Assignment 2

Kubra Iqbal

Due Date: Saturday, *October 6th, 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.

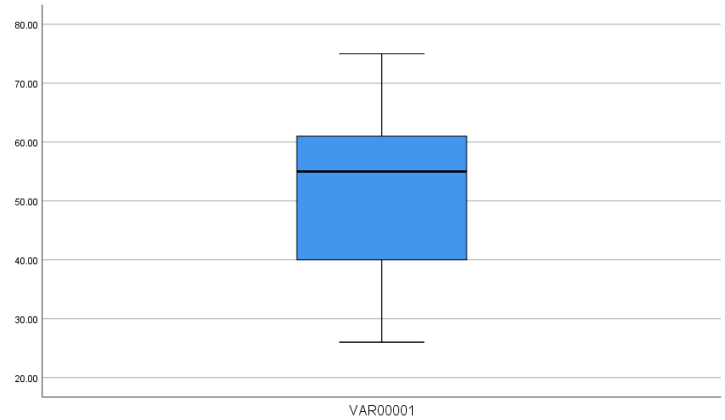
Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	26	26	29	29	40	45	50	55	60
%fat	10.5	30.5	8.8	20.8	32.4	26.9	30.4	30.2	33.2
Age	55	45	60	55	61	62	63	75	66
%fat	36.6	44.5	30.8	35.4	33.2	36.1	37.9	43.2	37.7

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

Case Processing Summary

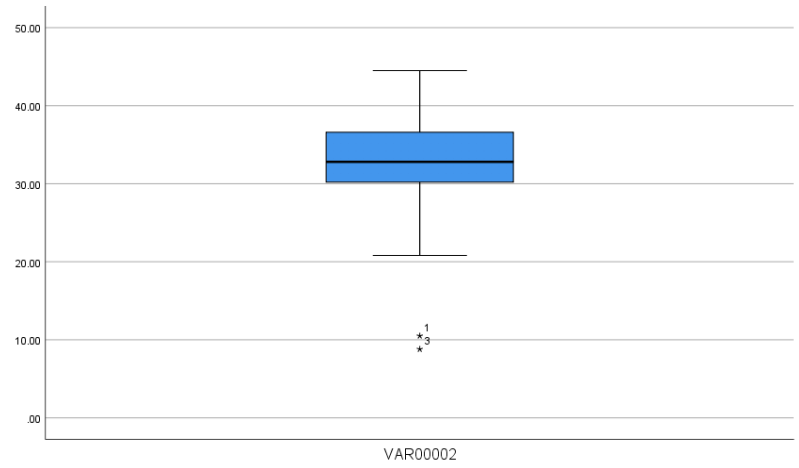
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
VAR00001	18	94.7%	1	5.3%	19	100.0%



The boxplot above is for age. No outliers for the Age Box Plot chart

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
VAR00002	18	100.0%	0	0.0%	18	100.0%



The boxplot above is for %age fat. This shows that it has two outliers : 10.5,8.8

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

	VAR00001	VAR00002	ZVAR00001	ZVAR00002	var
1	26.00	10.50	-1.61828	-2.15530	
2	26.00	30.50	-1.61828	-.05882	
3	29.00	8.80	-1.41693	-2.33351	
4	29.00	20.80	-1.41693	-1.07561	
5	40.00	32.40	-.67863	.14035	
6	45.00	26.90	-.34305	-.43619	
7	50.00	30.40	-.00746	-.06930	
8	55.00	30.20	.32813	-.09027	
9	60.00	33.20	.66372	.22421	
10	55.00	36.60	.32813	.58061	
11	45.00	44.50	-.34305	1.40872	
12	60.00	30.80	.66372	-.02737	
13	55.00	35.40	.32813039880148	.45482	
14	61.00	33.20	.73084	.22421	
15	62.00	36.10	.79795	.52820	
16	63.00	37.90	.86507	.71688	
17	75.00	43.20	1.67048	1.27245	
18	66.00	37.70	1.06642	.69592	
19					
20					

Variable 1 : Age

Variable 2 : %fat

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

Range : (new min(o), new max (1)

The Min – Max function is used to show a minimum value and maximum value from a specific database. The range is from Min to Max

ii. Z-score normalization

The normal range of all the data set z score normalization is (-∞,∞)

The range is (min-max)/standard devication, (max-mean)/standard devication. The data set range is from infinity to 1.

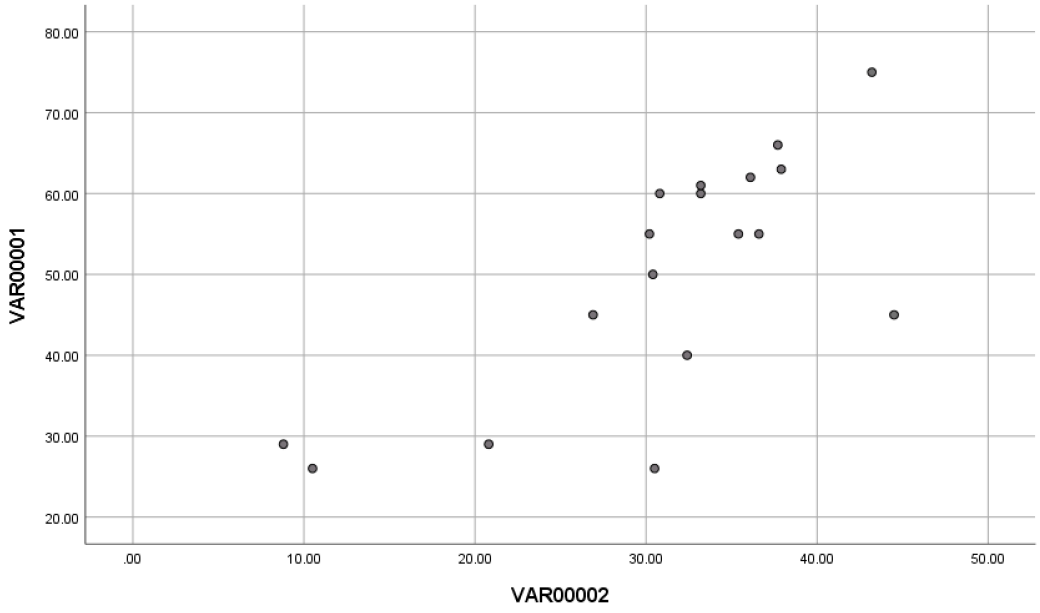
iii. Normalization by decimal scaling.

(-1,1)

The range for normalization by decimal scaling is in between (-1 to 1)

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

Graph



X axis : Body fat. Y Axis : Age

The above scatter plot shows a positive skewness. Body fat % is dependent on age. As the age keeps increasing, the body fat keeps increasing as well.

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?

Nonparametric Correlations

Correlations

			VAR00001	VAR00002
Kendall's tau_b	VAR00001	Correlation Coefficient	1.000	.571**
		Sig. (2-tailed)	.	.001
		N	18	18
	VAR00002	Correlation Coefficient	.571**	1.000
		Sig. (2-tailed)	.001	.
		N	18	18
Spearman's rho	VAR00001	Correlation Coefficient	1.000	.703**
		Sig. (2-tailed)	.	.001
		N	18	18
	VAR00002	Correlation Coefficient	.703**	1.000
		Sig. (2-tailed)	.001	.
		N	18	18

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

Correlation matrix shows a positive relationship between both the variables.

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:

8, 13, 14, 15, 17, 37, 55, 60, 77, 95, 208, 218

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

Bin 1 : 8, 13, 14, 15

Bin 2 : 17, 37, 55, 60

Bin 3 : 77, 95, 208, 218

Mean 12.5, 42.5, 149.5

Smoothing by bin :

Bin 1 : 13, 13, 13, 13

Bin 2 : 42, 42, 42 , 42

Bin 3 : 150, 150, 150, 150

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

Min = 8

Max = 218

N= 4

(218 – 8)/4 = 52.5

Bin 1	Range (8, 60.5)	8, 13, 14, 15, 17, 37, 55, 60
Bin 2	Range (60.5, 113)	77, 95
Bin 3	Range(113, 165.5)	
Bin 4	Range (165.5, 218)	208,218

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.

I would use the employed/unemployed variable to classify the data back into either plus class or circle class. This is because it shows that there is distinctive grouping/cluster. I would also utilize the clustering technique with partitioning the object into groups or clusters so that the given objects within a cluster are similar to each other and dissimilar to objects in other clusters.

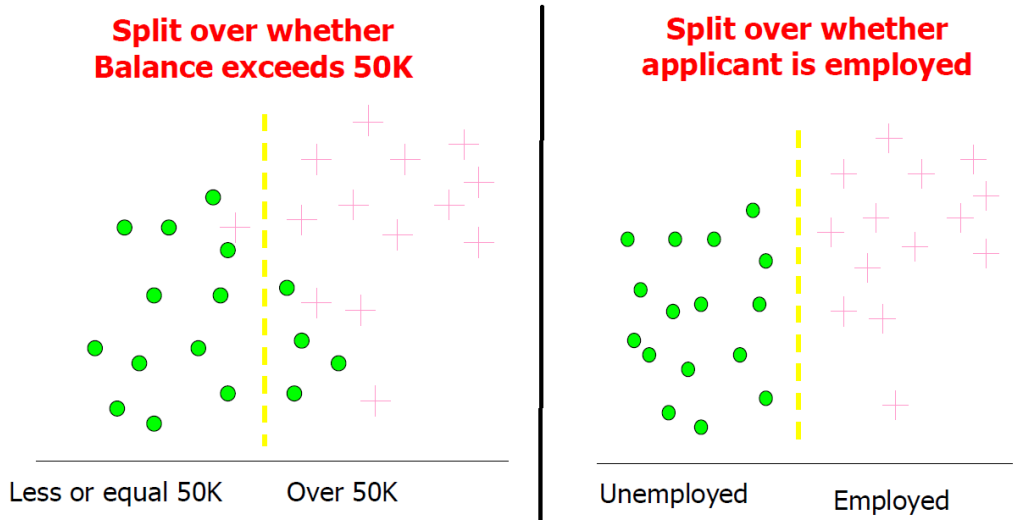


Figure 1: Data Plots for Problem 3.a.

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	1	I
0	0	1	II
1	0	0	II

Figure 2: Data for Problem 3.b

S1=2
S2=2
S= S1+S2
= 4

$I(S1,S2) = -(S1/S)\log_2(S1/S) - (S2/S)\log_2(S2/S)$
$= -(2/4)\log_2(2/4) - (2/4)\log_2(2/4)$
$= -0.5(-1) - 0.5(-1)$
$= 1$

For X;
S11= 2, S21= 1
$I(S11, S21) = -(2/3)\log_2(2/3) - (1/3)\log_2(1/3)$
$= 0.3899 + 0.528$
$= 0.9179$

S12= 0, S22= 1
$I(S12, S22) = -(0/1)\log_2(0/1) - (1/1)\log_2(1/1)$
$= 0 - 1(0)$
$= 0$

$E(X) = (3/4)(0.9179) + (1/4)(0)$
$= 0.6884$

$G(X) = I(S1,S2) - E(X)$
$= 1 - 0.6884$

= 0.3115……(a)

For Y:
S11= 2, S21= 0
$I(S11, S21) = - (2/2) \log_2 (2/2) - (0/2) \log_2 (0/2)$
$= 0$

S12= 0 S22= 2
$I(S12, S22) = - (0/2) \log_2 (0/2) - (2/2) \log_2 (2/2)$
$= 0$

$E(Y) = (2/4)(0) + (2/4)(0)$
$= 0$
$G(Y) = I(S1,S2) – E(Y)$
= 1 – 0 = 1……(b)

For Z;
$I(S12, S22) = - (0/1) \log_2 (0/1) - (1/1) \log_2 (1/1)$
$= 0 – 1(0)$
$= 0$
$E(Z) = (3/4)(0.9179) + (1/4)(0)$
$= 0.6884$

$G(Z) = I(S1,S2) – E(Z)$
$= 1 – 0.6884$
$= 0.3115$

As interpreted from the calculation G(Y) is the highest. Variable Y should be picked for classification
As seen from (1), (2), (3) ; G(Y) is the highest
Therefore, we would choose variable Y for classification.

Problem 4 (10 points): Download the Spotify Dataset along with the description from D2L.

a) (5 points) Describe the data in terms of number of attributes, number of cases, class distribution. Is there any correlation between features? Explain your answer.

The data is combined with the type of data that defines the music from Spotify. It consists of the name of the song, the ID of the song is unique for every song with a URI for each song. With different kinds, the music is divided into several attributes. The “mode” is one of the number of cases because it is binary : 0 and 1. Attributes include : Name, ID and Uri of the data. On the whole, the data has 17 attributes, 1421 cases and 4 classes.

==

b) (5points) Report the ranges for each numerical variable. Would you recommend normalizing the data? If yes, which approach would you apply? Justify your answer.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
0.838	1419	1.00	.00	1.00	.3993	.01012	.38135	.145
0.602	1419	.91	.06	.97	.5561	.00503	.18953	.036
475680	1419	4445704.00	54333.00	4500037.00	283275.9831	8953.35020	337269.2676	1.138E+11
0.302	1419	1.00	.00	1.00	.5549	.00781	.29408	.086
0.907	1419	1.00	.00	1.00	.2801	.01041	.39231	.154
8	1419	11.00	.00	11.00	5.1367	.09466	3.56597	12.716
0.113	1419	.96	.02	.98	.1916	.00440	.16579	.027
-11.627	1419	41.06	-41.81	-.75	-10.7379	.21209	7.98920	63.827
1	1419	1.00	.00	1.00	.5877	.01307	.49242	.242
0.0427	1419	.50	.02	.52	.0827	.00220	.08291	.007
119.758	1419	161.17	52.80	213.97	116.8557	.74221	27.95862	781.685
4	1419	4.00	1.00	5.00	3.8908	.01285	.48392	.234
0.3	1419	.97	.00	.97	.3983	.00686	.25834	.067
Valid N (listwise)	1419							

Based on the ranges for the attributes, normalizing the data would be the best idea. It would be best to normalize the data by decimal scale because the data has outliers which have a disadvantage with the max and min. The data is also not normally distributed which would cause the a disadvantage in using the Z-score method for this data.

Submission Instructions

1. Answer the problems and write your answers in a Word document.
2. Submit your file online at the website at **http://d2l.depaul.edu** and check your submission
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.
5. Late submissions are allowed with a 5%, 10%, and 15% penalty for a one day, two days, and three days, respectively.
6. No late work will be accepted after three days since the assignment was due.