Homework Assignment 2

Name: **Asrar Syed**

Due by: **September 13, 2024**

Course Section: **CSC 4740-002**

Q1: Explain the following normalization methods by using formula, the value ranges of output etc. (10 points)

1. **Min-Max Normalization:**
   1. Description of Method: Min-max normalization performs a linear transformation on the original data to scale it within a specific range, typically [0,1] or a custom range []. Given that and ​ are the minimum and maximum values of an attribute , this technique maps a value ​ from to a new normalized value ′​ within the specified range.
   2. **Formula:**

**Where:**

* + 1. is the original value.
    2. is the normalized value.
    3. & ​ are the minimum and maximum values of the attribute .
    4. & are the new desired minimum and maximum for normalization.
  1. **Explanation:** Min-max normalization keeps the relationships between the original data values. However, if a new data point falls outside the original range [], it could result in values outside the expected normalized range, which may cause an “out-of-bounds” error.

1. **Z-Score Normalization:**
   1. **Description of Method:** Z-Score Normalization is the process where the features are rescaled so that they have the property of a standard normal distribution with mean(μ) as 0 and standard deviation(σ) as 1.
   2. **Formula:**

**Where:**

* + 1. is the original value.
    2. is the mean of the data.
    3. is the standard deviation of the data.
    4. is the normalized value (z-score)
  1. **Explanation:** Z-score normalization standardizes data by converting each value into its z-score, which measures how far the value is from the mean in terms of standard deviations.

1. **Z-Score Normalization using Mean Absolute Deviation (MAD):**
   1. **Description of Method:** This is a variation of z-score normalization, but instead of using the standard deviation to measure spread, it uses the mean absolute deviation (MAD), which is less affected by outliers.
   2. **Formula:**

**Where:**

* + 1. is the original value.
    2. is the mean of the data.
    3. is the normalized value (z-score)
    4. (Mean Absolute Deviation) the average absolute difference between each data point and the mean, calculated as:
  1. **Explanation:** Z-score normalization using MAD scales the data based on how far each value is from the mean. This makes it more robust to outliers and similar to the regular z-score normalization, it centers the data around 0.

1. **Normalization by Decimal Scaling:**
   1. **Description of Method:** Decimal scaling shifts the decimal point of the values to scale them into a smaller range. The scaling factor is based on the largest value in the dataset.
   2. **Formula:**

**Where:**

* + 1. is the original value.
    2. is the smallest integer such that max
  1. **Explanation:** This method divides each value by a power of 10 so that the largest value in the dataset becomes less than 1. The number of decimal places is chosen based on the largest absolute value. This brings all values into a smaller range, typically between −1and 1, making it easier for some algorithms to process.

Q2: Use the methods below to normalize the following group of data: (10 points) 200, 300, 400, 600, 1000

1. **min-max normalization by setting min = 0 and max = 1**

|  |  |  |
| --- | --- | --- |
| **Data** | **Min-Max Normalization** | **Normalized Data** |
| 200 |  | 0 |
| 300 |  | 0.125 |
| 400 |  | 0.25 |
| 600 |  | 0.5 |
| 1000 |  | 1 |

1. **z-score normalization**

|  |  |  |
| --- | --- | --- |
| **Data** | **Z-Score Normalization** | **Normalized Data** |
| 200 |  | -1.06 |
| 300 |  | -0.707 |
| 400 |  | -0.354 |
| 600 |  | 0.354 |
| 1000 |  | 1.77 |

1. **z-score normalization using the mean absolute deviation instead of standard deviation**

|  |  |  |
| --- | --- | --- |
| **Data** | **Z-Score Normalization using MAD** | **Normalized Data** |
| 200 |  | -1.25 |
| 300 |  | -0.833 |
| 400 |  | -0.417 |
| 600 |  | 0.4117 |
| 1000 |  | 2.08 |

1. **normalization by decimal scaling**

|  |  |  |
| --- | --- | --- |
| **Data** | **Normalization using Decimal Scaling** | **Normalized Data** |
| 200 |  | 0.2 |
| 300 |  | 0.3 |
| 400 |  | 0.4 |
| 600 |  | 0.6 |
| 1000 |  | 1 |

Q3: Use a flowchart to explain the following procedures for attribute subset selection: (10 points)

(a) stepwise forward selection

A diagram of a process

Description automatically generated

Stepwise forward selection is a way to reduce the dimensionality of data by gradually selecting the most important attributes for a model. It starts with no attributes and adds one at a time, each chosen based on how much it improves the model’s performance. This process continues until adding more attributes doesn’t significantly help. The goal is to remove irrelevant or redundant attributes, making the model simpler and more efficient. Since testing all possible combinations of attributes would be too complex, stepwise selection uses a heuristic approach to find the best subset without having to try every combination.

Q4: The following table contains the attributes *name, trait-1, trait-2, trait-3, and trait-4*, etc. where *name* is an object identifier, attributes are describing personal traits of individuals who desire a penpal. Suppose that a service exists that attempts to find pairs of compatible penpals by computing the similarity of each pair.

**Compute the similarity of all pairs and which pair is the most compatible pair under**

**1) symmetric attributes**

The highest similarity score here is **0.571**, and two pairs have this score:

(Kevin, Eric): 0.571

(Caroline, Eric): 0.571

**2) asymmetric attributes assumption? (10 points)**

The highest similarity score here is **0.4**, for the pair:

(Kevin, Eric): 0.4

**Work for (self-reference in future)...**

**Note: Simple Matching (SMC) and Jaccard Coefficients (J)**

*I need to use* ***SMC*** *= number of matches / number of attributes for similarity in symmetric attributes*

*I need to* ***J*** *= number of 11 matches / number of non-zero attributes for similarity in asymmetric attributes.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Trait 1 | Trait 2 | Trait 3 | Trait 4 | Trait 5 | Trait 6 | Trait 7 |
| Kevin | P | P | P | N | N | N | P |
| Eric | P | N | P | P | N | N | N |
| Caroline | N | P | N | P | N | N | N |

**Note: We can let the value of P be set to 1 and the value N be set to 0.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kevin | Eric | | | |
|  | 1 | 0 | Sum |
| 1 | 2 | 2 | 4 |
| 0 | 1 | 2 | 3 |
| Sum | 3 | 4 | 7 |

|  |  |
| --- | --- |
| **Simple Matching (SMC) - Symmetric** | **Jaccard Coefficients (J) - Asymmetric** |
|  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Eric | Caroline | | | |
|  | 1 | 0 | Sum |
| 1 | 1 | 2 | 3 |
| 0 | 1 | 3 | 4 |
| Sum | 2 | 5 | 7 |

|  |  |
| --- | --- |
| **Simple Matching (SMC) - Symmetric** | **Jaccard Coefficients (J) - Asymmetric** |
|  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Caroline | Kevin | | | |
|  | 1 | 0 | Sum |
| 1 | 1 | 1 | 2 |
| 0 | 3 | 2 | 5 |
| Sum | 4 | 3 | 7 |

|  |  |
| --- | --- |
| **Simple Matching (SMC) - Symmetric** | **Jaccard Coefficients (J) - Asymmetric** |
|  |  |

Q5: The following is data collected from 30 participants from 10 attributes, (55 points)

Step 1: Please check data quality, do necessary cleaning and/or transformation steps, and explain what you did, and why (10 points) Hint: consider missing, outliers, scale/normalization

Missing Values: I removed the 5th row because it had too many missing values across multiple attributes, making it unreliable for analysis. Imputation was not feasible due to the extent of missing data, so removing the row ensures cleaner results.

Outliers: I detected mild outliers using the IQR method but chose not to adjust them as they didn’t significantly distort the data. If there was extreme outliers, I would have replaced the outlier values outside the bounds with the upper or lower limits of the IQR.

Scaling: I used z-score normalization for attributes with outliers to reduce their impact on the overall distribution. For attributes without outliers, I applied min-max scaling to maintain their original distribution while bringing them into a [0,1] range.

Step 2: Note: Python : be aware that **Each row of data array represents a variable** (5 points).

1. Compute covariance matrix of data among all variables after step1. (this should be a 10-by-10 matrix)
2. compute the total variance of data = sum of diagonal elements of covariance matrix (5 points).
3. compute correlation (Pearson’s correlation) between variable 1 and variable 2 (5 points).

Step 3: perform **Principal component analysis (PCA)** to generate a number of Principal Components (PCs) capturing >85% of total data variance.

1. Plot percentage of variances of each Principal Components(PC) in a decreasing order (5 points)
2. How many components do you need to capture > 85% total data variance? (5 points)
3. Plot the PC (or projection direction) of N components you selected (5 points, N lines in one plot or N separate plots)
4. Plot the generated (**NEW**) top P PC variables you selected (5 points, N lines in one plot or N separate plots)
5. Compute the covarion matrix of the NEW P PC variables (this should be P-by-P matrix), and compute the total variance of PCs ( sum of diagonal elements of covariance matrix ), compare this value with the total variance of data in Step2\_b, what % variance kept in PCs (5 points).
6. compute correlation (Pearson’s correlation) between new variable PC1 and new variable PC2 (5 points).

Q6: Graduate students Please apply UMAP or tSNE for visualization of clean data in 2D. Since we only have 30 data points, please select parameters with smaller neighbors. (5 points)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Fluid IQ** | **Crystallized IQ** | **Vocabulary** | **Inhibitory Control** | **Memory** | **Mental Flexibility** | **Processing Speed** | **Attention Problem** | **Anxiety Problem** | **Social Problems** |
| 1 | 122 | 77 | 131 | 81 | 86 | 86 | 80 | 8.6 | 7 | 8 |
| 2 | 103 | 77 | 98 | 69 | 97 | 84 | 57 | 10 | 8.5 | 9 |
| 3 | 148 | 91 | 153 | 89 | 109 | 87 | 67 | 7.8 | 8 | 7.2 |
| 4 | 137 | 107 | 142 | 106 | 105 | 102 | 94 | 7.6 | 6.6 | 5.6 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7.8 | 6.8 | 7.5 |
| 6 | 101 | 87 | 102 | 75 | 94 | 90 | 88 | 8.8 | 7.7 | 6.7 |
| 7 | 102 | 89 | 122 | 86 | 97 | 90 | 88 | 7.6 | 7.6 | 7 |
| 8 | 72 | 63 | 82 | 57 | 67 | 88 | 78 | 8.2 | 8.5 | 7 |
| 9 | 148 | 91 | 120 | 97 | 105 | 74 | 86 | 9.8 | 9 | 8.2 |
| 10 | 116 | 99 | 109 | 95 | 105 | 101 | 101 | 7.2 | 7.3 | 8.2 |
| 11 | 107 | 101 | 102 | 101 | 105 | 97 | 97 | 7.8 | 7.6 | 7 |
| 12 | 131 | 90 | 153 | 93 | 101 | 88 | 84 | 8.4 | 8.8 | 7 |
| 13 | 110 | 75 | 98 | 71 | 97 | 89 | 67 | 7.4 | 8.8 | 7.5 |
| 14 | 84 | 82 | 98 | 93 | 82 | 91 | 86 | 7.4 | 8.2 | 6.9 |
| 15 | 125 | 99 | 112 | 98 | 105 | 105 | 90 | 8.2 | 9.7 | 8 |
| 16 | 110 | 95 | 109 | 91 | 101 | 90 | 94 | 6.5 | 7.7 | 6.4 |
| 17 | 113 | 100 | 112 | 99 | 105 | 96 | 97 | 9.2 | 7 | 6.2 |
| 18 | 95 | 93 | 92 | 97 | 94 | 94 | 90 | 8.2 | 8.5 | 7.5 |
| 19 | 66 | 80 | 72 | 84 | 86 | 95 | 71 | 8 | 8.3 | 7 |
| 20 | 103 | 91 | 120 | 96 | 97 | 72 | 92 | 8 | 7.9 | 8.8 |
| 21 | 142 | 96 | 122 | 92 | 109 | 98 | 67 | 7 | 8.3 | 6.7 |
| 22 | 84 | 91 | 92 | 102 | 82 | 91 | 90 | 7.2 | 7.6 | 6.7 |
| 23 | 116 | 84 | 131 | 90 | 70 | 85 | 101 | 8.6 | 7.6 | 7.3 |
| 24 | 110 | 81 | 98 | 86 | 101 | 74 | 90 | 7.6 | 7.6 | 7.3 |
| 25 | 116 | 82 | 120 | 91 | 101 | 81 | 74 | 8 | 7.3 | 6.7 |
| 26 | 97 | 74 | 98 | 67 | 86 | 79 | 78 | 7.2 | 9.4 | 8.3 |
| 27 | 84 | 72 | 77 | 76 | 74 | 73 | 88 | 8 | 7.9 | 7 |
| 28 | 142 | 106 | 120 | 104 | 109 | 100 | 90 | 7.6 | 5.3 | 5.9 |
| 29 | 84 | 85 | 92 | 97 | 86 | 93 | 82 | 8 | 5.6 | 6.4 |
| 30 | 97 | 81 | 98 | 91 | 97 | 90 | 59 | 8.4 | 6.6 | 6.5 |