CS412hw1

Kaiqi Zheng

Due: September 20 2018

1 Problem 1

a. mid term max is 99, min term min is 75
final max is 100, final min is 77

b. mid term mean is 88.7, mid term mode is 96, midterm medium is 89.0
final mean is 88.2, final mode is 80,88, final medium is 88.0

c. midterm Q1 is 84.5, Q3 is 95.75, range is 11.25.
final Q1 is 81.25, Q3 is 93.75, range is 12.5

d. sample variance is

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

population variance is

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

midterm sample variance is 65.12, population variance is 58.61
final variance is 63.96, population variance is 57.56

e. midterm standard deviation
(sample): 8.070 (population): 7.656
final standard deviation
(sample): 7.997 (population): 7.587

2 Problem 2

a. After minmax normalization, the midterm score of no.1 student is 0.833, no.2 student is 0.458, no.3 student is 0.125

b. The midterm's variance population is 0.102

c. After z-score normalization, the score of no.4 student is 0.896, no.5 student is -0.422, no.6 student is -1.476

d. Final's variance population is 1.0

3 Problem 3

a. The co-variance is 20.1778.

b. Pearson correlation coefficient is 0.31266.

c. They are not independent. If independent, then the following result must be 0:

$$cov(M, F) = E[(M - E[M])(F - E(F))] \tag{3}$$

But it is not. So they are independent.

d. The manhattan distance is 75
Eculidean distance is 28.30
supreme distance is 16.0
cosine similarity is 0.995

e. Supreme distance means which student has the largest difference between midterm and final score.

f. Kullback-Leibler divergence is a measure of similarity between two distributions, not a suitable distance measure. And it will therefore lose information about the relationship among each student's performance between midterm and final, while the four distance metrics above are able to capture.
Jaccard distance is not suitable too. Because it is mainly used for binary variance or multiple variance.

4 Problem 4

The observe value is [200,80,20,3000], the expected value is [19, 261, 201, 2819]

a. I am a little bit confused about the description of this question. Does ”the correlation value for “purchasing beer” and “purchasing diaper”; ” means only for x1 item, or for the whole?
   chi square only for x1(”purchasing diaper” and ”purchasing beer”) is

$$(200 - 19)^2/19 = 1724.263 \tag{4}$$

   chi square for all is

$$\frac{(200 - 19)^2}{19} + \frac{(80 - 261)^2}{261} + \frac{(20 - 201)^2}{201} + \frac{(3000 - 2819)^2}{2819} = 2024.396 \tag{5}$$

b. because the chisquare is so large, the p value is very small, then they must not be independent, thus they are correlated.

c. use data in Table 2, we can obtain
   p0 = 200/3300 = 0.061,
   p1 = (80+20)/3300 = 0.030,
   p2 = 3000/3300 = 0.909

d. the KL divergence when using p to approximate q is 1.179