CS 412: Introduction to Data Mining, Spring 2018

Quiz 1

- There are 5 problems total worth 36 points as shown in each question.
- You must not communicate with other students during this test.
- No books, notes allowed.
- No other electronic device except calculators are allowed. You cannot use your mobile as calculators.
- This is a 30 minute exam.
- For decimal numbers, report up to 2 point decimal points without rounding.
- Do not turn this page until instructed to.
- There are several different versions of this exam.

1. Fill in your information:

Full Name: ______________________________

NetID: ______________________________

1/1. (6 points) What are the best distance measures/coefficients for each of the following applications.

A. Driving distance between two locations in Downtown Chicago

B. Comparing patients based on their cancer medical test (positive or negative)

C. Finding the most similar document to a given text document by a search engine

Solution. Each 2 points

A. Manhattan Distance

B. Jaccard Coefficient (asymmetric binary data)

C. cosine similarity

1/2. (6 points) What are the best distance measures/coefficients for each of the following applications.

A. Comparing patients based on their cancer medical test (positive or negative)

B. Finding the most similar document to a given text document by a search engine

C. Driving distance between two locations in Downtown Chicago

Solution. Each 2 points

A. Jaccard Coefficient (asymmetric binary data)

B. cosine similarity

C. Manhattan Distance

1/3. (6 points) What are the best distance measures/coefficients for each of the following applications.

A. Driving distance between two locations in Downtown Chicago

B. Finding the most similar document to a given text document by a search engine

C. Comparing patients based on their cancer medical test (positive or negative)

Solution. Each 2 points

A. Manhattan Distance

B. cosine similarity

C. Jaccard Coefficient (asymmetric binary data)

2/1. (6 points) Consider the following data points: $(3, 6, 6), (1, 5, 8)$

Compute the following tuple (L_1, L_2, L_∞) for this pair of data points. Show the steps of calculation.

Solution. Each 2 points. Writing down the formulas would get half the credit.

A. $L_1 = |1 - 3| + |5 - 6| + |8 - 6| = 5$

B. $L_2 = \sqrt{(1-3)^2 + (6-5)^2 + (6-8)^2} = 3$

C. $max(|1 - 3|, |5 - 6|, |8 - 6|) = 2$

2/2. (6 points) Consider the following data points: $(2, 4, 7), (2, 5, 9)$

Compute the following tuple (L_1, L_2, L_∞) for this pair of data points. Show the steps of calculation.

Solution. Each 2 points. Writing down the formulas would get half the credit.

A. $L_1 = |2 - 2| + |4 - 5| + |9 - 7| = 3$

B. $L_2 = \sqrt{(2-2)^2 + (4-5)^2 + (9-7)^2} = \sqrt{5}$

C. $max(|2 - 2|, |5 - 4|, |9 - 7|) = 2$

2/3. (6 points) Consider the following data points: $(10, 0, 15), (9, 3, 10)$

Compute the following tuple (L_1, L_2, L_∞) for this pair of data points. Show the steps of calculation.

Solution. Each 2 points. Writing down the formulas would get half the credit.

A. $L_1 = |10 - 9| + |3 - 0| + |10 - 15| = 9$

B. $L_2 = \sqrt{(10-9)^2 + (3-0)^2 + (10-15)^2} = \sqrt{35}$

C. $max(|10 - 9|, |3 - 0|, |10 - 15|) = 5$

3/1. (6 points) The mean and median of a *moderately skewed* (asymmetrical) dataset are 10 and 20 respectively. Determine the approximate value for the mode (the dataset has only one mode).

Solution. for a moderately skewed dataset, we have the following empirical relation:
$mean - mode \simeq 3 \times (mean - median)$
$\Rightarrow mode \simeq 3 \times median - 2 \times meann$
$\Rightarrow mode \simeq 40$

3/2. (6 points) The mean and median of a *moderately skewed* (asymmetrical) dataset are 5 and 10 respectively. Determine the approximate value for the mode (the dataset has only one mode).

Solution. for a moderately skewed dataset, we have the following empirical relation:
$mean - mode \simeq 3 \times (mean - median)$
$\Rightarrow mode \simeq 3 \times median - 2 \times meann$
$\Rightarrow mode \simeq 20$

3/3. (6 points) The mean and median of a *moderately skewed* (asymmetrical) dataset are 15 and 20 respectively. Determine the approximate value for the mode (the dataset has only one mode).

Solution. for a moderately skewed dataset, we have the following empirical relation:
$mean - mode \simeq 3 \times (mean - median)$
$\Rightarrow mode \simeq 3 \times median - 2 \times meann$
$\Rightarrow mode \simeq 30$

4/1. (9 points) The following table shows the attendance and students' performance on an exam in a CS class. Conduct a correlation analysis between "attendance" and "exam performance" and report χ^2 value. Write down the steps you take as well.

	Pass	Fail	**Total**
Attended	25	5	30
Skipped	5	10	15
Total	30	15	45

Solution.

A. Writing the e_{ij} correctly: 2 points

Calculating the expected frequencies (e_{ij}) correctly: each e_{ij} 1 point

	Pass	Fail	**Total**
Attended	25 **(20)**	5 **(10)**	30
Skipped	5 **(10)**	10 **(5)**	15
Total	30	15	45

$$e_{ij} = \frac{count(A = a_{ij}) \times count(B = b_{ij})}{n}$$

$$e_{11} = \frac{count(Attended) \times count(Pass)}{45} = \frac{30 \times 30}{45} = 20$$

$$e_{21} = \frac{count(Skipped) \times count(Pass)}{45} = \frac{15 \times 30}{45} = 10$$

$$e_{12} = \frac{count(Attended) \times count(Fail)}{45} = \frac{30 \times 15}{45} = 10$$

$$e_{22} = \frac{count(Skipped) \times count(Fail)}{45} = \frac{15 \times 15}{45} = 5$$

B. Writing the χ^2 formula: 2 points Calculating the χ^2: 1 point $\chi^2 = \frac{(25-20)^2}{20} + \frac{(5-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(10-5)^2}{5} = 1.25 + 2.5 + 2.5 + 5 = 11.25$

4/2. (9 points) The following table shows the attendance and students' performance on an exam in a CS class. Conduct a correlation analysis between "attendance" and "exam performance" and report χ^2 value. Write down the steps you take as well.

	Pass	Fail	**Total**
Attended	75	15	90
Skipped	15	30	45
Total	90	45	135

Solution.

A. Writing the e_{ij} correctly: 2 points

Calculating the expected frequencies (e_{ij}) correctly: each e_{ij} 1 point

	Pass	Fail	**Total**
Attended	75 **(60)**	15 **(30)**	90
Skipped	15 **(30)**	30 **(15)**	15
Total	90	45	135

$$e_{ij} = \frac{count(A = a_{ij}) \times count(B = b_{ij})}{n}$$

$$e_{11} = \frac{count(Attended) \times count(Pass)}{135} = \frac{90 \times 90}{135} = 60$$

$$e_{21} = \frac{count(Skipped) \times count(Pass)}{135} = \frac{45 \times 90}{135} = 30$$

$$e_{12} = \frac{count(Attended) \times count(Fail)}{135} = \frac{90 \times 75}{135} = 30$$

$$e_{22} = \frac{count(Skipped) \times count(Fail)}{135} = \frac{45 \times 45}{135} = 15$$

B. Writing the χ^2 formula: 2 points Calculating the χ^2: 1 point $\chi^2 = \frac{(75-60)^2}{60} + \frac{(15-30)^2}{30} + \frac{(15-30)^2}{30} + \frac{(30-15)^2}{15} = 3.75 + 7.5 + 7.5 + 15 = 33.75$

4/3. (9 points) The following table shows the attendance and students' performance on an exam in a CS class. Conduct a correlation analysis between "attendance" and "exam performance" and report χ^2 value. Write down the steps you take as well.

	Pass	Fail	**Total**
Attended	50	10	60
Skipped	10	20	30
Total	60	30	90

Solution.

A. Writing the e_{ij} correctly: 2 points

Calculating the expected frequencies (e_{ij}) correctly: each e_{ij} 1 point

	Pass	Fail	**Total**
Attended	50 **(40)**	10 **(20)**	60
Skipped	10 **(20)**	20 **(10)**	30
Total	60	30	90

$$e_{ij} = \frac{count(A = a_{ij}) \times count(B = b_{ij})}{n}$$

$$e_{11} = \frac{count(Attended) \times count(Pass)}{90} = \frac{60 \times 60}{90} = 40$$

$$e_{21} = \frac{count(Skipped) \times count(Pass)}{90} = \frac{30 \times 60}{90} = 20$$

$$e_{12} = \frac{count(Attended) \times count(Fail)}{90} = \frac{60 \times 30}{90} = 20$$
$$e_{22} = \frac{count(Skipped) \times count(Fail)}{90} = \frac{30 \times 30}{90} = 10$$

B. Writing the χ^2 formula: 2 points Calculating the χ^2: 1 point $\chi^2 = \frac{(50-40)^2}{40} + \frac{(10-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(20-10)^2}{10} = 2.5 + 5 + 5 + 10 = 22.5$

5/1. (9 points) For the following data points, 10, 40, 45, 50, 55

A. If we normalize the dataset using z-score normalization, what would be the normalized value of 40?

B. If we normalize the dataset using z-score normalization that employs "mean absolute deviation" instead of "standard deviation", what would be the normalized value of 40?

C. After analysing the data, we found that there is an outlier in this dataset (i.e., 10). Which of the above normalization techniques is more robust to the outlier? Why?

Solution. Each part 3 points.

A. In calculating the standard deviation for the z-score normalization, both sample (n-1) or population (n) standard deviation would work. Here, I used the population std.

mean = 40 std (population)= 15.81 std (sample)= 17.67

$$\frac{40 - mean}{std} = 0$$

B. MAD = 12

$$\frac{40 - mean}{MAD} = 0$$

C. The second option (using MAD) because when computing the mean absolute deviation, the deviations from the mean (i.e., $|x_i - mean|$) are not squared; hence, the effect of outliers is somewhat reduced.

5/2. (9 points) For the following data points, 5, 25, 30, 35, 40

A. If we normalize the dataset using z-score normalization, what would be the normalized value of 30?

B. If we normalize the dataset using z-score normalization that employs "mean absolute deviation" instead of "standard deviation", what would be the normalized value of 30?

C. After analysing the data, we found that there is an outlier in this dataset (i.e., 5). Which of the above normalization techniques is more robust to the outlier? Why?

Solution. Each part 3 points

A. In calculating the standard deviation for the z-score normalization, both sample (n-1) or population (n) standard deviation would work. Here, I used the population std.

mean = 27 std (population)= 12.08 std (sample)= 13.5

$$\frac{30 - mean}{std} = 0.24$$

B. MAD = 9.6

$$\frac{30 - mean}{MAD} = 0.31$$

C. The second option (using MAD) because when computing the mean absolute deviation, the deviations from the mean (i.e., $|x_i - mean|$) are not squared; hence, the effect of outliers is somewhat reduced.

5/3. (9 points) For the following data points, 10, 35, 45, 55, 65

A. If we normalize the dataset using z-score normalization, what would be the normalized value of 35?

B. If we normalize the dataset using z-score normalization that employs "mean absolute deviation" instead of "standard deviation", what would be the normalized value of 35?

C. After analysing the data, we found that there is an outlier in this dataset (i.e., 10). Which of the above normalization techniques is more robust to the outlier? Why?

Solution. Each part 3 points

A. In calculating the standard deviation for the z-score normalization, both sample (n-1) or population (n) standard deviation would work. Here, I used the population std.

mean = 42 std (population)= 18.86 std (sample)= 21.09

$$\frac{35 - mean}{std} = -0.37$$

B. MAD = 15.6

$$\frac{35 - mean}{MAD} = -0.44$$

C. The second option (using MAD) because when computing the mean absolute deviation, the deviations from the mean (i.e., $|x_i - mean|$) are not squared; hence, the effect of outliers is somewhat reduced.