,j) = \frac{r + s}{q + r + s + t}$$

		object j		
		1	0	sum
object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	

q is number of $(1,1)$, r is the number of $(1,0)$, s is number of $(0,1)$, t is the number of $(0,0)$

For this question, $d(i,j) = \frac{15 + 40}{150 + 15 + 40 + 3300} \approx 0.017$

(b) The Jaccard coefficient is the coefficient $sim(i,j)$ which can be computed as

$sim(i,j) = \frac{q}{\;} = 1 - d(i,j)$

(d) [illegible] degrees of freedom are $(2-1) \times (2-1) = 1$

(b) The Jaccard coefficient is[6] the coefficient $sim(i,j)$ which can be computed as

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)$$

For this question, we can obtain $sim(i,j) = \frac{150}{150+15+40} \approx 0.732$

(c) Let (A_i, B_j) denotes the joint event that attribute A takes on value a_i and attribute B takes on value b_j.

The χ^2 value is computed as

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

where o_{ij} is the observed frequency of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A=a_i) \times count(B=b_j)}{n}$$

where n is the number of data tuples.

	B_1	B_0	Total
A_1	150 (9)	40 (181)	190
A_0	15 (156)	3300 (3159)	3315
Total	165	3340	3505

Then,

$$\chi^2 = \frac{(150-9)^2}{9} + \frac{(40-181)^2}{181} + \frac{(15-156)^2}{156} + \frac{(3300-3159)^2}{3159}$$

$$\approx 2452.576$$

(d) For the 2×2 table, the degrees of freedom are $(2-1) \times (2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.05 significance level is $3.841 < 2452.576$, hence we can reject the null hypothesis.