**CSC 4740/6740 Data Mining**

**Assignment 1**

**Due Date: 11:59 pm, Thursday, September 22, 2022**

**Note**: Even though these statistics are simple to compute, which can be done manually, I suggest that you calculate them using programs, either Python or Matlab are recommended. You can call the API functions from any libraries. This will be better for you to get familiar with these API functions. Real-life dataset will be large and computer programs are needed.

1. (10 points) Suppose we have the BestBuy customer data in the following table.

|  |  |
| --- | --- |
| **Customer** | **Age** |
| David | 46 |
| Lisa | 25 |
| Michael | 27 |
| Susan | 27 |
| William | 28 |
| Mat | 36 |
| James | 53 |
| Kevin | 27 |
| Paul | 18 |
| Anthony | 25 |

* 1. Please calculate the mean, median, and mode.

mean = 31.2

median = 27

mode = 27

1. (25 points) Suppose we have the climate data for Atlanta in the following table. Climate data for Atlanta

|  |  |
| --- | --- |
| **Month** | **Temperature (**℉**)** |
| Jan | 52.3 |
| Feb | 56.6 |
| Mar | 64.6 |
| Apr | 72.5 |
| May | 79.9 |
| Jun | 86.4 |
| Jul | 89.1 |
| Aug | 88.1 |
| Sep | 82.2 |
| Oct | 72.7 |
| Nov | 63.6 |
| Dec |  |

2.1) Please compute the five-number summary of this dataset.

max = 89.1, min = 52.3, q1 = 61.85, q2 = 72.6, q3 = 83.25

* 1. Will there be outliers if we use boxplot to visualize the five-number summary? If yes, please indicate which data objects are outliers. Please briefly explain your answers.

No, there will not be any outliers, as the maximum values and the minimum values are below and above their respective thresholds. The threshold for an outlier is if it is less than the difference between first quartile and 1.5\*inter quartile range: (q1 – 1.5 \* IQR) or if it is greater than the sum of third quartile and 1.5\*inter quartile range: (q3 + 1.5\*IQR). The minimum value “52.3” is greater than 29.48, as calculated by (q1 – 1.5 \* IQR) and the maximum value “89.1” is less than 115.35, as calculated by (q3 + 1.5 \* IQR). There are no outliers in this dataset, when visualized by boxplot.

2.3) Please visualize the data by using plot function in Matlab or some similar functions in other software. You can use any software. Based on the plotted curve, please also briefly describe the visualization result.

Chart, box and whisker chart

Description automatically generated

Here, we see the minimum and maximum values represented at the 2 ends of the plot. The line of the plot between 60 and 65 is q1, the middle orange line is q2 (median), and the line between 80 and 85 is q3.

Chart, line chart

Description automatically generated

This plotting of the data looks like a bell-curve, but is not a bell curve. The rate of change from February to June is constant, just like the rate of change during September to December. Although we can judge the data based on its skewed nature, we will need to normalize the dataset first. In all, this look like a bell curve, but is not one. One critical value is when month = July, that is where we have a global maximum temperature. There is a local minimum at January, but it can also be the global minimum because we know that the months are confined.

1. (15 points) Suppose we have the customers’ information in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer** | **David** | **Susan** | **Lisa** |
| **Profession** | Manager | Manager | Programmer |
| **Education** | B.Sc. | B.Sc. | M.Sc. |
| **Hobbies** | Golf | Swimming | Swimming |

* 1. Which types of attributes are there in the table?

Customer is a nominal attribute, profession is a discrete attribute, education is a binary attribute (if the possible degrees are B.Sc. and M.Sc.); hobbies are nominal attributes.

* 1. Please compute the similarity values between “David” and “Susan”.

Matches = profession, education

Similarity = matches/total = 2/3

* 1. Please compute the similarity values between “Susan” and “Lisa”.

Matches = Hobbies

Similarity = matches/total = 1/3

1. (15 points) Suppose we have the patients’ information in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Patient** | **Tom** | **Mat** | **Lucy** |
| **Fever** | Yes | No | Yes |
| **Cough** | No | Yes | Yes |
| **Sleepy** | Yes | No | No |
| **Headache** | Yes | Yes | No |
| **Running nose** | Yes | Yes | No |
| **Fatigue** | Yes | Yes | Yes |
| **Sweaty** | Yes | No | Yes |
| **Dizziness** | Yes | Yes | Yes |

* 1. Which types of attributes are there in the table?

They are all binary attributes, as they only have 2 possible values: “yes” or “no.”

* 1. Compute the similarity values between “Tom” and “Mat”

Matches/Total = 4/8 = 1/2

* 1. Compute the similarity values between “Mat” and “Lucy”.

Matches/total = 4/8 = 1/2

1. (15 points) Suppose we have the Fisher’s iris data in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Flower** | **A** | **B** | **C** |
| **Sepal Length** | 5.1 | 7.0 | 4.8 |
| **Sepal Width** | 3.5 | 3.2 | 3.4 |
| **Petal Length** | 1.4 | 4.7 | 1.9 |
| **Petal Width** | 0.2 | 1.4 | 0.2 |

Please choose one similarity measure and solve the following problems.

* 1. Which types of attributes are there in the table?

They are all numerical attributes, as they have numbers.

* 1. Which type of similarity measure do you choose?

I can use the cosine similarity to solve this, as it is useful to find similarities even if the two points are far away with respect to Euclidean distances (although this is not a problem here, as we do not have a very large dataset). It also captures the orientation of the subjects and not their magnitudes. The similarity is calculated based on the angles between data objects.

* 1. Compute the similarity values between “A” and “B”;

similarity = [5.1\*7 + 3.5\*3.2 + 1.4\*4.7 + 0.2\*1.4]/() = 0.9283 0.93 (high similarity!)

* 1. Compute the similarity values between “B” and “C”.

similarity = [4.8\*7 + 3.4\*3.2 + 1.9\*4.7 + 0.2\*1.4]/() = 0.9512 0.95 (high similarity!)

1. (15 points) Suppose we have the customer information in the loan company in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer** | **Kevin** | **John** | **Daniel** |
| **Credit Score Range** | Excellent | Very good | Good |
| **Salary Range** | High | Very High | Medium |
| **Age** | Senior | Middle Age | Young |

The ranking options within each attribute are provided in the following tables.

**Credit Score Range**

**Salary Range**

**Age**

Excellent

Very High

Senior

Very good

High

Middle Age

Good

Medium

Young

Fair

Low

Poor

* 1. Which types of attributes are there in the table?

They are all ordinal type of attributes, as they are all words, but have a relative ranking with respect to one another.

Rankings are as follows:



* 1. Compute the similarity values between “Kevin” and “John”.

Kevin = [1, 0.67, 1]

John = [0.75, 1, 0.5]

Similarity = 0.9113 0.91 (high similarity)

* 1. Compute the similarity values between “John” and “Daniel”.

John = [0.75, 1, 0.5]

Daniel = [0.5, 0.33, 0]

Similarity = 0.8741 0.87 (decent similarity)

1. (5 points) Please normalize the following dataset by using the min-max normalization method. The new range should be [0, 1].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient** | **Tom** | **Mat** | **Lucy** | **Brian** |
| **Height (feet)** | 5.7 | 6.2 | 5.1 | 6.4 |

First, I have decided to sort the given data, thereby converting it to:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient** | **Lucy** | **Tom** | **Mat** | **Brian** |
| **Height (feet)** | 5.1 | 5.7 | 6.2 | 6.4 |

Now, by employing the simple formula: (value – min)/(max-min) for every value in our dataset, I have produced:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient** | **Lucy** | **Tom** | **Mat** | **Brian** |
| **Height (feet)** | 0 | 0.4615384615384617 | 0.8461538461538461 | 1 |