นายชวัลวิทย์ เกียรติณัฐกร
6510503310

Suppose that the data for analysis includes the attribute *Number*. The *Number* values for the dataset are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Answer the following questions.

(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

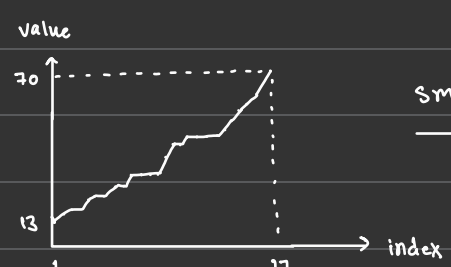
Partition into bins

1: 13, 15, 16
2: 16, 19, 20
3: 20, 21, 22
4: 22, 25, 25
5: 25, 25, 30
6: 33, 33, 35
7: 35, 35, 35
8: 36, 40, 45
9: 46, 52, 70

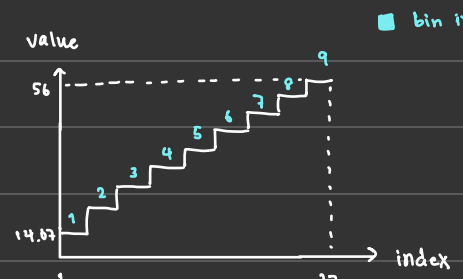
Smoothing by bin means:

1: 14.67, 14.67, 14.67
2: 18.33, 18.33, 18.33
3: 21, 21, 21
4: 24, 24, 24
5: 26.67, 26.67, 26.67
6: 33.67, 33.67, 33.67
7: 35, 35, 35
8: 40.33, 40.33, 40.33
9: 56, 56, 56

หลังจากทำการ smoothing data แล้วพบว่า data ที่อยู่ใน bin เดียวกันจะมีค่าเหมือนกันทั้งหมด ส่งผลให้ความซับซ้อนของ data ลดลง ลดผลกระทบของค่า outlier ที่มีต่อ data สังเกตได้จากค่า 70 ที่ถูกแปลงเป็น 56 โดยสรุปแล้วหลังจากทำการ smoothing เป็นการลด noise ที่มีผลต่อการวิเคราะห์ data ให้วิเคราะห์ได้ง่ายมากยิ่งขึ้น



กราฟคร่าวๆ ของ data ก่อน smoothing



กราฟคร่าวๆ ของ data หลัง smoothing

(b) How might you determine outliers in the data?

n (data size) = 27

Q1 position = $\frac{27+1}{4}$ = 7

Q1 value = 20

Q3 position = $\left(\frac{27+1}{4}\right) \times 3$ = 21

Q3 value = 35

IQR = Q3 - Q1

= 35 - 20 = 15

Q1 - 1.5 IQR = 20 - 1.5(15) = -2.5

Q3 + 1.5 IQR = 35 + 1.5(15) = 57.5

จาก Q1 - 1.5 IQR = -2.5 และ Q3 + 1.5 IQR = 57.5

จะได้ว่า ค่าที่ไม่อยู่ในช่วง [-2.5, 57.5] เป็น outlier

ดังนั้น มีค่า outlier ค่าเดียวคือ 70 #

(c) What other methods are there for data smoothing?

นอกจากวิธี bin means ก็สามารถใช้วิธีอื่นได้เช่นกัน

1.) bin medians : ทุกค่าใน bin จะถูกแทนที่ด้วยค่า median ในแต่ละ bin นั้น ๆ

Partition into bins :

1: 13,15,16
2: 16,19,20
3: 20,21,22
4: 22,25,25
5: 25,25,30
6: 33,33,35
7: 35,35,35
8: 36,40,45
9: 46,52,70

Smoothing by bin medians :

1: 15,15,15
2: 19,19,19
3: 21,21,21
4: 25,25,25
5: 25,25,25
6: 33,33,33
7: 35,35,35
8: 40,40,40
9: 52,52,52

2.) bin boundaries : ใช้ค่า min, max ของแต่ละ bin ในการคำนวณ โดยมีหลักการแทนค่าดังนี้ หากค่าที่พิจารณามีค่าใกล้เคียงกับ min มากกว่า max ก็จะแทนด้วยค่า min และในทางกลับกันหากมีค่าใกล้กับ max มากกว่า ก็จะแทนด้วยค่า max

Partition into bins:

1: 13,15,16
2: 16,19,20
3: 20,21,22
4: 22,25,25
5: 25,25,30
6: 33,33,35
7: 35,35,35
8: 36,40,45
9: 46,52,70

Smoothing by bin boundaries :

1: 13,16,16
2: 16,20,20
3: 20,20,22 ← เลือกแทนด้วยค่า min
4: 22,25,25
5: 25,25,30
6: 33,33,35
7: 35,35,35
8: 36,36,45
9: 46,46,70

(d) Use min-max normalization to transform the value 35 for *Number* onto the range [0.0, 1.0]

จาก $x_{i,new} = \frac{x_i - min(x)}{max(x) - min(x)}$ $\quad x_i \in x$

พิจารณาโดยตัด 70 ที่เป็น outlier ออก

และ $min(x) = 13$, $max(x) = 52$

จะได้ $35_{new} = \frac{35 - 13}{52 - 13}$

$= \frac{22}{39} \approx 0.56$ #

(e) Use z-score normalization to transform the value 35 for *Number*

จาก $x_{i,new} = \frac{x_i - mean(x)}{stdev(x)}$ $\quad x_i \in x$

พิจารณาโดยตัด 70 ที่เป็น outlier ออก

และ $mean(x) = \frac{\sum x_i}{N} \approx 28.42$, $stdev(x) = \frac{\sum (x_i - \mu)^2}{N} \approx 10.37$

จะได้ $35_{new} = \frac{35 - 28.42}{10.37} \approx 0.63$ #

(g) Compare min-max and z-score normalization techniques.

z-score normalization: เหมาะกับการ normalize ข้อมูลแบบ normal หรือใกล้เคียงกับแบบ normal โดยเป็นการ represent ค่าค่าต่าง ๆ อยู่ห่างจากค่า mean อย่างไร (หลังจาก normalize ข้อมูลที่มีค่าเท่ากับ mean จะมีค่าเท่ากับ 0) และช่วงของค่าที่ normalize แล้วไม่ได้ถูกกำหนดช่วงไว้ เป็นได้ทั้งบวกและลบ อีกทั้งยัง sensitive กับ outlier น้อยกว่า min-max normalization

min-max normalization: สามารถใช้กับข้อมูลที่ไม่เป็น normal ได้ หลังจาก normalize แล้ว ค่าที่ได้จะมีช่วงที่ชัดเจน ในกรณีที่มีเฉพาะค่าบวกจะใช้ช่วง [0,1] ส่วนกรณีที่ข้อมูลมีค่าลบด้วยจะใช้ช่วง [-1,1] และ sensitive กับ outlier มากกว่า z-score normalization เหมาะกับการนำไปใช้ Neutral network, Linear regression เป็นต้น