CS 412 Data Mining (22Sp): Assignment 1

Xu Ke (UIN: 675776713 NetID: kex5)

Jan. 28, 2022

1 Problem 1. True or False

1.1 Binary attribute is a type of continuous attribute.

$\boxed{False}$ Binary attributes are a special case of discrete attributes.(Lecture 2 Page 12)

1.2 For an unimodal curve that is not symmetric, empirically, its mean and median is always larger than its mode.

$\boxed{False}$ It should depend on whether left skewed distribution or right skewed distribution. (Lecture 2 Page 17)

1.3 Given a set of data points, its median is equals to its second quartile.

$\boxed{True}$ The second quartile should be 50^{th} percentile, thus identical to median. (Lecture 2 Page 22)

1.4 Given two real-valued vectors, if their histogram are exactly the same, these two vectors are the same.

$\boxed{False}$ Histogram graphs display of tabulated frequencies shown as bars for data in specific range, but data in the bars can vary as long as it is in the range. (Lecture 2 Page 24)

1.5 The similarity of two data points must be in the range of [0, 1].

$\boxed{False}$ Similarity often falls in the range [0,1], where 0 stands for no similarity and 1 stands for completely similar, but it depends on the similarity function chosen. (Lecture 2 Page 32)

1.6 Given a vector of real numbers, if a raw number in the vector is negative, its corresponding normalized value will be negative after z-score standardization.

$\boxed{False}$ Since $z = \frac{x-\mu}{\sigma}$, where x stands for raw score to be standardized, μ stands for mean of the population and σ stands for standard deviation, thus it will be negative when the raw score is below the mean, and positive when above. (Lecture 2 Page 34)

1.7 Given two random variables X_1 and X_2, if their covariance is 0, they are independent from each other.

$\boxed{False}$ If X_1 and X_2 are independent, $\sigma_{12} = 0$, but only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence. (Lecture 2 Page 53)

1.8 Given two data points, their L_1 distance is always no smaller than their L_∞ distance ("supremum" distance).

$\boxed{True}$ Since in L_1 norm all component will contribute to the result, while in L_∞ only the maximum difference will have any effect, which implies that the L_1 distance is always no smaller than their L_∞ distance ("supremum" distance) for the given two data points. (Lecture 2 Page 37)

1.9 KL divergence is a metric.

$\boxed{False}$ KL divergence is not a metric since it is not symmetric, i.e.the KL from p(x) to q(x) is generally not the same as the KL from q(x) to p(x). (Lecture 2 Page 58)

1.10 Principal Component Analysis (PCA) is a feature extraction method.

$\boxed{True}$ Principle Component Analysis (PCA) is a common feature extraction method in data science. Technically, PCA finds the eigenvectors of a covariance matrix with the highest eigenvalues and then uses those to project the data into a new subspace of equal or less dimensions. (Lecture 2 Page 87)

2 Problem 2. Basic Statistics and Normalization

Table 1: Midterm and Final Scores of 10 Students.

Student No.	1	2	3	4	5	6	7	8	9	10
Midterm	86	86	98	89	94	90	88	75	96	66
Final	92	77	100	85	95	92	87	92	100	70

2.1 What are mean, median and mode of midterm scores?

$$Mean = \frac{86+86+98+89+94+90+88+75+96+66}{10} = \boxed{86.8} \tag{1}$$

$$Median = \frac{88+89}{2} = \boxed{88.5} \tag{2}$$

$$Mode = \boxed{86} \tag{3}$$

2.2 What are first quartile, third quartile and inter-quartile range of midterm scores?

$$FirstQuartile : Q_1 = \boxed{86} \tag{4}$$

$$ThirdQuartile : Q_3 = \boxed{94} \tag{5}$$

$$Inter - QuartileRange : IQR = Q_3 – Q_1 = 94 - 86 = \boxed{8} \tag{6}$$

2.3 After min-max normalization on midterm scores, what are the normalized midterm scores? What are the sample variance and sample standard deviation of the normalized midterm scores?

According to the definition of Min-max normalization:

$$v' = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A \tag{7}$$

In this case we need to map [66,98] to [0,1], thus we will have $min_A = 66$,$max_A = 98$, $newmin_A = 0$ and $newmax_A = 1$. Then for the midterm scores for those 10 students, we will have:

$$v_1 = \frac{86-66}{98-66}(1-0) + 0 = \boxed{0.625} \tag{8}$$

$$v_2 = \frac{86-66}{98-66}(1-0) + 0 = \boxed{0.625} \tag{9}$$

$$v_3 = \frac{98-66}{98-66}(1-0)+0 = \boxed{1} \tag{10}$$

$$v_4 = \frac{89-66}{98-66}(1-0)+0 = \boxed{0.71875} \tag{11}$$

$$v_5 = \frac{94-66}{98-66}(1-0)+0 = \boxed{0.875} \tag{12}$$

$$v_6 = \frac{90-66}{98-66}(1-0)+0 = \boxed{0.75} \tag{13}$$

$$v_7 = \frac{88-66}{98-66}(1-0)+0 = \boxed{0.6875} \tag{14}$$

$$v_8 = \frac{75-66}{98-66}(1-0)+0 = \boxed{0.28125} \tag{15}$$

$$v_9 = \frac{96-66}{98-66}(1-0)+0 = \boxed{0.9375} \tag{16}$$

$$v_{10} = \frac{66-66}{98-66}(1-0)+0 = \boxed{0} \tag{17}$$

thus, the mean value of normalized midterm scores will be:

$$Mean : \mu = \frac{0.625+0.625+1+0.71875+0.875+0.75+0.6875+0.28125+0.9375+0}{10} = 0.65 \tag{18}$$

thus, the sample variance and sample standard deviation of the normalized midterm scores will be:

$$\begin{aligned} Variance : s = &\frac{(0.625-0.65)^2+(0.625-0.65)^2+(1-0.65)^2+(0.71875-0.65)^2+(0.875-0.65)^2}{10-1} \\ &+\frac{(0.75-0.65)^2+(0.6875-0.65)^2+(0.28125-0.65)^2+(0.9375-0.65)^2+(0-0.65)^2}{10-1} = \boxed{0.0924} \end{aligned} \tag{19}$$

$$StandardDeviation : \sigma = \sqrt{Variance} = \boxed{0.3040} \tag{20}$$

2.4 After z-score normalization on final scores, what are the normalized final scores? What are the population variance and population standard deviation of the normalized final scores?

According to the definition of Z-score normalization:

$$v' = \frac{v-\mu}{\sigma} \tag{21}$$

In this case, for the final scores, we will have:

$$Mean : \mu = \frac{92+77+100+85+95+92+87+92+100+70}{10} = 89 \tag{22}$$

$$\begin{aligned} Variance : s = &\frac{(92-89)^2+(77-89)^2+(100-89)^2+(85-89)^2+(95-89)^2}{10} \\ &+\frac{(92-89)^2+(87-89)^2+(92-89)^2+(100-89)^2+(70-89)^2}{10} = 83 \end{aligned} \tag{23}$$

$$StandardDeviation : \sigma = \sqrt{Variance} = 9.1104 \tag{24}$$

Then for the final scores for those 10 students, we will have:

$$v_1' = \frac{92-89}{9.1104} = \boxed{0.3293} \tag{25}$$

$$v_2' = \frac{77-89}{9.1104} = \boxed{-1.3172} \tag{26}$$

$$v_3' = \frac{100-89}{9.1104} = \boxed{1.2074} \tag{27}$$

$$v_4' = \frac{85-89}{9.1104} = \boxed{-0.4391} \tag{28}$$

$$v_5' = \frac{95-89}{9.1104} = \boxed{0.6586} \tag{29}$$

$$v_6' = \frac{92-89}{9.1104} = \boxed{0.3293} \tag{30}$$

$$v_7' = \frac{87-89}{9.1104} = \boxed{-0.2195} \tag{31}$$

$$v_8' = \frac{92-89}{9.1104} = \boxed{0.3293} \tag{32}$$

$$v_9' = \frac{100-89}{9.1104} = \boxed{1.2074} \tag{33}$$

$$v_{10}' = \frac{70-89}{9.1104} = \boxed{-2.0855} \tag{34}$$

Thus, the mean of normalized final scores will be:

$$\begin{aligned} Mean : \mu &= \frac{0.3293 + (-1.3172) + 1.2074 + (-0.4391) + 0.6585}{10} \\ &+ \frac{0.3293 + (-0.2195) + 0.3293 + 1.2074 + (-2.0855)}{10} = 0 \end{aligned} \tag{35}$$

Thus, the population variance and population standard deviation of the normalized final scores will be:

$$\begin{aligned} Variance : s &= \frac{(v_1'-0)^2 + (v_2'-0)^2 + (v_3'-0)^2 + (v_4'-0)^2 + (v_5'-0)^2}{10} \\ &+ \frac{(v_6'-0)^2 + (v_7'-0)^2 + (v_8'-0)^2 + (v_9'-0)^2 + (v_{10}'-0)^2}{10} \\ &= \frac{(0.3293)^2 + (-1.3172)^2 + (1.2074)^2 + (-0.4391)^2 + (0.6586)^2}{10} \\ &+ \frac{(0.3293)^2 + (-0.2195)^2 + (0.3293)^2 + (1.2074)^2 + (-2.0855)^2}{10} = \boxed{1} \end{aligned} \tag{36}$$

$$Standard Deviation : \sigma = \sqrt{Variance} = \boxed{1} \tag{37}$$

Alternatively, it is kind of tricky that actually z-score is to some extent identical to the normalized Gaussian distribution, which we've already learned in ECE313 and ECE314. Thus, after this step, the normalized data should satisfy the standard Gaussian distribution, which is N(0,1). That is to say, the normalized final scores have mean 0 and variance 1, which obviously implies standard deviation to be also 1.

3 Problem 3. Similarity

3.1 Given the Age and %Fat data of 6 adults (Table 2), please answer the following questions.

Table 2: *Age* and *%Fat* of 6 Adults.

No.	1	2	3	4	5	6
Age	22	25	35	43	49	53
%Fat	9.5	11.5	30.5	24.2	29.4	33.2

3.1.1 What is the sample covariance of Age and %Fat?

According to the definition, the sample covariance between X_1 and X_2 should be:

$$\hat{\sigma}_{12} = \frac{1}{n}\Sigma_{i=1}^{n}(x_{i1} - \hat{\mu_1})(x_{i2} - \hat{\mu_2}) \tag{38}$$

In this case for mean of Age and %Fat, we have:

$$Meanof Age : \mu_{Age} = \frac{22+25+35+43+49+53}{6} = \frac{227}{6} \tag{39}$$

$$Meanof\%Fat : \mu_{\%Fat} = \frac{9.5+11.5+30.5+24.2+29.4+33.2}{6} = 23.05 \tag{40}$$

Then, we will have the sample covariance of Age and %Fat, which is:

$$\begin{aligned}\hat{\sigma}_{Age\%Fat} = \frac{1}{5}\Sigma_{i=1}^{6}(x_{iAge} - \hat{\mu}_{Age})(x_{i\%Fat} - \hat{\mu}_{\%Fat}) = \frac{1}{5}((22 - \frac{227}{5})(9.5 - 23.05) \\ +(25 - \frac{227}{5})(11.5 - 23.05) + (35 - \frac{227}{5})(30.5 - 23.05) + (43 - \frac{227}{5})(24.2 - 23.05) \\ +(49 - \frac{227}{5})(29.4 - 23.05) + (53 - \frac{227}{5})(33.2 - 23.05)) = \boxed{114.4900}\end{aligned} \tag{41}$$

3.1.2 What is the sample correlation of Age and %Fat?

According to the definition, the sample correlation between X_1 and X_2 should be:

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \frac{\Sigma_{i=1}^{n}(x_{i1} - \hat{\mu_1})(x_{i2} - \hat{\mu_2})}{\sqrt{\Sigma_{i=1}^{n}(x_{i1} - \hat{\mu_1})^2\Sigma_{i=1}^{n}(x_{i2} - \hat{\mu_2})^2}} \tag{42}$$

In this case for sample variance of Age and %Fat, we have:

$$\begin{aligned}\hat{\sigma}_{Age} = \frac{1}{6}\Sigma_{i=1}^{6}(x_{iAge} - \hat{\mu}_{Age})^2 = \frac{1}{6}((22 - \frac{227}{6})^2 + (25 - \frac{227}{6})^2 \\ +(35 - \frac{227}{6})^2 + (43 - \frac{227}{6})^2 + (49 - \frac{227}{6})^2 + (53 - \frac{227}{6})^2) = 134.1889\end{aligned} \tag{43}$$

$$\begin{aligned}\hat{\sigma}_{\%Fat} = \frac{1}{6}\Sigma_{i=1}^{6}(x_{\%Fat} - \hat{\mu}_{\%Fat})^2 = \frac{1}{6}((9.5 - 23.05)^2 + (11.5 - 23.05)^2 \\ +(30.5 - 23.05)^2 + (24.2 - 23.05)^2 + (29.4 - 23.05)^2 + (33.2 - 23.05)^2) = 86.1958\end{aligned} \tag{44}$$

Thus, the sample correlation of Age and %Fat will be:

$$\hat{\rho}_{Age\%Fat} = \frac{\hat{\sigma}_{Age\%Fat}}{\hat{\sigma}_{Age}\hat{\sigma}_{\%Fat}} = \frac{95.4083}{\sqrt{134.1889 * 86.1958}} = \boxed{0.8873} \tag{45}$$

3.1.3 Based on your calulcation, are Age and %Fat correlated? Why? If so, are they positively correlated or negatively correlated?

$\boxed{Yes}$ the Age and %Fat are positively correlated since we have conclusion from the previous question that $\hat{\rho}_{Age\%Fat} = 0.8873$, which is clearly greater than 0, and thus positively correlated (Lecture 2 Page 55).

3.1.4 If we represent the data of adult No.2 and No.3 as two vectors $(25, 11.5)^T$ and $(35, 30.5)^T$, respectively, calculate their (1) Manhattan distance, (2) Euclidean distance and (3) "supremum" distance.

$$ManhattanDistance : L_1 = |25 - 35| + |11.5 - 30.5| = \boxed{29} \tag{46}$$

$$EuclideanDistance : L_2 = \sqrt{|25 - 35|^2 + |11.5 - 30.5|^2} = \boxed{21.4709} \tag{47}$$

$$SupremumDistance : L_\infty = \lim_{p \to \infty} \sqrt{|25 - 35|^p + |11.5 - 30.5|^p} = max(10, 19) = \boxed{19} \tag{48}$$

3.2 Suppose we are studying the relationship between (not) eating pizza and (not) drinking Pepsi. You have distributed questionnaires and the following results are collected from 165 guests at a restaurant, as shown in Table.3. Now, you conjecture that the guest who eats a piazza won't indicate that s/he also drinks Pepsi, which is our null hypothesis. Please answer the following questions using your knowledge on χ^2 calculation.

Table 3: Questionnaire results

	Eating pizza	Not eat pizza
Drink Pepsi	60	20
Not drink Pepsi	15	70

3.2.1 What is the expected value of a guest who eats pizza also drinks Pepsi?

Above all, we can extend our table shown as following and make null hypothesis that the two distributions are independent:

	Eating Pizza	Not Eat Pizza	Sum
Drink Pepsi	60 (X_1)	20 (X_2)	80
Not Drink Pepsi	15 (X_3)	70 (X_4)	85
Sum	75	90	165

Then we can have the following set of equations:

$$\begin{cases} X_1 : X_2 = X_3 : X_4 = 75 : 90 \\ X_1 : X_3 = X_2 : X_4 = 80 : 85 \end{cases} \tag{49}$$

Solve the equations above, we will have:

$$\begin{cases} X_1 = \dfrac{400}{11} = 36.3636 \\ X_2 = \dfrac{480}{11} = 43.6364 \\ X_3 = \dfrac{425}{11} = 38.6364 \\ X_4 = \dfrac{510}{11} = 46.3636 \end{cases} \tag{50}$$

Thus, the expected value of a guest who eats pizza also drinks Pepsi is $\boxed{36.3636}$

3.2.2 What is the expected value of a guest who eats pizza but doesn't drink Pepsi?

We can easily get the expected value of a guest who eats pizza but doesn't drink Pepsi is $\boxed{38.6364}$ (as above)

3.2.3 What is χ^2 value?

According to the definition of χ^2 test, we have:

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i} \tag{51}$$

Thus, in this case, we will have:

$$\chi^2 = \frac{(60 - 36.3636)^2}{36.3636} + \frac{(20 - 43.6364)^2}{43.6364} + \frac{(15 - 38.6364)^2}{38.6364} + \frac{(70 - 46.3636)^2}{46.3636} = \boxed{54.6499} \tag{52}$$

3.2.4 Does the null hypothesis stand at the confidence level of 0.001? Please justify your answer.

$\boxed{No}$ since according to the hint, in this case, we have Degrees of Freedom(DF) = 1, and from the given reference table, we know that for DF=1, when $\chi^2 = 54.6499$ which is much greater than 10.8280, thus the probability(P) should be much smaller than 0.001. Thus, We can reject the null hypothesis of independence at a confidence level of 0.001 (Lecture 2 Page 48), which implies that the null hypothesis fails and two distributions are correlated.

4 Problem 4. Data Cleaning

Table 4: Student Grades

ID	Level	GPA	Major	Grade
001	Graduate	3.7	CS	A
002	Graduate	3.0	CS	A
003	Undergraduate		ECE	A
-004		3.5	ECE	A

4.1 (True or False) Does the ID column contain noisy data? Please justify your answer.

$\boxed{True}$ It is clear that the data "-004" in ID column should be "004" instead since it is inconsistent with others.

4.2 For the 'Level' column, if we fill in the missing value with the value with highest probability, what would it be? Please justify your answer.

It would be $\boxed{Graduate}$ since for remaining data, we have probability $\frac{2}{3}$ for "Graduate" and only $\frac{1}{3}$ for "Undergraduate" in "Level" column. Thus, "Graduate" has the highest probability.

4.3 For the 'GPA' column, if we fill in the missing value with the attribute mean, what would it be? Please justify your answer.

It would be $\boxed{3.4}$ since we can compute it as follows easily:

$$Mean: \mu = \frac{3.7 + 3.0 + 3.5}{3} = 3.4 \tag{53}$$

4.4 Suppose the grade should be A if GPA $\leq$ 3.0. Does the 'Grade' column contain intentional data? Why?

$\boxed{Yes}$ The "Grade" for the student with ID "003" should be intentional data since actually the GPA for that student is missing and the "A" should be kind of default value for this student.

5 Problem 5. Principle Component Analysis (PCA)

5.1 What is the first principal component?

Suppose we have 3 data points in a 3-dimensional Euclidean space:

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \tag{54}$$

$$x_2 = \begin{bmatrix} -3 \\ -6 \\ -9 \end{bmatrix} \tag{55}$$

$$x_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{56}$$

In this case, we can start from m by n (in this case, 3 by 3) data matrix X, which is:

$$X = \begin{bmatrix} 1 & -3 & 0 \\ 2 & -6 & 0 \\ 3 & -9 & 0 \end{bmatrix} \tag{57}$$

then, we can take the mean of matrix X, namely $\hat{X}$, which is:

$$\hat{X} = \begin{bmatrix} \frac{1+(-3)+0}{3} \\ \frac{2+(-6)+0}{3} \\ \frac{3+(-9)+0}{3} \end{bmatrix} = \begin{bmatrix} -\frac{2}{3} \\ -\frac{4}{3} \\ -2 \end{bmatrix} \tag{58}$$

Then, we can recenter, that is subtract mean from each row of X, which is:

$$X_c = X - \hat{X} = \begin{bmatrix} 1 & -3 & 0 \\ 2 & -6 & 0 \\ 3 & -9 & 0 \end{bmatrix} - \begin{bmatrix} -\frac{2}{3} \\ -\frac{4}{3} \\ -2 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} & -\frac{7}{3} & \frac{2}{3} \\ \frac{10}{3} & -\frac{14}{3} & \frac{4}{3} \\ 5 & -7 & 2 \end{bmatrix} \tag{59}$$

Then, we can compute covariance matrix, $\Sigma = \frac{1}{m} X_c X_c^T$, which is:

$$\begin{aligned} \Sigma = \frac{1}{m} X_c X_c^T &= \frac{1}{3} \begin{bmatrix} \frac{5}{3} & -\frac{7}{3} & \frac{2}{3} \\ \frac{10}{3} & -\frac{14}{3} & \frac{4}{3} \\ 5 & -7 & 2 \end{bmatrix} \begin{bmatrix} \frac{5}{3} & \frac{10}{3} & 5 \\ -\frac{7}{3} & -\frac{14}{3} & -7 \\ \frac{2}{3} & \frac{4}{3} & 2 \end{bmatrix} \\ &= \frac{1}{3} \begin{bmatrix} \frac{26}{3} & \frac{52}{3} & 26 \\ \frac{52}{3} & \frac{104}{3} & 52 \\ 26 & 52 & 78 \end{bmatrix} = \begin{bmatrix} \frac{26}{9} & \frac{52}{9} & \frac{26}{3} \\ \frac{52}{9} & \frac{104}{9} & \frac{52}{3} \\ \frac{26}{3} & \frac{52}{3} & 26 \end{bmatrix} \end{aligned} \tag{60}$$

Then, we need to find the eigenvalues of Σ, which will satisfy $|M - \lambda I| = 0$:

$$\begin{bmatrix} \frac{26}{9} & \frac{52}{9} & \frac{26}{3} \\ \frac{52}{9} & \frac{104}{9} & \frac{52}{3} \\ \frac{26}{3} & \frac{52}{3} & 26 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} \frac{26}{9} - \lambda & \frac{52}{9} & \frac{26}{3} \\ \frac{52}{9} & \frac{104}{9} - \lambda & \frac{52}{3} \\ \frac{26}{3} & \frac{52}{3} & 26 - \lambda \end{bmatrix} = D \tag{61}$$

Then, we will have to solve $det(D) = 0$, and then we will get:

$$-\lambda^3 + \frac{364}{9}\lambda^2 = 0 \tag{62}$$

Thus, we will easily have:

$$\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = \frac{364}{9} \tag{63}$$

In order to get the first principal component, we will choose $\lambda = \frac{364}{9}$, which is the highest eigenvalue, plug it back into the equation above then we will have to solve:

$$\begin{bmatrix} \frac{26}{9} - \frac{364}{9} & \frac{52}{9} & \frac{26}{3} \\ \frac{52}{9} & \frac{104}{9} - \frac{364}{9} & \frac{52}{3} \\ \frac{26}{3} & \frac{52}{3} & 26 - \frac{364}{9} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 0 \tag{64}$$

Thus, we will get the eigenvector or namely the first principal component, which is:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \boxed{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}} \tag{65}$$

5.2 What is the difference between PCA and attribute subset selection?

Even though both PCA and attribute subset selection are techniques for dimensionality reduction, attribute subset selection is a kind of feature selection technique while PCA is a kind of feature extraction technique. More specifically, PCA transforms the data in the high-dimensional space to a space of fewer dimensions but attribute subset selection finds a subset of the original variables without transforming them.