Universitat Politècnica de Catalunya · Barcelona Tech - UPC
Escola d'Enginyeria de Barcelona Est EEBE
Homework 3 (Data Management)

Sheik, Abdullahi

October 2023

Problem 2.1

(a) Smoothing by Bin Means

Smoothing by bin means is a method used to smooth data, which helps in reducing the noise within the data. It involves dividing data into bins of a certain size (in this case, a depth of 3) and then replacing all values within a bin by the bin's mean. Here's how 'd do it for the given data:

1. First, we divide the data into bins of depth 3. Given your data set, that would be:
   - Bin 1: 13, 15, 16
   - Bin 2: 16, 19, 20
   - Bin 3: 20, 21, 22
   - Bin 4: 22, 25, 25
   - Bin 5: 25, 25, 30
   - Bin 6: 33, 33, 35
   - Bin 7: 35, 35, 35
   - Bin 8: 36, 40, 45
   - Bin 9: 46, 52, 70

2. Next, we calculate the mean of each bin:

$$
\begin{aligned}
\text{Bin 1 mean} &= \frac{13+15+16}{3} = 14.67 \\
\text{Bin 2 mean} &= \frac{16+19+20}{3} = 18.33 \\
\text{Bin 3 mean} &= \frac{20+21+22}{3} = 21 \\
\text{Bin 4 mean} &= \frac{22+25+25}{3} = 24 \\
\text{Bin 5 mean} &= \frac{25+25+30}{3} = 26.67 \\
\text{Bin 6 mean} &= \frac{33+33+35}{3} = 33.67 \\
\text{Bin 7 mean} &= \frac{35+35+35}{3} = 35 \\
\text{Bin 8 mean} &= \frac{36+40+45}{3} = 40.33 \\
\text{Bin 9 mean} &= \frac{46+52+70}{3} = 56
\end{aligned}
$$

3. Now, we replace all original values in each bin with the corresponding bin mean:

   - Smoothed data: 14.67, 14.67, 14.67, 18.33, 18.33, 18.33, 21, 21, 21, 24, 24, 24, 26.67, 26.67, 26.67, 33.67, 33.67, 33.67, 35, 35, 35, 40.33, 40.33, 40.33, 56, 56, 56

Comment: This technique simplifies the data and helps in reducing the variability within the data. By smoothing the data, we lose some detailed information (like exact individual ages), but it helps to understand the underlying patterns or distributions by reducing noise. However, the choice of bin depth is crucial because too large a bin can oversimplify the data, while too small a bin might not smooth the noise effectively.

(b) Determining Outliers

Determining outliers in the data typically involves statistical methods. Here are a few common methods:

1. Standard Deviation Method: If a value has a distance from the mean that is higher than a certain number of standard deviations, it can be considered an outlier.

2. Interquartile Range (IQR): You first find the first quartile (Q1) and the third quartile (Q3) in the data. The difference (Q3 - Q1) is the IQR.

An outlier is typically defined as a data point that is located outside the fences “Q1 - 1.5 * IQR” or “Q3 + 1.5 * IQR”.

3. Box Plots: This is a visual method of identifying outliers. Points outside the 'whiskers' or 'fences' of the box plot are considered outliers.

In the context of your data, the Standard Deviation Method and IQR would be most effective. For example, using the IQR method, you'd calculate Q1 and Q3, then find the IQR, and any data points outside Q1 - 1.5*IQR or Q3 + 1.5*IQR would be considered outliers. Given the large gap between the majority of your ages and the maximum age of 70, there's a possibility that the 70 could be an outlier, but you'd need to perform the actual calculations or visual inspection through a box plot to be certain.

Problem 2.2

Normalization Methods

Normalization is a technique used to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Here are the value ranges of the normalization methods you've listed:

(a) Min-Max Normalization:

Min-Max Normalization, also known as rescaling, typically rescales the data in the range of [0, 1]. The general formula is:

$$x_{\text{new}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x_{new} is the normalized value, x is the original value, $\min(x)$ is the smallest value in the feature column, and $\max(x)$ is the largest value in the feature column.

Range: [0, 1] (or it can be scaled to any other range [a, b] by multiplying by $b - a$ and then adding a).

(b) Z-score Normalization:

Z-score normalization (or standardization) is the process of rescaling data so that it has a mean (μ) of 0 and a standard deviation (σ) of 1. The formula for z-score normalization is:

$$z = \frac{x - \mu}{\sigma}$$

where z is the z-score normalized value, x is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature.

Range: Typically, most of the data will fall within the range of [-3, 3], but it's not limited to this range; the z-score can technically be any value from negative to positive infinity.

(c) Z-score Normalization Using Mean Absolute Deviation:

When using the Mean Absolute Deviation (MAD) instead of the standard deviation, the formula changes to:

$$z = \frac{x - \mu}{\text{MAD}}$$

where MAD $=$ median($|X_i -$ median$(X)|$) or the average (mean) absolute deviations of the data points from their mean.

Range: As with standard z-scores, most of the data will fall within a certain range (often [-3, 3] if the data follows a normal distribution), but the actual possible range is from negative to positive infinity.

(d) Normalization by Decimal Scaling:

Normalization by decimal scaling changes the range of values by moving the decimal point. The number of decimal places moved depends on the maximum absolute value in the dataset. The formula is:

$$x_{\text{new}} = \frac{x}{10^j}$$

where j is the smallest integer such that $\max(|x_{\text{new}}|) < 1$; it's determined by the maximum absolute value in the dataset.

Range: The values will be in the range of [-1,1] but not limited to this range, as values are not squeezed as in min-max normalization. However, the goal is often to get values within this range.

Problem 2.3

Data Normalization Methods

To normalize the data set [200, 300, 400, 600, 1000] using the specified methods, we need to perform the following calculations:

Original Data

200, 300, 400, 600, 1000

Calculate Basic Statistics

- Minimum (min): 200
- Maximum (max): 1000
- Mean (μ): $(200 + 300 + 400 + 600 + 1000)/5 = 2500/5 = 500$

- Standard Deviation (σ): First, find the variance: $\frac{1}{N}\sum(x_i - \mu)^2$, then take the square root to find σ.
  - Variance: $(1/5) * [(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2] = (1/5) * [90000 + 40000 + 10000 + 10000 + 250000] = (1/5) * [390000] = 78000$
  - Standard Deviation (σ): $\sqrt{78000} \approx 279.51$
- Mean Absolute Deviation (MAD): Calculate the absolute deviations from the mean, then find their mean.
  - MAD: $\frac{1}{N}\sum|x_i - \mu| = (1/5) * [|200-500| + |300-500| + |400-500| + |600-500| + |1000-500|] = (1/5) * [300 + 200 + 100 + 100 + 500] = (1/5) * [1200] = 240$

(a) Min-Max Normalization

The formula for min-max normalization is:

$$x_{\text{new}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

For each x in [200, 300, 400, 600, 1000]:

- 200: $(200 - 200)/(1000 - 200) = 0/800 = 0.0$
- 300: $(300 - 200)/(1000 - 200) = 100/800 = 0.125$
- 400: $(400 - 200)/(1000 - 200) = 200/800 = 0.25$
- 600: $(600 - 200)/(1000 - 200) = 400/800 = 0.5$
- 1000: $(1000 - 200)/(1000 - 200) = 800/800 = 1.0$

Min-Max Normalized Data: 0.0, 0.125, 0.25, 0.5, 1.0

(b) Z-Score Normalization

The formula for z-score normalization is:

$$z = \frac{x - \mu}{\sigma}$$

For each x in [200, 300, 400, 600, 1000]:

- 200: $(200 - 500)/279.51 \approx -1.07$
- 300: $(300 - 500)/279.51 \approx -0.71$
- 400: $(400 - 500)/279.51 \approx -0.36$
- 600: $(600 - 500)/279.51 \approx 0.36$
- 1000: $(1000 - 500)/279.51 \approx 1.79$

Z-Score Normalized Data: -1.07, -0.71, -0.36, 0.36, 1.79

(c) Z-Score Normalization Using Mean Absolute Deviation

The formula is:

$$z = \frac{x - \mu}{\text{MAD}}$$

For each x in [200, 300, 400, 600, 1000]:

- 200: $(200 - 500)/240 = -1.25$
- 300: $(300 - 500)/240 = -0.83$
- 400: $(400 - 500)/240 = -0.42$
- 600: $(600 - 500)/240 = 0.42$
- 1000: $(1000 - 500)/240 = 2.08$

Z-Score (MAD) Normalized Data: -1.25, -0.83, -0.42, 0.42, 2.08

(d) Normalization by Decimal Scaling

The formula is:

$$x_{\text{new}} = \frac{x}{10^j}$$

j is the smallest integer such that $\max(|x_{\text{new}}|) < 1$. Here, j would be 4 because the maximum absolute value in the original data is 1000, and we need to move the decimal point 4 places to make it less than 1.

For each x in [200, 300, 400, 600, 1000]:

- 200: $200/10^4 = 0.02$
- 300: $300/10^4 = 0.03$
- 400: $400/10^4 = 0.04$
- 600: $600/10^4 = 0.06$
- 1000: $1000/10^4 = 0.1$

Decimal Scaled Data: 0.02, 0.03, 0.04, 0.06, 0.1

These are the normalized data using the four methods mentioned. Each method has its specific use cases, advantages, and disadvantages depending on the context in which it's used.

Problem 2.4

To perform the different normalization techniques on the value 35 from the given age data, we first need to recall the original data and compute some basic statistics.
Original Age Data: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

1. **Calculate basic statistics:**
   - Minimum (min): 13
   - Maximum (max): 70
   - Mean (μ): $\frac{1}{27}\sum_{i=1}^{27} x_i = \frac{1}{27} * 802 = 29.70$ (approximated)
   - Standard Deviation (σ): Given as 12.94

Now, we can proceed with the normalizations.

(a) **Min-Max Normalization:**

The formula for min-max normalization to the range [0,1] is:

$$x_{\text{new}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

For the age 35:

$$x_{\text{new}} = \frac{35 - 13}{70 - 13} = \frac{22}{57} \approx 0.386$$

(b) **Z-Score Normalization:**

The formula for z-score normalization is:

$$z = \frac{x - \mu}{\sigma}$$

For the age 35:

$$z = \frac{35 - 29.70}{12.94} \approx \frac{5.3}{12.94} \approx 0.410$$

(c) **Normalization by Decimal Scaling:**

The formula for decimal scaling is:

$$x_{\text{new}} = \frac{x}{10^j}$$

Here, j is determined by the maximum absolute value in the dataset. The highest value in the data is 70, so $j = 2$ (as $10^2 = 100$, and $70/100$ is less than 1).

For the age 35:

$$x_{\text{new}} = \frac{35}{10^2} = \frac{35}{100} = 0.35$$

(d) **Comment on Preferred Method:**

The choice of normalization method really depends on the specific use case and the nature of the data.

- **Min-Max Normalization** is simple and bounds the data into a specific range, which is particularly useful for algorithms that are sensitive to the scale of the input data (like neural networks) or when consistency of input data range across different features is required. However, it is sensitive to outliers, as extreme values can distort the normalization of all other data points.
- **Z-Score Normalization (Standardization)** is beneficial when the data has a Gaussian distribution, or when you want to assume a Gaussian distribution for your machine learning algorithm (as many algorithms expect this). It handles outliers by not bounding values to a specific range.
- **Normalization by Decimal Scaling** is the least common method used in practice for machine learning as it doesn't handle outliers well and doesn't relate data to a standard statistical distribution. However, it can be useful for simply reducing the size of numbers while maintaining their relative scale.

In the context of age data in this example, if we expect other features to be scaled between 0 and 1, or if we are using algorithms sensitive to feature scale, min-max normalization might be preferable for consistency. However, if we are concerned about potential outliers or want to maintain a representation of the data's distribution, z-score normalization could be a better choice. The presence of a significant age outlier (age 70) in this dataset might make z-score normalization a more attractive choice as it reduces the influence of this outlier during the rescaling process.

Problem 2.5

Solve the problems using the given data for age and body fat percentage.

Z-Score Normalization (a)

The z-score of a value essentially tells you how many standard deviations you are from the mean of the dataset. The formula for calculating the z-score is:

$$z = \frac{(X - \mu)}{\sigma}$$

where:

- X is a value from the dataset,

- μ is the mean of the dataset, and
- σ is the standard deviation of the dataset.

First, we need to calculate the mean (μ) and standard deviation (σ) for both 'age' and '%fat'.

1. 'age' attribute:

Mean (μ_{age}) is calculated as follows:

$$\mu_{\text{age}} = \frac{(23+23+27+27+39+41+47+49+50+52+54+54+56+57+58+58+60+61)}{18} = \frac{895}{18} \approx 49$$

To calculate the standard deviation, we need the variance first (the average of the squared differences from the Mean). Let's calculate the variance for 'age':

$$\sigma^2_{\text{age}} = \frac{\sum(x_i - \mu_{\text{age}})^2}{n}$$

$$\sigma^2_{\text{age}} = \frac{(23-49.72)^2 + (23-49.72)^2 + \ldots + (61-49.72)^2}{18}$$

$$\sigma^2_{\text{age}} \approx 165.35$$

Now, the standard deviation (σ_{age}) will be the square root of the variance:

$$\sigma_{\text{age}} = \sqrt{165.35} \approx 12.86$$

2. '%fat' attribute:

Mean (μ_{fat}) is calculated as follows:

$$\mu_{\text{fat}} = \frac{(9.5+26.5+7.8+17.8+31.4+25.9+27.4+27.2+31.2+34.6+42.5+28.8+33.4+30.2+34.1+3}{18}$$

Similarly, we calculate the variance for '%fat':

$$\sigma^2_{\text{fat}} = \frac{\sum(x_i - \mu_{\text{fat}})^2}{n}$$

$$\sigma^2_{\text{fat}} = \frac{(9.5-30.54)^2 + (26.5-30.54)^2 + \ldots + (35.7-30.54)^2}{18}$$

$$\sigma^2_{\text{fat}} \approx 74.61$$

Now, the standard deviation (σ_{fat}) will be the square root of the variance:

$$\sigma_{\text{fat}} = \sqrt{74.61} \approx 8.64$$

Now, we can calculate the z-scores for both attributes.

Correlation Coefficient (b)

Pearson's correlation coefficient (r) between two variables is calculated by the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

where:

- x_i and y_i are individual data points,
- μ_x and μ_y are the means of the respective datasets,
- n is the total number of data points.

The covariance between two variables is calculated by the following formula:

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Based on the correlation coefficient value (r):

- If $r > 0$, it indicates a positive correlation.
- If $r < 0$, it indicates a negative correlation.
- If $r = 0$, it indicates no correlation.