**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 |

1. (2 points) Draw the box-plots for age and %fat. Interpret the distribution of the data.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

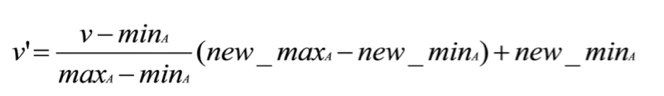
* Age:
  1. Data distribution is skewed to left
  2. The values range is between 26 and 75.
  3. The first quartile = 37.25, median = 55, third quartile = 61.25.
  4. There is no outlier
* %FAT :
  1. Data distribution is normal.
  2. The values range between 8.8 and 44.5.
  3. The first quartile = 29.37, median = 32.8 and third quartiles = 36.87.
  4. There are two denoted outliers denoted at 8.8 and 10.5.

1. (2 points) Normalize the two attributes based on z-score normalization.

A screenshot of a computer

Description automatically generated

1. (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.
   * 1. Min-max normalization

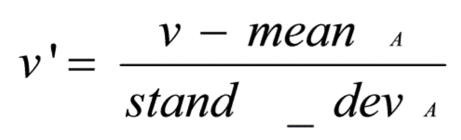


Range [min, max] = [0, 1].

When using the min-max normalization, the values are forced into a specific range.

The advantage of this method is when outliers are present in the data.

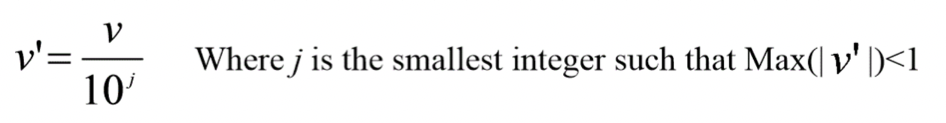
* + 1. Z-score normalization



The range for Z-score normalization is [-∞, ∞].

Z-score normalization is useful when the data is normally distributed.

* + 1. Normalization by decimal scaling.



The range for normalization by decimal scaling is between -1 and 1.

1. (2 points) Draw a scatter-plot based on the two variables and interpret the relationship between the two variables.

A picture containing white, light, large, water

Description automatically generatedBased on the scatterplot we can see a positive correlation between Age and %Fat.

1. (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?

A screenshot of a cell phone

Description automatically generated

Correlation Matrix:

The two values are strongly correlated with Pearson correlation coefficient of 0.735

Covariance:

The covariance tells us how much the two variables vary together. This also shows a strong relationship between Age and %Fat.

**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:

1. (2.5 points) equal-depth partitioning with 4 values per bin

N=12 12/4 = 3

1. Bin 1: 8, 13, 14, 15
2. Bin 2: 17, 37, 55, 60
3. Bin 3: 77, 95, 208, 218

Smoothing by Bin means:

1. Bin 1: 12.5, 12.5, 12.5, 12.5
2. Bin 2: 42.25, 42.25, 42.25, 42.25
3. Bin 3: 149.5, 149.5, 149.5, 149.5
4. (2.5 points) equal-width partitioning with 4 bins

Min=8; Max=218; N=4

W=(218-8)/4 = 52.5

* 1. Bin 1: 8, 13, 14, 15, 17, 37, 55, 60 Range [8, 60.5]
  2. Bin 2: 77, 95 Range [ 60.5, 113]
  3. Bin 3: Range [113, 165.5]
  4. Bin 4 : 208, 2018 Range [165.5, 218]

Smoothing by Bin means:

1. Bin 1: 27.375, 27.375, 27.375, 27.375, 27.375, 27.375, 27.375, 27.375
2. Bin 2: 86, 86
3. Bin 3:
4. Bin 4: 213, 213

When using the equal-width method, each bin will have the same size range of values. In this case, the equal-width binning method does not work well because the majority of the values fall into the first bin and there is an empty bin present. This indicated that the data is not normally distributed.

**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.

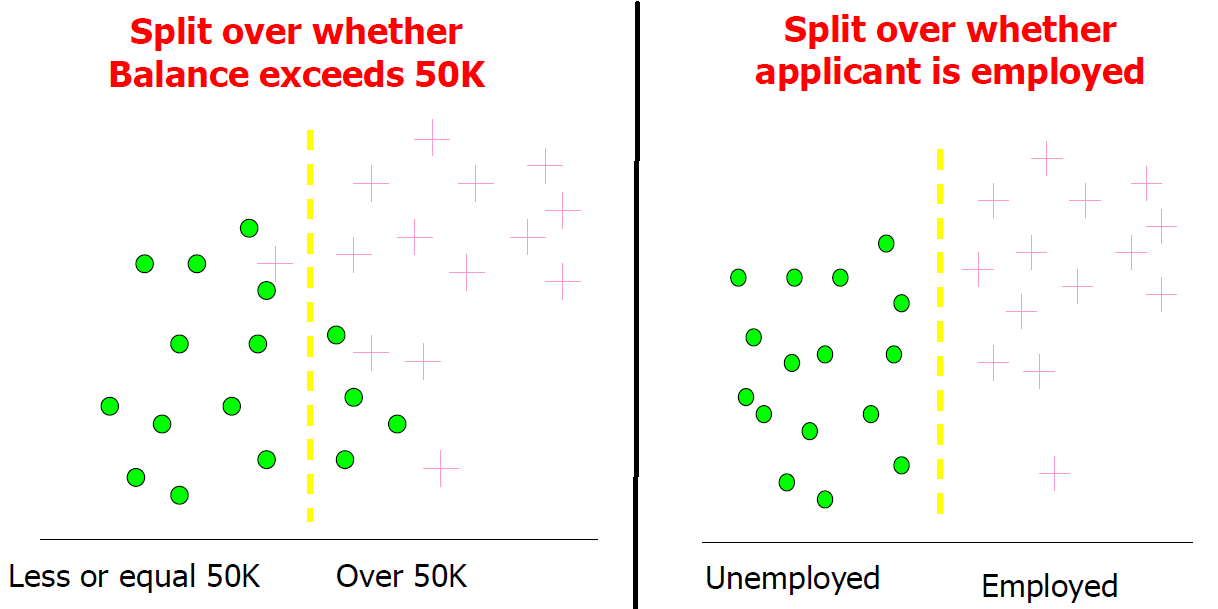


Figure 1: Data Plots for Problem 3.a.

I would select the employment status variable which would be easier to classify into 2 classes because the two classes are perfectly clustered. As there are 2 classes we can go with binary classification.

1. (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 | Ĉ |
| 1 | 1 | 1 | I |
| 1 | 1 | 1 | I |
| 0 | 0 | 1 | II |
| 1 | 0 | 0 | II |

Figure 2: Data for Problem 3.b

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

variable Y has an accuracy rate of 100%. So Variable Y would be the best variable to classify the data because as it has the highest accuracy rate and perfect classification. So, it is the best choice to classify the data.

Variables X and Z both have a 75% accuracy rate.

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

1. (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.

number of attributes: 13

number of cases: 1420

class distribution:

|  |  |
| --- | --- |
| Mood | Count |
| dinner | 467 |
| dinner, party | 3 |
| dinner, workout | 1 |
| party | 225 |
| party, workout | 52 |
| sleep | 362 |
| workout | 310 |

A close up of a piece of paper

Description automatically generatedA screenshot of a cell phone

Description automatically generated

|  |  |
| --- | --- |
| positive correlation: | negative correlations: |
| Energy and loudness  Instrumentalness and acousticness  Danceability and loudness  Danceability and valence | Instrumentalness and danceability  Instrumentalness and energy  Instrumentalness and loudness  Instrumentalness and valence  Acousticness and danceability  Acousticness and energy  Acousticness and loudness |

1. (5points) Report the ranges for each numerical variable. Would you recommend normalizing the data? If yes, which approach would you apply? Justify your answer.

A screenshot of a cell phone

Description automatically generated

Yes, I recommend normalizing the data.

As there are outliers Min- max approach fails and the data is not distributed normally Z-Score don’t work

So, we can select with decimal scale normalization