CS412: Introduction to Data Mining **Fall 2016**

Solution of Assignment 1

1. This dataset includes the records of students' exam scores (sampled from the population) for the past few years of an online course. For each row in the dataset, the first column corresponds to students' id, the second column is the mid-term score, and the third column is the final score. Each row is split by tab. Based on the dataset, estimate the following statistics. If the result is not an integer, round it to 3 decimal places.

   a. (10') Max, min

   b. (15') First quartile Q1, median, third quartile Q3.

   c. (5') The mean score.

   d. (5') The mode score.

   e. (5') Variance.

   f. (5') Do you think the model score consistent with empirical formula that (mean - mode = 3 * (mean - median))? (1 for yes, 0 for no).

Answer:

   a. Max = 100, Min = 26

   b. Q1 = 65, Q2 = 74, Q3 = 83

   c. Mean = 73.698

   d. Mode = 78

   e. Variance = 176.115 or 175.939

   f. 0 (no)
   You can refer to https://piazza.com/class/is2n71wakk671d?cid=63 for a detailed explanation.

2. Use the dataset provided in Question 1. Normalize the final scores using z-score normalization (divided by the empirical standard deviation). If the result is not an integer, round it to 3 decimal places.

   a. (8') Before normalization, the variance is:
   After normalization, the variance is:

   b. (4') For an original score of 90, the corresponding score after normalization is:

Answer:

a. Before normalization, the variance is: 116.227 or 116.111
After normalization, the variance is: 1

b. For an original score of 90, the corresponding score after normalization is: 0.241

3. Correlation analysis. Use the dataset provided in Question 1. If the result is not an integer, round it to 3 decimal places.

a. (4') Pearson's correlation coefficient between midterm scores and final scores is:

b. (4') Covariance between midterm scores and final scores is:

Answer:

a. Pearson's correlation coefficient between midterm scores and final scores is: 0.710
Note: The formula of Pearson's correlation coefficient is

$$\begin{aligned} r_{x,y} &= \frac{\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{(n-1)\sigma_x \sigma_y} \\ &= \frac{E(xy) - E(x)E(y)}{\sqrt{E(x^2) - E(x)^2} - \sqrt{E(y^2) - E(y)^2}} \end{aligned}$$

b. Covariance between midterm scores and final scores is: 101.601 or 101.499
Note: The formula to calculate covariance is

$$cov_{x,y} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

OR

$$cov_{x,y} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

4. Given the inventories of two supermarkets King Kullen (KK) and J Sainsbury (JS), compare the similarity between this two supermarkets by using the different proximity measures. if the result is not integer, then round it to 3 decimal places.

a. (5') Given 200 items, the following table summarizes how many items are supplied by each supermarket.
The Jaccard coefficient of J Sainsbury and King Kullen is:

	J Sainsbury		
		0	1
King Kullen	0	43	31
	1	19	107

Table 1: Item supplement summary

b. For the inventories of two supermarkets, each column consists of the number of specific items sold at each supermarket. Compare the similarity between the two supermarkets using the different proximity measures. If the result is not an integer, round it to 3 decimal places.

i. (15') Based on all items, what's the Minkowski distance of different h values:
   1. h = 1.
   2. h = 2
   3. h = ∞.

ii. (5') Cosine similarity:

iii. (5') KullbackLeibler divergence. We denote that there are i_1 of item1 in J Sainsbury, and j_1 of item1 in King Kullen. Assume that there is a customer who will pick up a product by random, the probability of this customer to pick up item 1 in J Sainsbury is i_1 / $(i_1 + ... + i_{100})$. Based on this probability distribution, calculate the KullbackLeibler divergence of these two supermarkets P(J Sainsbury || King Kullen):

Answer:

a. Jaccard = 0.682
   Note: The Jaccard index is

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{107}{19 + 31 + 107} = 0.682$$

b. For the inventories of two supermarkets, each column consists of the number of specific items sold at each supermarket. Compare the similarity between the two supermarkets using the different proximity measures. If the result is not an integer, round it to 3 decimal places.

   i. (15') Based on all items, what's the Minkowski distance of different h values:
      1. $h = 1, d = 6073$
      2. $h = 2, d = 708.615$
      3. $h = \infty, d = 170$ Note: The formulas of Minkowski Distance (for h = 1, h = 2) for is as following:

$$d_h = (\sum_{i=1}^{n} (|x_i - y_i|^h)^{1/h}$$

      When $h = \infty$,

$$d_\infty = max_i |x_i - y_i|$$

   ii. (5') Cosine = 0.844
      Note: The cosine similarity can be calculated as following:

$$consine(x,y) = \frac{xy}{\|x\|\|y\|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

   iii. (5') 0.204
      Note: Get the distribution of KK p_{KK} and JS p_JS, the KL Divergence can be obtained by

$$D(p_{JS} \| p_{KK}) = \sum_k ln(\frac{p_{JS}(i_k)}{p_{KK}(i_k)}) p_{JS}(i_k)$$

5. The following table is a summary of customers' purchase history of diapers and beer. Calculate the chi-square correlation value. If the result is not an integer, round it to 3

	buy diaper	do not buy diaper
buy beer	1346	430
do not buy beer	133	32974

decimal places.

Answer: 23593.796
Note:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

In which E_i is the expected value, O_i is the observed value.