Assignment 2

Lavinia Wang #1473704

Due Date: Saturday, *February 3rd, 2018, by midnight*

Total number of points: 35 points

Problem 1 (10 points): This problem is an example of data preprocessing needed in a data mining process.

Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a. (2 points) Draw the box-plots for age and %fat. Interpret the distribution of the data.

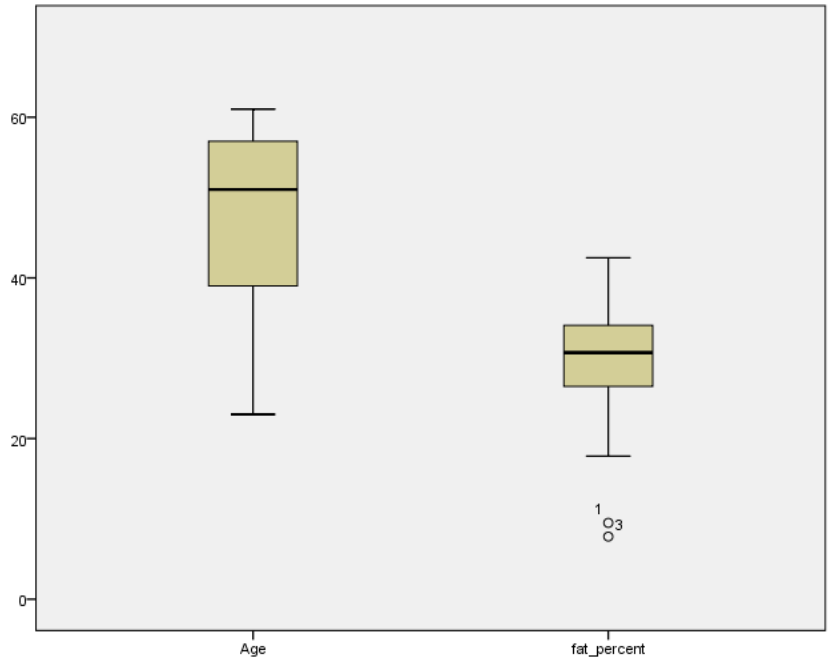

Figure 1: Boxplot of Age & Fat Percentage Distribution

From figure 1, we know that the minimum age is about 24 and the maximum age is roughly 61, giving a range of 37. The median of distribution is 51. The 4 sections of the box plot are uneven in size. Data fallen in the first quartile is much more than that in any of the second, third and fourth quartile. Thus, the distribution is slightly skewed to the left. There are no obvious outliers.

From the right plot, we know that the minimum fat percent is about below 10 (approximately 8) and the maximum fat percent is roughly 43, giving a range of 35. The median of distribution is 30. Data fallen in the first and forth quartile is much more than that in the second and third quartile. Thus, the distribution is

skewed to the left. There are two possible outliers marked 1 and 3. Fat Percent box plot is much lower than age and has smaller standard deviation.

b. (2 points) Normalize the two attributes based on z-score normalization.

Age	-1.77	-1.77	-1.47	-1.47	-0.56	-0.41	0.04	0.19	0.27
%fat	-2.03	-0.19	-2.25	-1.17	0.24	-0.30	-0.19	-0.19	0.24
Age	0.42	0.57	0.57	0.72	0.80	0.87	0.87	1.03	1.10
%fat	0.67	1.54	0.02	0.46	0.13	0.56	0.46	1.32	0.78

Table 1: Normalized age and body fat data

Note: $\mu(age) = 46.44$ $\sigma(age) = 13.22$ $\mu(\%fat) = 28.78$ $\sigma(\%fat) = 9.25$

c. (2 points) Regardless of the original ranges of the variables, normalization techniques transform the data into new ranges that allow to compare and use variables on the same scales. What are the values ranges of the following normalization methods? Explain your answer.

i. Min-max normalization

ii. Z-score normalization

iii. Normalization by decimal scaling.

i. Min-max normalization: The value range is [new_min_A, new_max_A] or commonly [0.0, 1.0] or [-1.0, 1.0] which preserves the relationships among the original data values.

ii. Z-score normalization: The value range is [-infinity, infinity] although it is very unlikely to get extreme values.

iii. Normalization by decimal scaling: The value range is [-1, 1] which moves the decimal point of values of attribute A.

d. (2 points) Draw a scatter-plot based on the two variables and interpret the relationship between the two variables.

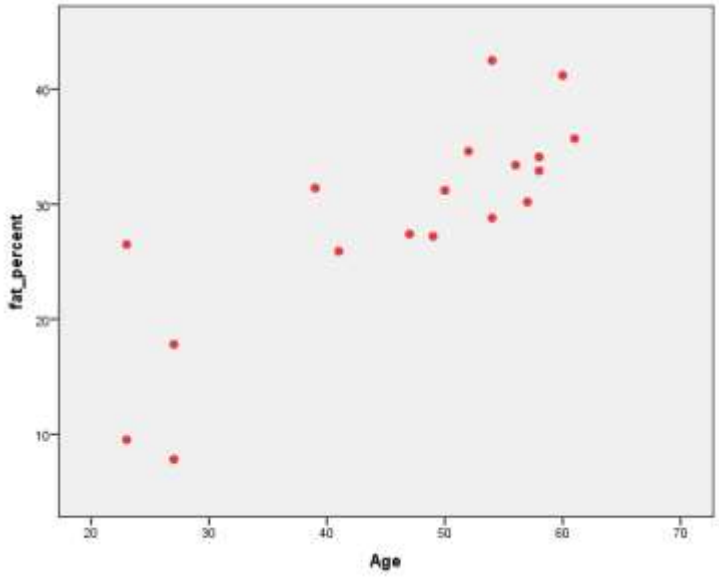

Figure 2: Scatterplot of Age vs Fat Percent (Original Values)

As age increases, fat percentage increases correspondingly. So we can conclude that there is a positive linear relationship between age and fat percentage indicating that the older person is in age, the more body fat percentage.

e. (2 points) Calculate the correlation matrix. Are these two attributes positively or negatively correlated? Calculate the covariance matrix. How is the correlation matrix different from the covariance matrix?

Correlations

		Age	fat_percent
Age	Pearson Correlation	1	.818**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	2970.444	1700.333
	Covariance	174.732	100.020
	N	18	18
fat_percent	Pearson Correlation	.818**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	1700.333	1455.945
	Covariance	100.020	85.644
	N	18	18

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2: Correlation and Covariance Table of Age and Fat Percent

From the correlation matrix, we find the correlation coefficient between the two variables is 0.818 which indicates high positive correlation.

Correlation matrix has a fixed range [-1, 1] where -1 means the two variables are inversely correlated and 1 means the two variables are positively correlated with 0 meaning no correlation. In covariance matrix, the minimum is 0 and no maximum bound. If the variables are not normalized meaning they have different scales, for instance, height and weight, in meters and kilograms respectively, you will get a different covariance from when you do it in other units.

Problem 2 (5 points): This problem is an example of data preprocessing needed in a data mining process.

Suppose a group of 12 sales price records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50,55,72,92,204,215

Partition them into bins by each of the following method, smooth the data and interpret the results:

a. (2.5 points) equal-depth partitioning with 4 values per bin

Mean is subject to outliers, choosing boundary, mean, median depends on each case

Original data	5,10,11,13	15,35,50,55	72,92,204,215
Smooth by mean	9.75, 9.75, 9.75, 9.75	38.75, 38.75, 38.75, 38.75	143.25, 143.25, 143.25, 143.25

Smooth by median	10.5, 10.5, 10.5, 10.5	42.5, 42.5, 42.5, 42.5	148, 148, 148, 148
Smooth by boundary	5, 13, 13, 13	15, 15, 55, 55	72, 72, 215, 215

Table 3: Equal-depth partitioning with 4 values per bin

a. (2.5 points) equal-width partitioning with 4 bins

215-5=210 210/4=52.5 [5, 57.5) [57.5, 110) [110, 162.5) [162.5, 215]

Original data	5,10,11,13,15,35,50,55	72, 92		204, 215
Smooth by mean	24.25, 24.25, 24.25, 24.25, 24.25, 24.25, 24.25, 24.25	82, 82		209.5, 209.5
Smooth by median	30, 30, 30, 30, 30, 30, 30, 30	82, 82		209.5, 209.5
Smooth by boundary	5, 5, 5, 5, 5, 55, 55, 55	72, 92		204, 215

Table 4: Equal-widh partitioning with 4 bins

These approaches to binning and smoothing allow the removal of noise (random error or variance in a measured variable) from the data. Binning methods help smooth a data value by associated it with its "neighborhood" or the values surrounding it. In equal-depth partitioning, smoothing by mean or by median indeed “smooth” the data, as 204 and 215 in the original data are two extreme values and dominate the original mean. In equal-width partitioning, either way of smoothing didn’t really remove the noise. If we use 3 bins instead of 4 bins, then we can have bigger width to have greater effect of smoothing.

Problem 3 (10 points):

a) (2 points) Figure 1 illustrates the plots for some data with respect to two variables: balance and employment status. If you have to select one of these two variables to classify the data into two classes (circle class and plus class), which one would you select? Is there any approach/criterion that you can use to support your selection? Explain your answer.

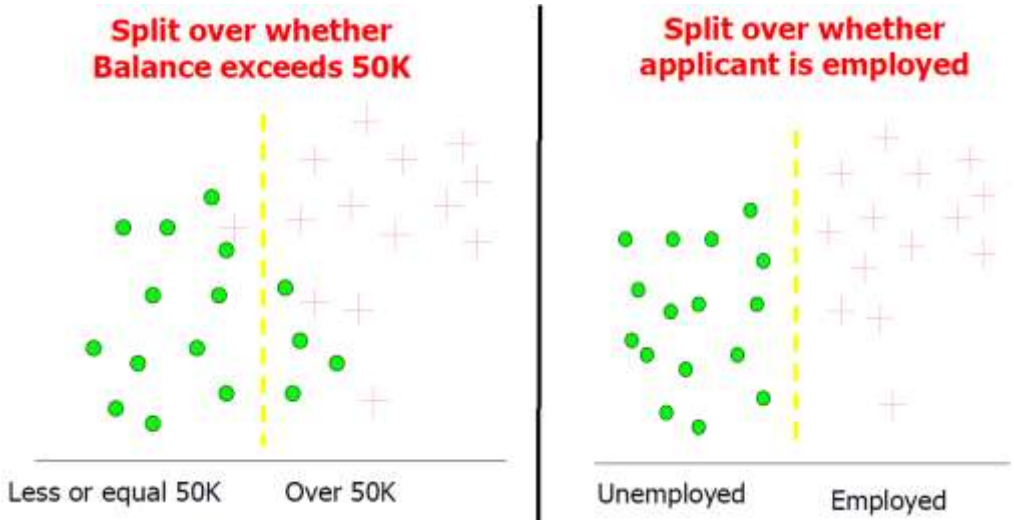

Figure 1: Data Plots for Problem 3.a.

I would use the unemployed / employed variable to classify the data because there seems to be a distinctive grouping/cluster. It's obvious that in balance some data can't be classified using just one criterion. I would consider using the clustering technique which partitions the objects into groups so that objects within a cluster are similar to one another.

b) (8 points) For the data in Figure 2 with three variables (X, Y, and Z) and two classes (I and II): which variable you would choose to classify the data? Show all the steps of your calculations and interpret your answer.

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Figure 2: Data for Problem 3.b

Calculate expected info required to classify an arbitrary tuple:

$I(S_1, S_2) = I(2, 2) = -2/4\log_2(2/4) - 2/4\log_2(2/4) = 1$

Calculate entropy of each attribute:

Attribute X:

For X = "1": $S_{11} = 2$, $S_{21} = 1$, $I(S_{11}, S_{21}) = -2/3\log_2(2/3) – 1/3\log_2(1/3) = 0.9183$

For X = "0": $S_{12} = 0$, $S_{22} = 1$, $I(S_{12}, S_{22}) = -0/1\log_2(0/1) – 1/1\log_2(1/1) = 0$

Attribute Y:

For Y = "1": $S_{11} = 2$, $S_{21} = 0$, $I(S_{11}, S_{21}) = -2/2\log_2(2/2) – 0/2\log_2(0/2) = 0$

For Y = "0": $S_{12} = 0$, $S_{22} = 2$, $I(S_{12}, S_{22}) = -0/2\log_2(0/2) – 2/2\log_2(2/2) = 0$

Attribute Z:

For Z = "1": $S_{11} = 1$, $S_{21} = 1$, $I(S_{11}, S_{21}) = -1/2\log_2(1/2) - 1/2\log_2(1/2) = 1$

For Z = "0": $S_{12} = 1$, $S_{22} = 1$, $I(S_{12}, S_{22}) = -1/2\log_2(1/2) - 1/2\log_2(1/2) = 1$

Calculate expected info required to classify a given sample if S is partitioned according to the attribute:

$E(X) = ¾\ I(S_{11}, S_{21}) + ¼\ I(S_{12}, S_{22}) = 0.688725$

$E(Y) = 2/4 I(S_{11}, S_{21}) + 2/4 I(S_{12}, S_{22}) = 0$

$E(Z) = 2/4 I(S_{11}, S_{21}) + 2/4 I(S_{12}, S_{22}) = 1$

Calculate information gain for each attribute:

$Gain(X) = I(S_1, S_2) - E(X) = 0.0817$

$Gain(Y) = I(S_1, S_2) - E(Y) = 1$

$Gain(Z) = I(S_1, S_2) - E(Z) = 0$

Attribute Y should be chosen to classify the data. Class I data both have attribute Y equal to "1" and class II data have attribute Y equal to "0". From information gain calculation, attribute Y has the highest value while the other two attributes are either close or equal to 0.

Problem 4 (10 points): Download the Wine Recognition Dataset from: https://archive.ics.uci.edu/ml/machine-learning-databases/wine/ The data are the results of chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. There are 178 instances in total with 59, 71, and 48 instances in class 1, class 2, and class 3, respectively.

NOTE: 1st attribute is class identifier (1-3)

a) (5 points) Are the variables (except the class variable) correlated? Explain your answer.

Yes. From table 5 pair-wised correlation matrix, we can find some weak or strong correlations between variables.

The most obvious and strongest positive correlation is between flavonoid and total phenols, with 0.864. Total phenols is also strongly positively correlated with proanthocyanins and OD280/OD315 of diluted wines. Flavonoid is also strongly positively correlated with proanthocyanins, hue and OD280/OD315 of diluted wines and negatively correlated with nonflavonoid phenols. Alcohol is positively correlated with color intensity and proline.

The strongest negative correlation is between malic acid and hue with -0.561, meaning the higher level of malic acid, the lighter the hue.

OD280/OD315 of diluted wines is the variable with most correlations with other variables, including total phenols, flavonoids, nonflavonoid phenols, proanthocyanins and hue.

Correlations

		Alcohol	Malic acid	Ash	Ash alcalinity	Magnesium	Total phenols	Flavonoids	Nonflavonoid phenols	Proanthocyanins	Color intensity	Hue	Diluted wines	Proline
Alcohol	Pearson Correlation	1												
Malic acid	Pearson Correlation	.100	1											
Ash	Pearson Correlation	.211	.165	1										
Ash alcalinity	Pearson Correlation	-.303	.286	.447	1									
Magnesium	Pearson Correlation	.259	-.049	.287	-.072	1								
Total phenols	Pearson Correlation	.285	-.334	.128	-.318	.208	1							
Flavonoids	Pearson Correlation	.230	-.409	.114	-.347	.187	**.864**	1						
Nonflavonoid phenols	Pearson Correlation	-.151	.292	.187	.359	-.252	-.448	**-.536**	1					
Proanthocyanins	Pearson Correlation	.128	-.218	.008	-.191	.227	**.611**	**.650**	-.363	1				
Color intensity	Pearson Correlation	**.548**	.250	.259	**.020**	.199	-.056	-.174	.140	**-.027**	1			
Hue	Pearson Correlation	-.075	**-.561**	**-.075**	-.273	**.052**	.433	**.543**	-.262	.294	**-.523**	1		
Diluted wines	Pearson Correlation	**.057**	-.367	**.002**	-.268	**.047**	**.700**	**.786**	**-.502**	**.513**	-.436	**.567**	1	
Proline	Pearson Correlation	**.641**	-.190	.223	-.437	.388	.496	.491	-.309	.326	.316	.235	.306	1

Table 5: Correlation matrix of 13 variables

Note:|Value| > 0.5 are bold and highlighted in red and the symmetric part are removed for clear visualization

b) (5points) Report the ranges for each of the variables. Would you recommend to normalize the data? If yes, which approach would you apply? Justify your answer.

Statistics

	Alcohol	Malic acid	Ash	Ash alcalinity	Magnesium	Total phenols	Flavonoids	Nonflavonoid phenols	Proanthocyanins	Color intensity	Hue	Diluted wines	Proline
Range	3.80	5.06	1.87	19.4	92	2.90	4.74	.53	3.17	11.7200	1.230	2.73	1402
Min	11.03	.74	1.36	10.6	70	.98	.34	.13	.41	1.2800	.480	1.27	278
Ma	14.83	5.80	3.23	30.0	162	3.88	5.08	.66	3.58	13.0000	1.710	4.00	1680

Table 6: Range of 13 variables

In this case, we are trying to solve a classification problem. As indicated in the last questions, we have several variables that are highly correlated to each other. So we want to apply feature extraction algorithm PCA (Principal Component Analysis) to achieve dimensionality reduction. The application of PCA requires data to be normalized. From table 6, we can see that variable Proline has its range/scale of [278, 1680] whereas Ash has its range/scale of only [1.36, 3.23]. Such huge difference indicates that we must normalize the data.

Z-score normalization is the most common way to apply with scale range [0, 1] and value range [-infinity, infinity]. This method preserve range (maximum and minimum) and introduce the dispersion of standard deviation / variance.
Min-Max normalization is also recommended and we need wo make sure the scale for all variables are the same as the new min-max range could be [0,1] or [-1, 1].