CS 412: Spring'21
Introduction To Data Mining

Assignment 1

Xiangcan Li

1. (a) The maximum is the largest number in the dataset, while the minimum is the smallest number in the dataset. The largest number. The largest exam score is 100 and the smallest exam score is 37. Hence, the maximum is 100 and the minimum is 37.

   (b) First quartile Q1 is 25th percentile, the median is 50th percentile and third quartile Q3 is 75th percentile. Then the first quartile Q1 is 68.0, the median is 77.0 and the third quartile is 87.0.

   (c) The mean for the dataset $\{x_i\}_{i=1}^n$ is

   $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

   Then the mean is 76.715.

   (d) Mode is the value that appears most often in the dataset. Then the mode is 77 and 83.

   (e) The variance for the dataset $\{x_i\}_{i=1}^n$ is

   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

   Then the variance is 173.279.

2. (a) From previous exercise, we know that the variance of midterm-original is 173.279.
   To normalize a dataset $\{x_i\}_{i=1}^n$, for each x_i, we have

   $$\hat{x_i} = \frac{x - \text{mean}(\{x\})}{\text{std}(\{x\})}.$$

   Then after normalization, the variance is 1.000.

   The variance of midterm-original is relatively large, while the variance of midterm-normalized is very close to 1.

   (b) To normalize a data x_i, we have

   $$\hat{x_i} = \frac{x - \text{mean}(\{x\})}{\text{std}(\{x\})}.$$

   Then for an original midterm score of 90, s the corresponding score after normalization is 1.009.

(c) The Pearson's correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Then the Pearson's correlation coefficient between midterm-original and finals-original is 0.544.

(d) The Pearson's correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Then the Pearson's correlation coefficient between midterm-normalized and finals-original is 0.544.

(e) The covariance is

$$\text{cov} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

Then the covariance between midterm-original and finals-original is 78.254.

3. (a) The Minkowski distance of order h between two points $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is defined as:

$$D(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^h\right)^{\frac{1}{h}}.$$

(i) $h = 1 : D(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$.

Then the Minkowski distance of the vectors for CML and CBL with order 1 is 6152.

(ii) $h = 2 : D(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{\frac{1}{2}}$.

Then the Minkowski distance of the vectors for CML and CBL with order 2 is 715.328.

(iii) $h = \infty : \lim_{h\to\infty} \left(\sum_{i=1}^{n} |x_i - y_i|^h\right)^{\frac{1}{h}} = \max_{i=1}^{n} |x_i - y_i|$.

Then the Minkowski distance of the vectors for CML and CBL with order reaching infinity is 170.

(b) The cosine similarity between two vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is defined as:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}.$$

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

(c) For discrete probability distributions P and Q defined on the same probability space, $\mathcal{X}$, the Kullback-Leibler (KL) divergence from Q to P is defined to be

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

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 $\frac{i_1}{i_1 + ... + i_{100}}$.
Note that

$$\sum_{j=1}^{100} \frac{i_j}{i_1 + ... + i_{100}} = 1,$$

then the probability of a person randomly picking up Book j in that library consists of a discrete distribution.

Hence, the Kullback-Leibler (KL) divergence between CML and CBL by constructing probability distributions for each library is 0.207.

4. (a) The distance between the binary attributes Buy Beer and Buy Diaper by assuming they are symmetric binary variables is

$$d(\text{Buy Beer}, \text{Buy Diaper}) = \frac{r+s}{q+r+s+t}.$$

From the table, we have $q = 150, r = 40, s = 15, t = 3300$. Then the distance is 0.016.

(b) The Jaccard coefficient between Buy Beer and Buy Diaper is

$$\frac{q}{q+r+s}.$$

From the table, we have $q = 150, r = 40, s = 15, t = 3300$. Then the distance is 0.732.

(c) The test statistics is defined as

$$Q = \sum_{\text{cells}} \frac{(O-E)^2}{E},$$

where O denotes the observed cell frequency and E = row total $\cdot$ column total/grand total.
Then the $\mathcal{X}^2$ statistic for the contingency table is 2468.183.

(d) The degree of freedom d.f. is defined as

$$\text{d.f.} = (\text{Number of rows} - 1) \cdot (\text{Number of columns} - 1).$$

Then the degree of freedom is $(2-1) \times (2-1) = 1$.
Given $\alpha = 0.05$ and d.f. $= 1$, we have $\mathcal{X}_\alpha = 3.841$.

	$P(X \leq x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

We reject the null hypothesis when the test statistics is larger than or equal to $\mathcal{X}_\alpha$. We observe that $2468.183 > 3.841$. Hence, we reject the null hypothesis.