

Tutorial 6

Tutorial on Preprocessing

Luke Chang

The University of Auckland

April 2021

Topics

1 Data Cleaning - Noisy Data

2 Data Transformation by Normalization

Data Cleaning - Noisy Data

Question 1

Give the following data (in ascending order) for the attribute *age*:

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 Use *smoothing by bin means* to smooth these data with 3 bins. Illustrate your steps. Comment on the effect of this technique for the given data.
- 2 How might you determine *outliers* in the data?
- 3 What other methods are there for data smoothing?

Question 1.1

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

Use *smoothing by bin means* to smooth these data with 3 bins.

Step 1: Sort the data in ascending order

Step 2: Compute the number of samples per bin using the **equal-frequency** strategy: $27/3 = 9$

Step 3: Partition into **equal-frequency** bins:

- **Bin 1:** 13, 15, 16, 16, 19, 20, 20, 21, 22
- **Bin 2:** 22, 25, 25, 25, 25, 30, 33, 33, 35
- **Bin 3:** 35, 35, 35, 36, 40, 45, 46, 52, 70

Step 4: Smoothing by bin **means**:

- **Bin 1:** 18, 18, 18, 18, 18, 18, 18, 18, 18
- **Bin 2:** 28, 28, 28, 28, 28, 28, 28, 28, 28
- **Bin 3:** 44, 44, 44, 44, 44, 44, 44, 44, 44

The Effect of Binning

Question 1.1

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

Use *smoothing by bin means* to smooth these data with 3 bins.

After smoothing by bin means:

- **Bin 1:** 18, 18, 18, 18, 18, 18, 18, 18, 18
- **Bin 2:** 28, 28, 28, 28, 28, 28, 28, 28, 28
- **Bin 3:** 44, 44, 44, 44, 44, 44, 44, 44, 44

The larger the width, the greater the effect of the smoothing.

The widths for bins are 9, 13 and **35**, respectively.

Binning has the greatest effect on Bin 3 ($\text{Width}_{\text{Bin 3}} = 70 - 44 = 26$).

Remove Outliers by Clustering

Question 1.2

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

How might you determine *outliers* in the data?

Outlier analysis using Clustering: We split the data into 3 groups, or “clusters”. Once we obtain the mean and standard deviation for each cluster, then we can use the 95% rule and set the cut-off point at 2σ .

Step 1: Compute the absolute deviation, $\text{abs}(x - \mu)$

- **Bin 1:** 5, 3, 2, 2, 1, 2, 2, 3, 4
- **Bin 2:** 6, 3, 3, 3, 3, 1, 5, 5, 7
- **Bin 3:** 9, 9, 9, 8, 4, 1, 2, 8, 26

Step 2: Compute means and standard deviations (We had the mean from binning.)

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Remove Outliers by Clustering

Question 1.2

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

How might you determine *outliers* in the data?

	Mean	σ	2σ
Bin 1	18	2.9	6
Bin 2	28	4.4	9
Bin 3	44	10.9	22

Step 3: Find all data points that are above the threshold

- **Bin 1:** 5, 3, 2, 2, 1, 2, 2, 3, 4
- **Bin 2:** 6, 3, 3, 3, 3, 1, 5, 5, 7
- **Bin 3:** 9, 9, 9, 8, 4, 1, 2, 8, **26**

We identify the data point “70” is an outlier.

Other Methods for Data Smoothing

Question 1.3

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

What other methods are there for data smoothing?

What tools are available?

- Binning: Done
- Regression: Not the best choice for 1D data.
- Clustering: Unsupervised learning, e.g., K-means algorithm (The example is in the notebook.)
- Moving average

Moving Average

Question 1.3

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

Smooth the data using moving average method with a fixed window size of 3.

Step 1: Sort the data in ascending order

Step 2: Add padding based on the window size

13, 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, 70

Step 3: Update each data point based on its neighbour values

$$x'_i = \frac{x_{i-1} + x_i + x_{i+1}}{3}$$

$$x'_1 = \frac{13 + 13 + 15}{3} = \frac{41}{3} \approx 13.7 = 14$$

After smoothing (Round to 0 d.p.)

14, 15, 16, 17, 18, 20, 20, 21, 22, 23, 24, 25, 25, 27, 29, 32, 34, 34, 35, 35, 35, 37, 40, 44, 48, 56, 64

Data Transformation by Normalization

Question 2

Use these methods to normalize the following data points:

200, 300, 400, 600, 1000

- ① Min-max normalization by setting $\min = 0$ and $\max = 1$
- ② Z-score normalization
- ③ Z-score normalization using the mean absolute deviation instead of standard deviation
- ④ Decimal scaling

What are the value ranges of the following normalization methods?

Min-max normalization

Question 2.1

200, 300, 400, 600, 1000

Min-max normalization by setting $\min = 0$ and $\max = 1$

Question: What are the value ranges of min-max normalization?

You can set *min* and *max* to any value, but the most commonly used ranges are $[0, 1]$ and $[-1, 1]$.

Step 1: Find $\min = 200$ and $\max = 1000$

Step 2: Apply the formula to each data point

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$0, \frac{300 - 200}{800}, \frac{400 - 200}{800}, \frac{600 - 200}{800}, 1$$
$$0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1$$

Z-score Normalization

Question 2.2

Apply z-score normalization to 200, 300, 400, 600, 1000

Question: What are the value ranges of z-score normalization?

Centred at 0. It can be any real number. However, it's unlikely to get extreme values.

Step 1: Compute mean μ and standard deviation σ

$$\mu = \frac{200 + 300 + 400 + 600 + 1000}{5} = 500$$

$$\sigma = \sqrt{\frac{1}{5} \sum_{i=1}^5 (x_i - \mu)^2} = \sqrt{\frac{(-300)^2 + (-200)^2 + (-100)^2 + 100^2 + 500^2}{5}} = \sqrt{80000} = 282.8$$

Step 2: Apply the z-score formula to each data point

$$z_i = \frac{x_i - \mu}{\sigma}$$

Results: $-1.06, -0.71, -0.35, 0.35, 1.77$ (round at 2 d.p.)

Z-score Normalization Using the Mean Absolute Deviation

Question 2.3

Apply z-score normalization using the **mean absolute deviation** to 200,300,400,600,1000

Step 1: We already have $\mu = 500$.

Step 2: Compute absolute deviation

X	deviation	absolute deviation
200	-300	300
300	-200	200
400	-100	100
600	100	100
1000	500	500

Step 3: Compute Mean Absolute Deviation (M.A.D)

$$s_X = \frac{300 + 200 + 100 + 100 + 500}{5} = \frac{1200}{5} = 240$$

Z-score Normalization Using the Mean Absolute Deviation

Question 2.3

Apply z-score normalization using the **mean absolute deviation** to 200,300,400,600,1000

Step 4: Apply the z-score normalization using the mean absolute deviation to each data point

$$z'_i = \frac{x_i - \mu}{s_X}$$

Results: $-1.25, -0.83, -0.42, 0.42, 2.08$ (round at 2 d.p.)

Motivation: Since $|x_i - \mu|$ is not squared, the effect of outliers is reduced.

Question 2.4

Apply decimal scaling to 200,300,400,600,1000

Question: What are the value ranges of decimal scaling?
Within the range of $(-1, 1)$; Exclude -1 and 1.

A value, x_i , of X is normalized to x'_i by computing:

$$x'_i = \frac{x_i}{10^j}$$

Where j is the smallest integer such that $\max(|x'_i|) < 1$.

- 1 Find the range of the data points: $[200, 1000]$.
- 2 The maximum absolute value is 1000.
- 3 Therefore, we divide each value by 10000, (i.e., $j = 4$)

Results: 0.02, 0.03, 0.04, 0.06, 0.1