**ASSIGNMENT# 2**

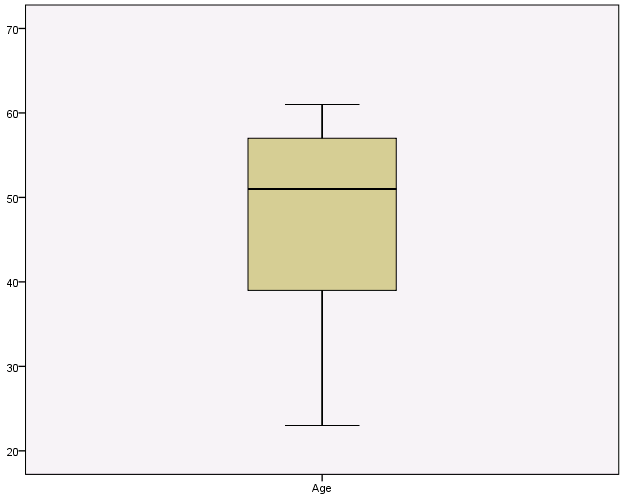
**Problem 1 (15 points):** This problem is an example of data preprocessing needed in a data mining process.

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

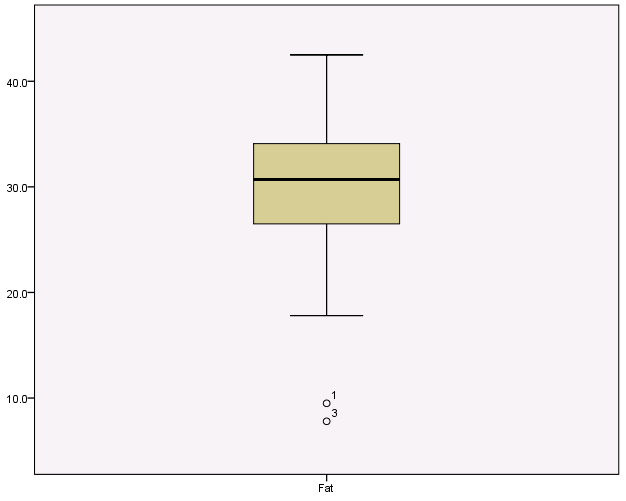
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 |
| %fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |
| Age | 52 | 54 | 54 | 56 | 57 | 58 | 58 | 60 | 61 |
| %fat | 34.6 | 42.5 | 28.8 | 33.4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

* 1. **(3 points) Draw the box-plots for age and %fat. Interpret the distribution of the data.**

The box plot for age, it is evident that the distribution is skewed to the left. The age samples ranges from 23 to 61, median is 51, Q1 is 39 and Q3 is 57. Also it can be noted from the graph, there are no outliers detected.

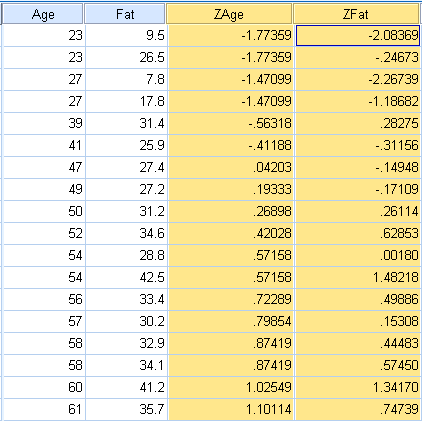


The box plot for %fat, it is evident that the distribution is not skewed. The %fat samples ranges from 7.8 to 42.5, median is 30.7, Q1 is 26.5 and Q3 is 34.1. Also it can be noted from the graph, there are two outliers detected at 7.8 and 9.5.



* 1. **(3 points) Normalize the two attributes based on z-score normalization.**

All the normalized values fall in the interval of [-3, 3], there are no outliers detected.



* 1. **(3 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.**

* + 1. **Min-max normalization**

The formula used for Min- Max normalization is



Thus, for each value v from the original interval [min, max] will be mapped with a new value v’ into a new interval [new\_min, new\_max].

* + 1. **Z-score normalization.**

This is used when data is normally distributed. The original values v are normalized into new values v’ using mean and standard deviation.

V’ =

About 99.7% of normalized values fall in the interval of [-3, 3].

* + 1. **Normalization by decimal scaling.**

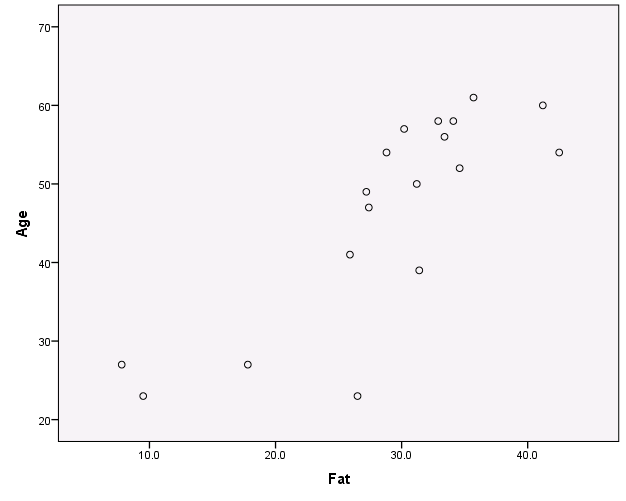
The formula used for decimal scaling normalization

V1 = where j is the smallest integer such that Max (|v’|) < 1.

It moves the decimal point of values so that the new values will be in the range [-1, 1].

* 1. **(3 points) Draw a scatter-plot based on the two variables and interpret the relationship between the two variables.**

The scatter plot for the Age and %fat are plotted in the graph. There is a positive relation between two variables. As the Age increases so does the %fat.



* 1. **(3 points) Calculate the correlation coefficient. Are these two attributes positively or negatively correlated? Compute the covariance matrix.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Correlations** | | | |
|  | | Age | Fat |
| Age | Pearson Correlation | 1 | .818\*\* |
| Sig. (2-tailed) |  | .000 |
| N | 18 | 18 |
| Fat | Pearson Correlation | .818\*\* | 1 |
| Sig. (2-tailed) | .000 |  |
| N | 18 | 18 |
|  | | | |
|  | | | |

The correlation coefficient of Age and %fat is 0.818. The two variables are positively correlated.

|  |  |  |  |
| --- | --- | --- | --- |
| **Correlations** | | | |
|  | | Age | Fat |
| 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 | 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 |
|  | | | |

**Problem 2 (10 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:

1. **(5 points) equal-depth partitioning with 3 values per bin**

In equal-depth partitioning, the range is divided into N intervals, where each bin contains approximately same number of samples. In this problem, a group of 12 sales prices records are divided into 4 bins, each containing 3 samples.

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as bin boundaries. Each bin value is then replaced by the closest boundary value. The larger the width, the greater the effect of the smoothing.

|  |  |  |
| --- | --- | --- |
|  | N Intervals | Smooth by boundaries |
| Bin 1 | 5, 10, 11 | 5, 11, 11 |
| Bin 2 | 13, 15, 35 | 13, 13, 35 |
| Bin 3 | 50, 55, 72 | 50, 50, 72 |
| Bin 4 | 92, 204, 215 | 92, 215, 215 |

1. **(5 points) equal-width partitioning with 3 bins**

In equal-width partitioning, the samples is divided into range of N intervals of equal size.

Width = (max-min)/N 🡺 (215-5)/3 = 70.

Thus in the above problem, the size of the each bin is 70.

Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72

Bin 2: 92

Bin 3: 204, 215

**Problem 3 (10 points):**

1. (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.

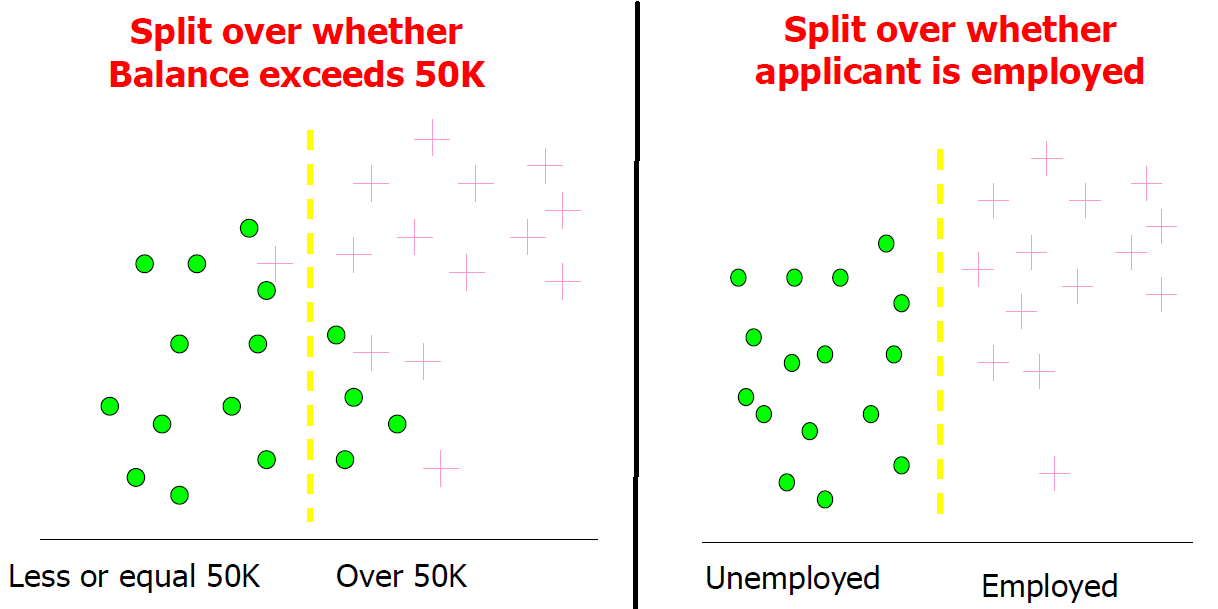
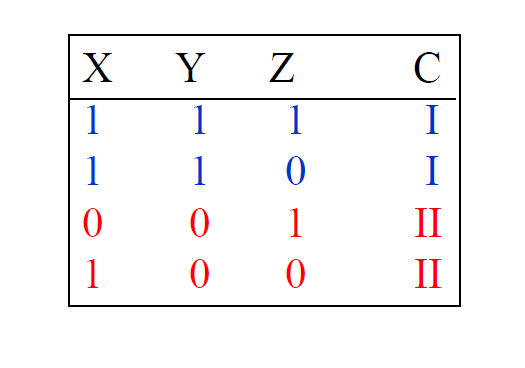


Figure 1: Data Plots for Problem 3.a.

The variable employment status is selected to classify the data into two classes of class circle and class plus. The Employed variable with Unemployed and Employed values, the boundary clearly distinguishes both classes 100 % accurately with no impurities. Whereas in case of balance variable the values are not distributed evenly resulting in outliers/ impurities. Hence the variable Employed is best suited for classifying the data.

1. (8 points) For the data in Figure 2 with three variables and two classes: which variable you would choose to classify the data? Show all the steps of your calculations and interpret your answer.



**Figure 2: Data for Problem 3.b**

**When we split on X variable**

Information Gain = 1

E(X)= - (1/3) log2 (1/3) – (2/3) log2 (2/3) -0

= 0.5284 + 0.39

= 0.9184

Gain = 1 – (3/4) 0.9184 – (1/4) 0

= 0.3112

**Split on Y variable**

E(Y) = 0+0

Gain = 1- (1/2) 0 – (1/2) 0 = 1 🡺 Best case

**Split on Z variable**

E (Z) = 1+1

Gain = 1- (1/2) 1 – (1/2) 1 = 0 🡺 Worst case

Hence variable Y is chosen to be the best to classify the data.