**ISCS 539 Data Analytics**

Assignment 1

1. What are the most appropriate scales for the following examples?

1. university students' exam marks

Relative scale is good for the student’s exam marks like 90, 80 or 70.

1. level of urgency in the emergency room of a hospital

Ordinal scale is a good scale to categorize the level of urgency in a hospital emergency room since it can be mild to severe.

1. classification of the animals in a zoo

Nominal scale is the most appropriate for this classification because an animal can either be a zebra or not.

1. carbon dioxide levels in the atmosphere

Absolute scale is appropriate in categorizing the amount of carbon dioxide levels in the atmosphere.

2. The binary representation of the object given in rows 3 and 4 of Table 1 with the three attributes, Refund, Marital Status and Cheat, using One-hot encoding?

|  |  |  |  |
| --- | --- | --- | --- |
| Tid | Refund | Marital Status | Cheat |
| 3 | No | Single | No |
| 4 | Yes | Married | No |

Using one hot encoding we convert the three attributes which are in nominal scale to relative scale. The process uses 1-of-n known as one-attribute-per-value conversion, which transforms n values of a nominal attribute into n binary attributes.

**Refund** attribute have two values either Yes or No, therefore there are two binary representations.

|  |  |  |
| --- | --- | --- |
| Tid | Refund | One hot encoding |
| 3 | No | 01 |
| 4 | Yes | 10 |

**Married status** has three values single, married or divorced, there are three binary representations.

|  |  |
| --- | --- |
| Single | 001 |
| Married | 010 |
| Divorced | 100 |

|  |  |  |
| --- | --- | --- |
| Tid | Married Status | One hot encoding |
| 3 | Single | 001 |
| 4 | Married | 010 |

**Cheat** has two vales and therefore two binary representations.

|  |  |
| --- | --- |
| Yes | 01 |
| No | 10 |

|  |  |  |
| --- | --- | --- |
| Tid | Cheat | One hot encoding |
| 3 | No | 10 |
| 4 | No | 10 |

3. The values of an attribute from the interval [−100, 100] can be mapped to the interval [0, 1] by normalization or standardization. Min-max rescaling can be done by subtracting the minimum value from every value and dividing the result by the amplitude: the difference between the maximum and minimum of the new values. Therefore, to convert a set of values in interval [-100,100] to values in the interval [0, 1], we simply subtract the smallest value ‘-100’ from all the values and divide by the amplitude ‘100-100’ which is zero.

4. For the following vectors, x and y, calculate covariance, Pearson’s correlation and Spearman’s rank correlation. Justify your answers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | x2 | Y | y2 | xy |
| 2 | 4 | -1 | 1 | -2 |
| -1 | 1 | 1 | 1 | -1 |
| 0 | 0 | -2 | 4 | 0 |
| 1 | 1 | 0 | 0 | 0 |
| -2 | 4 | 1 | 1 | -2 |
| -3 | 9 | 2 | 4 | -6 |
|  |  |  |  |  |

Mean of x (x ̅) = (2+(-1)+0+1+(-2)+(-3))/6 = -0.5

Mean of y (y ̅) = (-1+1+(-2)+0+1+2)/6 = 0.17

The covariance of x and y is -0.76 which is close to -1. This means the two attributes are negatively correlated therefore as x increases y decreases.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | x - | (x - x ̅)2 | y | y - | (y -y ̅)2 | (x - x ̅) (y -y ̅) |
| 2 | 2.5 | 6.25 | -1 | -1.17 | 1.37 | -2.925 |
| -1 | -0.5 | 0.25 | 1 | 0.83 | 0.69 | -0.415 |
| 0 | 0.5 | 0.25 | -2 | -2.17 | 4.71 | -1.085 |
| 1 | 1.5 | 2.25 | 0 | -0.17 | 0.03 | -0.255 |
| -2 | -1.5 | 2.25 | 1 | 0.83 | 0.69 | -1.245 |
| -3 | -2.5 | 6.25 | 2 | 1.83 | 3.35 | -4.575 |
|  |  |  |  |  |  |  |

The pearson correlation coefficient is -0.76 which is close to -1 this indicates the points form a decreasing line. This result is in agreement with the covariance coefficient of x and y.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | x | xr | y | yr | d = xr – yr | d2 |
| 1 | 2 | 6.0 | -1 | 2.0 | 4.0 | 16.0 |
| 2 | -1 | 3.0 | 1 | 4.5 | -1.5 | 2.25 |
| 3 | 0 | 4.0 | -2 | 1.0 | 3.0 | 9.0 |
| 4 | 1 | 5.0 | 0 | 3.0 | 2.0 | 4.0 |
| 5 | -2 | 2.0 | 1 | 4.5 | -2.5 | 6.25 |
| 6 | -3 | 1.0 | 2 | 6.0 | 5.0 | 25.0 |
|  |  |  |  |  |  |  |

The spearman correlation coefficient is -0.79 this corresponds to a decreasing monotonic trend between x and y.

The values of the covariance coefficient, pearson correlation coefficient and spearman correlation coefficient are -0.76,-0.76 and -0.79 respectively. This implies a negative correlation and as the value of x increases, the value of y decreases.