**Problem 1**

**Consider the following dataset (sorted in non-decreasing order):**

**<15, 38, 41, 44, 45, 51, 63, 81, 82, 95, 103, 125, 134, 138, 142>**

1. **Perform the equal width binning on the above data with 3 bins using the method that we discussed in the class. Note that the bin boundaries are integers in the textbook (to make the discussion simple). But, for this assignment your bin boundaries will include fractions. So, you must follow the example in the lecture slides. For each bin, show the bin interval, data values in the bin, and smoothed values using bin means, bin medians, and bin boundaries.**

|  |  |  |
| --- | --- | --- |
| Bin | Interval | Values |
| Bin 1 | [15,57.33) | 15,38,41,44,45,51 |
| Bin 2 | [57.33,99.66) | 63,81,82,95 |
| Bin 3 | [99.66,142] | 103,125,134,138,142 |

|  |  |  |  |
| --- | --- | --- | --- |
| Bin | Smoothed by bin mean | Smoothed by bin median | Smoothed by bin boundary |
| Bin 1 | 39,39,39,39,39,39 | 42.5,42.5,42.5,42.5,42.5,42.5 | 15,51,51,51,51,51 |
| Bin 2 | 80.25,80.25,80.25,80.25 | 81.5,81.5,81.5,81.5 | 63,95,95,95 |
| Bin 3 | 128.4,128.4,128.4,128.4,128.4, | 134,134,134,134,134 | 103,142,142,142,142 |

1. **Repeat the same with equal depth binning with 3 bins.**

|  |  |
| --- | --- |
| Bin | Values |
| Bin 1 | 15, 38, 41, 44, 45 |
| Bin 2 | 51, 63, 81, 82, 95 |
| Bin 3 | 103, 125, 134, 138, 142 |

|  |  |  |  |
| --- | --- | --- | --- |
| Bin | Smoothed by bin mean | Smoothed by bin median | Smoothed by bin boundary |
| Bin 1 | 36.6,36.6,36.6,36.6,36.6 | 41,41,41,41,41 | 15,51,51,51,51 |
| Bin 2 | 74.4,74.4,74.4,74.4,74.4 | 81,81,81,81,81 | 51,51,95,95,95 |
| Bin 3 | 128.4,128.4,128.4,128.4,128.4, | 134,134,134,134,134 | 103,142,142,142,142 |

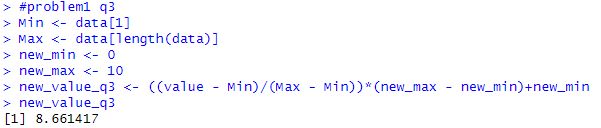
1. **If you transform the dataset into the interval of [0, 10] using Min-max normalization, what is the new value of 125?**

***See hw2.r***

Value before transform:

value

Transform process:



1. **If you transform the dataset using z-score normalization using the standard deviation, what is the new value of 125?**

***See hw2.r***

Value before transform:

value

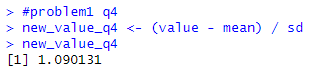
mean:

mean

sd:

sd

Transform process:



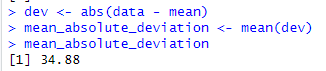
1. **If you transform the dataset using z-score normalization using the mean absolute deviation, what is the new value of 125?**

***See hw2.r***

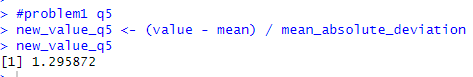
Value before transform:

value

Mean absolute deviation：



Transform process:



**Problem 2**

**This problem is a practice of calculating correlations between some input attributes (or predictive attributes) and the output attribute (or predictable attribute) in the a2-p2.csv dataset. Calculate following correlations:**

**correl(A1, A4)**

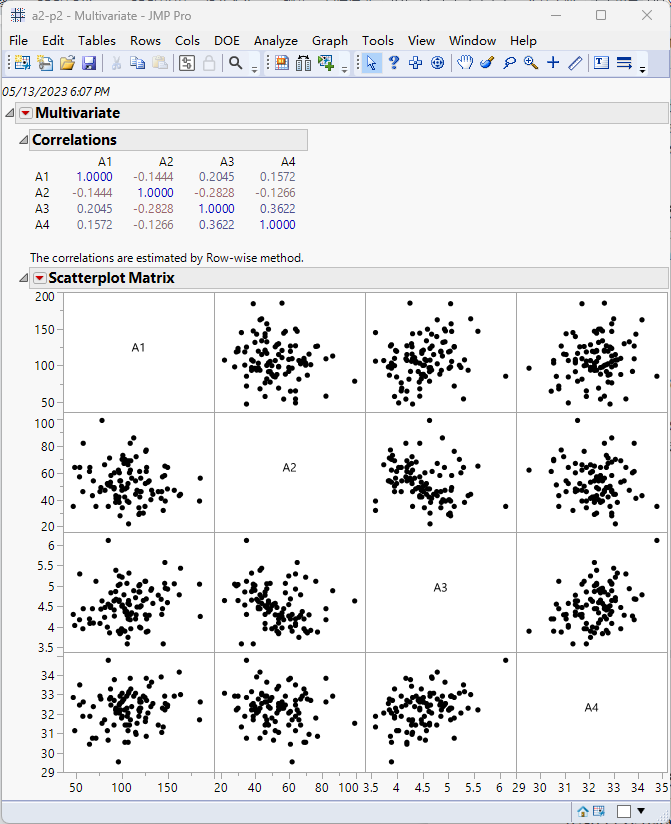
**correl(A2, A4)**

**correl(A3, A4)**

**Here, correl(X, Y) denotes the Pearson’s correlation coefficient between X and Y.**

**In your submission, include all three correlations, and indicate the attribute that has the**

**strongest correlation with A4.**

****

Because is much closer to 1 than any other, A3 has the strongest correlation with A4

**Problem 3**

1. **Determine whether there is a correlation between attribute A1 and attribute A4.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yes | No | Row Total |
| Old | 77 | 36 | 113 |
| Middle | 205 | 80 | 283 |
| Young | 35 | 5 | 40 |

Column for Yes = 35 + 205 + 77=317

Grand Total = 317 + 36 + 80 + 5 = 438

E(Young|Yes) = (40 \* 317) / 438 = 28.95

E(Young|No) = 40 - 28.95 = 11.05

E(Middle|Yes) = (285 \* 317) / 438 = 206.27

E(Middle|No) = 285 - 206.27 = 78.73

E(Old|Yes) = (113\* 317) / 438 = 81.78

E(Old|No) = 113 - 81.78 = 31.22

Compared 5.68684 with the p-value when and signficant level = 5%， which equal to 5.991, the Chi-square statistic is greater than the critical value, rejected, so, there is a significant correlation between the attributes A1 and A4

**(2) Determine whether there is a correlation between attribute A2 and attribute A4.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yes | No | Row Total |
| Low | 57 | 46 | 103 |
| Middle | 161 | 67 | 228 |
| High | 103 | 8 | 111 |

Column for Yes = 35 + 205 + 77=321

Grand Total = 321 + 46 + 67 + 8 = 442

E(Low|Yes) = (103 \* 321) / 442 = 74.8

E(Low|No) = 103 - 74.8 = 28.2

E(Middle|Yes) = (228 \* 321) / 442 = 165.58

E(Middle|No) = 228 - 165.58 = 62.42

E(High|Yes) = (111\* 321) / 442 = 80.61

E(High|No) = 111 - 80.61 = 30.39

Compared with the p-value when and signficant level = 5%， which equal to 5.991, the Chi-square statistic is much greater than the critical value, rejected, so, there is a significant correlation between the attributes A2 and A4

**Problem 4**

1. **.Derive classification rules using the 1R method which we discussed in the class.**

Attribute A1:

Low: 5 tuples, 2 Y’s and 3 N’s

Rules: if A1=Low, then class=Y, error rate=3/5

Medium: 4 tuples, 3 Y’s and 1 N’s

Rules: if A1=Medium, then class=Y, error rate=1/4

High: 3 tuples, 2 Y’s and 1 N’s

Rules: if A1=High, then class=Y,error rate=1/3

Total Error=5/12

Attribute A2:

Mild: 5 tuples, 3 Y’s and 2N’s

Rules: if A2=Mild, then class=Y,error rate=2/5

Cool: 4 tuples, 2 Y’s and 2N’s

Rules: if A2=Cool, then class=Y,error rate=2/4

Hot: 3 tuples, 2 Y’s and 1N’s

Rules: if A2=Hot, then class=Y,error rate=1/3

Total Error=5/12

Attribute A3:

East: 8 tuples, 7 Y’s and 1 N’s

Rules: if A3=East, then class=Y,error rate=1/8

West: 4 tuples, 0 Y’s and 4N’s

Rules: if A3=West, then class=N,error rate=0/4

Total Error=1/12

So the 1R is:

if A3= East, then class = Y

if A3= West, then class = N

1. **.Classify a new instance X = (A1 = Medium, A2 = Cool, A3 = East) using the rules.**

In this instance, A3=East, by the rule in the Q1, we can get that the class=Y.

**Problem 5**

**Suppose we have a new tuple X = (A1 = Medium, A2 = Cool, A3 = East). Predict the class label of X using Naïve Bayes classification. Reminder: your calculation must include probability information obtained from each of the three attributes.**

Class Y: 7/12

Class N: 5/12

For A1 = Medium:

P(A1 = Medium | Class = Y) = 4/7

P(A1 = Medium | Class = N) = 1/5

For A2 = Cool:

P(A2 = Cool | Class = Y) = 2/7

P(A2 = Cool | Class = N) = 2/5

For A3 = East:

P(A3 = East | Class = Y) = 6/7

P(A3 = East | Class = N) = 1/5

P(Class = Y | X) = P(Class = Y) \* P(A1 = Medium | Class = Y) \* P(A2 = Cool | Class = Y) \* P(A3 = East | Class = Y)

= (7/12) \* (4/7) \* (2/7) \* (6/7) = 0.040816

P(Class = N | X) = P(Class = N) \* P(A1 = Medium | Class = N) \* P(A2 = Cool | Class = N) \* P(A3 = East | Class = N)

= (5/12) \* (1/5) \* (2/5) \* (1/5) = 0.003333

These probabilities are not yet normalized, but we can see that P(Class = Y | X) is larger than P(Class = N | X). Therefore, we would predict the class label of X to be Y. However, if you want to have a proper probability distribution, you need to normalize these probabilities:

P'(Class = Y | X) = P(Class = Y | X) / (P(Class = Y | X) + P(Class = N | X)) = 0.040816 / (0.040816 + 0.003333) = 0.924528

P'(Class = N | X) = P(Class = N | X) / (P(Class = Y | X) + P(Class = N | X)) = 0.003333 / (0.040816 + 0.003333) = 0.075472

So, after normalization, the predicted class label of X remains as 'Y' with a probability of 0.924528.

**Problem 6**

***Weka***

