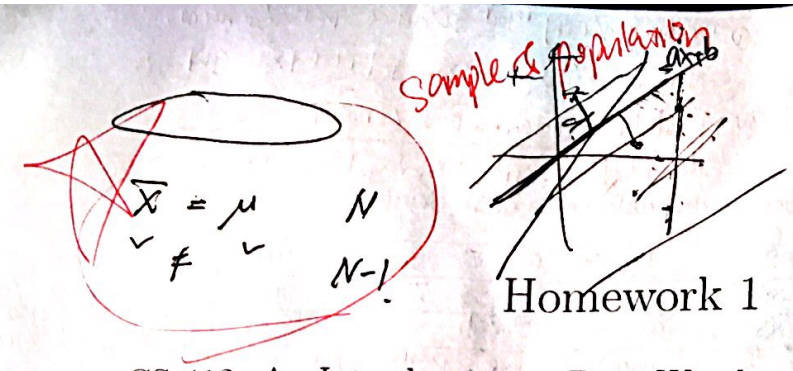

Homework 1

CS 412: An Introduction to Data Warehousing and Data Mining
Fall 2018

Handed In: Sep 20th, 2018

- Feel free to talk to other members of the class when doing the homework. We are more concerned that you learn how to solve the problem 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 Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- The homework is due at 11:59 PM on the due date. We will be using Compass (`http://compass2g.illinois.edu`) for collecting homework assignments. Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment. We do NOT accept late homework!
- The homework should be submitted in pdf format. If you use additional source code for solving problems, you are required to submit them and use the file names to identify the corresponding questions. For instance, 'Problem1.netid.py' refers to the python source code for Problem 1, replace netid with your netid. Compress all the files (pdf and source code files) into one zip file. Submit the compressed file ONLY. (If you did not use any source code, submitting the pdf file without compression will be fine)
- For each question, you will NOT get full credit if you only give out a final result. Necessary calculation steps are required. If the result is not an integer, round your result to 3 decimal places.

Problem 1. (28 points total)
Table 1 provides the information of 10 randomly sampled students' performances on midterm and final exams of an online course. Compute the following statistical properties for **both midterm scores and final scores**:

(a) (4 points) Max and min;

(b) (6 points) Mean, mode and median;

(c) (6 points) First quartile, third quartile and inter-quartile range;

(d) (6 points) Variance (sample & population);

Table 1: Midterm and Final Scores of 10 Students.

Student No.	1	2	3	4	5	6	7	8	9	10
Midterm	95	86	78	99	84	90	88	75	96	96
Final	88	88	90	95	85	77	99	80	100	80

(e) (6 points) Standard deviation (sample & population).

Problem 2. (20 points total)

Now we would like to normalize the data in Table 1 for further analyses.

(a) (6 points) After applying min-max normalization of **midterm scores**, find the normalized midterm scores of students No. 1, No. 2 and No. 3;

(b) (4 points) Find the variance (population) of min-max normalized midterm scores of all students;

(c) (6 points) After applying z-score normalization of **final scores**, find the normalized final scores of students No. 4, No. 5 and No. 6;

(d) (4 points) Find the variance (population) of z-score normalized final scores of all students.

Problem 3. (34 points total)

Next we are interested in the relationships between midterm scores and final scores in Table 1.

(a) (5 points) Find the covariance (population) between the midterm scores and final scores;

(b) (5 points) Find Pearson's correlation coefficient (use population standard deviation and covariance) between the midterm scores and final scores;

(c) (4 points) Assuming random variable M denotes midterm scores and random variable F denotes final scores, can you say M and F are independent based on the above results? Why?

(d) (8 points) Assuming $\boldsymbol{m} = [95 \quad 86 \quad \dots \quad 96]^T$ and $\boldsymbol{f} = [88 \quad 88 \quad \dots \quad 80]^T$ represent each student's midterm and final score respectively, compute (1) Manhattan distance, (2) Euclidean distance, (3) "supremum" distance and (4) cosine similarity of $\boldsymbol{m}$ and $\boldsymbol{f}$;

(e) (4 points) Select one distance measure from the four used in part (d), explain in plain English what that measure means in this scenario.

(f) (8 points) Do you think using Kullback-Leibler divergence will be a good choice as another distance measure in part (d)? If yes, show the Kullback-Leibler divergence from $\boldsymbol{m}$ to $\boldsymbol{f}$; if no, explain why. How about using Jaccard coefficient?

Table 2: Purchase history of beer and diapers.

	purchased diaper	not purchased diaper
purchased beer	200	80
not purchased beer	20	3000

Problem 4. (18 points total)
Table 2 shows 3,300 pieces of purchase history in a local market within one month on beer and diapers. We are interested in whether the purchase of these two items are correlated.

(a) (5 points) Calculate the χ^2 correlation value for "purchasing beer" and "purchasing diaper";

(b) (5 points) Based on the result you obtain in (a), do you think "purchasing beer" and "purchasing diaper" are independent or correlated? Feel free to refer to `https://www.medcalc.org/manual/chi-square-table.php` for values of the χ^2 distribution;

(c) (3 points) Assume a new customer comes and we do not have any information of his/her purchase history. Therefore, to estimate his/her purchase behavior, we use the previous data in Table 2. Let $\boldsymbol{p} = [p_0 \quad p_1 \quad p_2]^T$, where p_0, p_1, p_2 denote the probability that the new customer will purchase both beer and diaper, either beer and diaper and neither beer and diaper, respectively. Find $\boldsymbol{p}$;

(d) (5 points) Now assume we know that the new customer in part (c) has the following purchase behavior: $\boldsymbol{q} = [0.5 \quad 0.3 \quad 0.2]^T$. Find the Kullback-Leibler divergence value that represents the information loss when we use $\boldsymbol{p}$ to approximate $\boldsymbol{q}$.