Yiduo Feng

CS 699

Assignment 2

Date: 5/24/2022

**Problem 1** (**15 points)** 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.

Sorted data: 15, 38, 41, 44, 45, 51, 63, 81, 82, 95, 103, 125, 134, 138, 142

• Partition into equal‐width bins: (142 – 15) / 3 = 42.33

• Bin intervals are: [15, 57.33), [57.33, 99.66), [99.66, 142]

‐ Bin 1: 15, 38, 41, 44, 45, 51

‐ Bin 2: 63, 81, 82, 95

‐ Bin 3: 103, 125, 134, 138, 142

• Smoothing by bin means:

‐ Bin 1: 39, 39, 39, 39, 39, 39

‐ Bin 2: 80.25, 80.25, 80.25, 80.25

‐ Bin 3: 128.4, 128.4, 128.4, 128.4, 128.4

• Smoothing by bin boundaries:

‐ Bin 1: 15, 51, 51, 51, 51, 51

‐ Bin 2: 63, 95, 95, 95

‐ Bin 3: 103, 142, 142, 142, 142

(2) Repeat the same with equal depth binning with 3 bins.

• Sorted data: 15, 38, 41, 44, 45, 51, 63, 81, 82, 95, 103, 125, 134, 138, 142

• Partition into equal‐depth bins:

‐ Bin 1: 15, 38, 41, 44, 45

‐ Bin 2: 51, 63, 81, 82, 95

‐ Bin 3: 103, 125, 134, 138, 142

• Smoothing by bin means:

‐ Bin 1: 36.6, 36.6, 36.6, 36.6, 36.6

‐ Bin 2: 74.4, 74.4, 74.4, 74.4, 74.4

‐ Bin 3: 128.4, 128.4, 128.4, 128.4, 128.4

• Smoothing by bin boundaries:

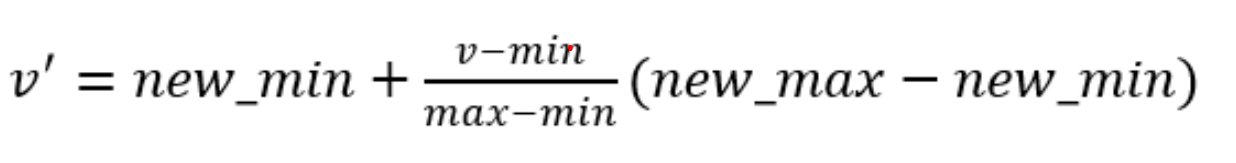
‐ Bin 1: 15, 45, 45, 45, 45

‐ Bin 2: 51, 51, 95, 95, 95

‐ Bin 3: 103, 142, 142, 142, 142

(3) If you transform the dataset into the interval of [0, 10] using Min-max

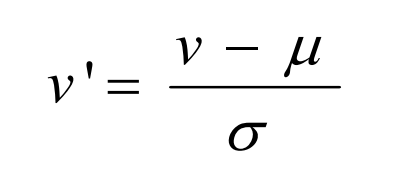
normalization, what is the new value of 125?



0 + ((125 - 15)/(142 - 15))\*(10 - 0) = 8.661417

(4) If you transform the dataset using z-score normalization using the standard

deviation, what is the new value of 125?



σ = (15+38+41+44+45+51+63+81+82+95+103+125+134+138+142)/ 15

= 1197/15

= 41.46289

μ = Σ(xi - μ)^2/N

= ((15 - 79.8)^2 + ... + (142 - 79.8)^2)/15

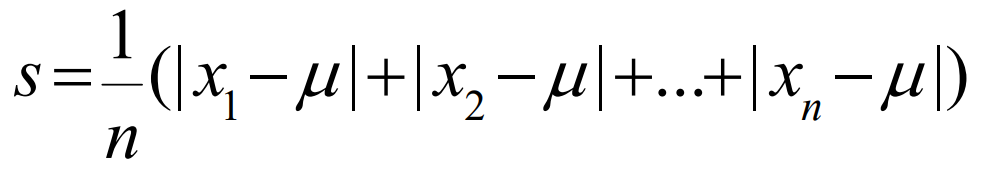
= 24068.4/15

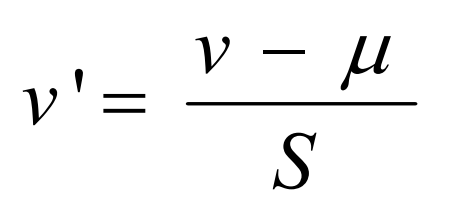
= 79.8

(125 - 79.8)/41.46289 = 1.090131

(5) If you transform the dataset using z-score normalization using the mean absolute

deviation, what is the new value of 125?





S = (|15 - 79.8|+...+|142 - 79.8|)\*1/15

= (64.8 + 41.8 + 38.8 + 35.8 + 34.8 + 28.8 + 16.8 + 1.2 + 2.2 + 15.2 + 23.2 + 45.2 + 54.2 + 58.2 + 62.2)\*1/15

= 523.2/15

= 34.88

(125 - 79.8)/34.88 = 1.295872

**Problem 2 (15 points)** 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:

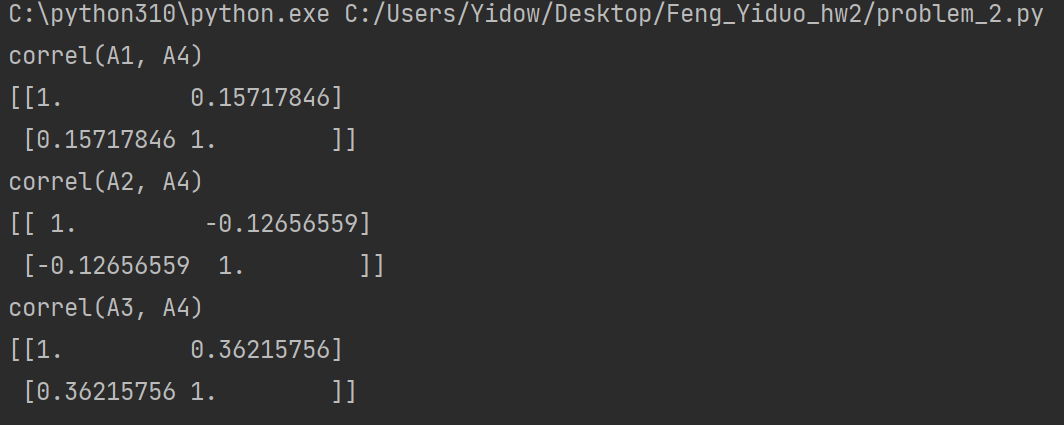
correl(A1, A4)

correl(A2, A4)

correl(A3, A4)

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

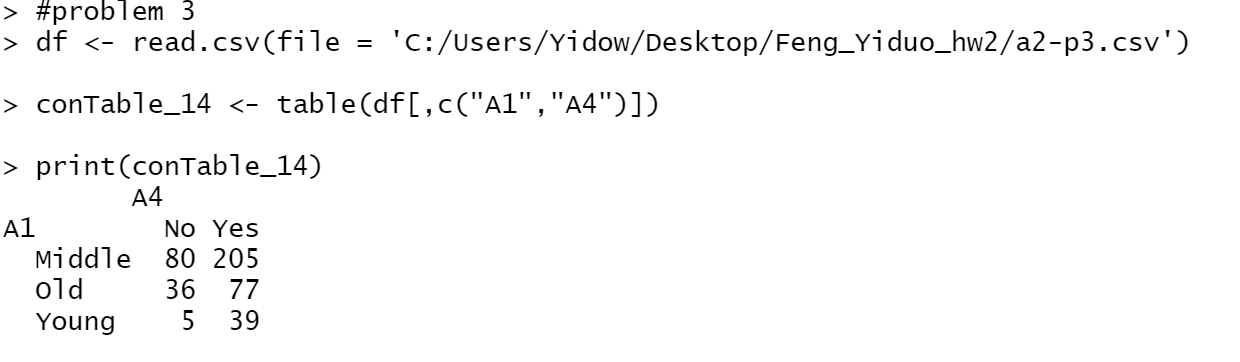
In your submission, include all three correlations, and indicate the attribute that has the strongest correlation with A4.



According to the results above, A3 has the strongest correlation with A4.

**Problem 3 (15 points)** This problem is a practice of determining correlation between two nominal attributes using the chi-square test, which we discussed in the class. Consider the *a2-p3.csv* dataset.

1. Determine whether there is a correlation between attribute *A*1 and attribute *A*4.



According to the results above, we can get the contingency table.

|  |  |  |  |
| --- | --- | --- | --- |
| A1 A4 | No | Yes | Sum |
| Middle | 80(78.02036) | 205(206.9796) | 285 |
| Old | 36(30.93439) | 77(82.06561) | 113 |
| Young | 5(12.04525) | 39(31.95475) | 44 |
| Sum | 121 | 321 | 442 |

Calculate expected values:

(285\*121)/442 = 78.02036

(285\*321)/442 = 206.9796

(113\*121)/442 = 30.93439

(113\*321)/442 = 82.06561

(44\*121)/442 = 12.04525

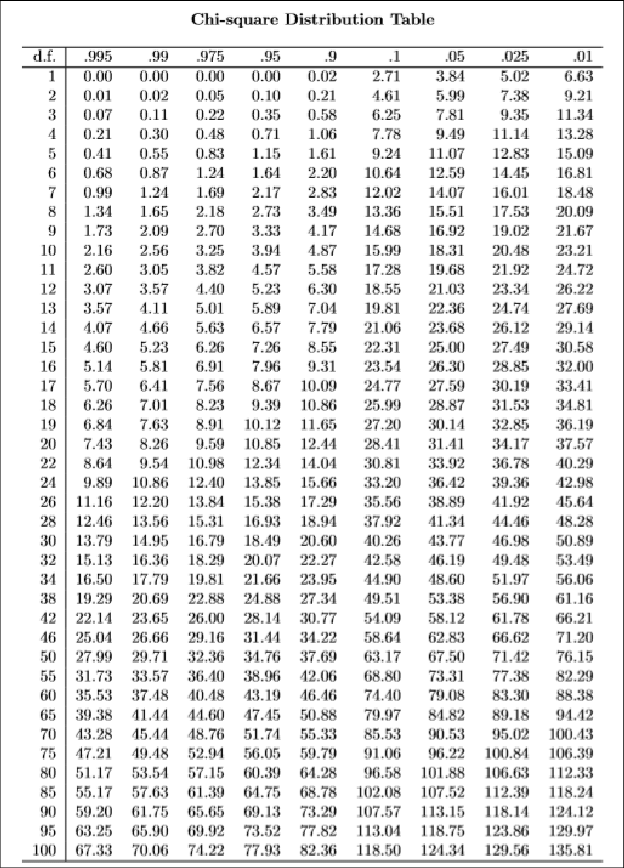
(44\*321)/442 = 31.95475

Calculate X2 test statistic:

X^2 = ((80-78.02036)^2)/78.02036+((205-206.9796)^2)/206.9796+((36-30.93439)^2)/30.93439+((77-82.06561)^2)/82.06561+((5-12.04525)^2)/12.04525+((39-31.95475)^2)/31.95475= 6.885419964

Alpha = 0.05

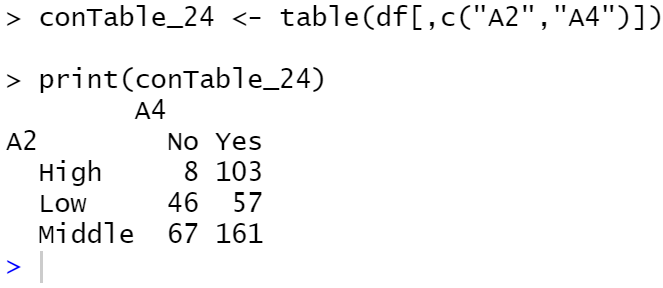
Degree of freedom = (num\_rows – 1) \* (num\_cols – 1) = 2 \* 1 = 2



Look up the chi‐square distribution table, X^2 0.05,2 = 5.99 < 6.885

So X^2 > X^2 df,alpha, there is a correlation between A1 and A4.

1. Determine whether there is a correlation between attribute *A*2 and attribute *A*4.



According to the results above, we can get the contingency table.

|  |  |  |  |
| --- | --- | --- | --- |
| A2 A4 | No | Yes | Sum |
| High | 8 | 103 | 111 |
| Low | 46 | 57 | 103 |
| Middle | 67 | 161 | 228 |
| Sum | 121 | 321 | 442 |

Calculate expected values:

(111\*121)/442 = 30.38688

(111\*321)/442 = 80.61312

(103\*121)/442 = 28.19683

(103\*321)/442 = 74.80317

(228\*121)/442 = 62.41629

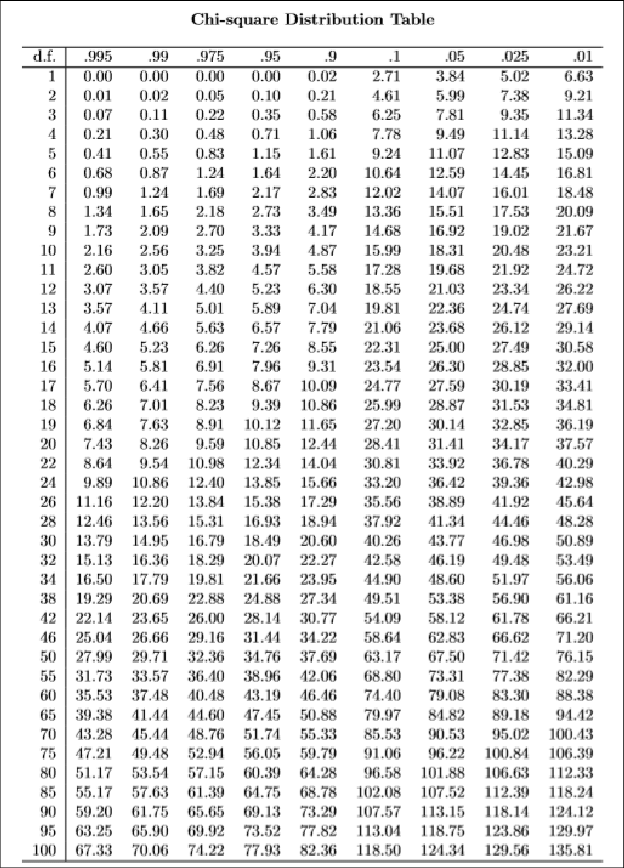
(228\*321)/442 = 165.5837

Calculate X2 test statistic:

X^2 = ((8-30.38688)^2)/30.38688+((103-80.61312)^2)/80.61312+((46-28.19683)^2)/28.19683+((57-74.80317)^2)/74.80317+((67-62.41629)^2)/62.41629+((161-165.5837)^2)/165.5837= 38.65144

Alpha = 0.05

Degree of freedom = (num\_rows – 1) \* (num\_cols – 1) = 2 \* 1 = 2



Look up the chi‐square distribution table, X^2 0.05,2 = 5.99 < 38.65144

So X^2 > X^2 df,alpha, there is a correlation between A2 and A4.

You can use any tool when creating contingency tables. However, you must calculate the chi-square test statistic yourself, including the expected values. You must not use

software, such as JMP Pro, R, or Python, to obtain the value of the test statistic. Use 5% significance level.

**Problem 4 (20 points)** Consider the following dataset:

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

Make rules for each attribute and calculate error rate:

Attribute A1

A1= Medium→ yes(3/4), error rate = 1/4

A1= Low→ no(3/5), error rate = 2/5

A1= High→ yes (2/3), error rate = 1/3

total error = 4/12

Attribute A2

A2= Mild→ yes(3/5), error rate = 2/5

A2= Cool→ no(2/4), error rate = 2/4

(arbitrary tie breaking)

A2= Hot→ yes (2/3), error rate = 1/3

total error = 5/12

Attribute A3

A3= East→ yes(7/8), error rate = 1/8

A3= West→ no(4/4), error rate = 0/4

total error = 1/12

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

According to (1),

A1 = Medium -> yes

A2 = Cool -> no/yes (the error rate is 2/4)

A3 = East -> yes

So class label of X = yes

**Problem 5 (20 points)** Consider the following dataset:

Suppose we have a new tuple *X* = (*A1* = Medium, *A2* = Cool, *A3* = East). Predict the class label of *X* using Naïve Bayes classification.

• Class prior probabilities are:

P(C1) = P( yes) = 7/12 = 0.5833333

P(C2) = P(no) = 5/12= 0.4166667

• For yes class:

P(A1= Medium| class = yes) = 3/7= 0.4285714

P(A2= Cool| class = yes) = 2/7 = 0.2857143

P(A3= East| class = yes) = 7/7 = 1.0

P(X|class = yes) = 0.4285714 \* 0.2857143 \* 1.0 = 0.122449

• For no class

P(A1= Medium| class = no) = 1/5= 0.2

P(A2= Cool| class = no) = 2/5 = 0.4

P(A3= East| class = no) = 1/5 = 0.2

P(X|class = no) = 0.2\*0.4\*0.2 = 0.016

P(X|Ci)\*P(Ci) :

P(X|class = yes) \* P(class = yes) = 0.122449 \* 0.5833333 = 0.07142858

P(X|class = no) \* P(class = no) = 0.016 \* 0.4166667 = 0.006666667 < 0.07142858

The model predicts that X belongs to class yes.

**Problem 6 (15 points)** This problem is a practice of performing classification using a

tool. You may use Weka, JMP Pro, R, or Python. If you use Weka or JMP Pro, you must submit relevant screenshots. If you use R or Python, you must submit the R or Python programs you used for this assignment. If you never used data mining (or machine learning) tools for classification, we suggest that you use either Weka or JMP Pro. How to perform classification with Naïve Bayes using Weka or JMP Pro is described in a separate file. If you learned how to use R or Python for classification in other courses, then you may use R, Python, Weka, or JMP Pro.

(3). If you use R or Python

 Split the dataset into training and test sets with the 66%-34% ratio.

 Run Naïve Bayes on *autism-adult-a2.csv*

 Calculate the prediction accuracy on the validation dataset.

 Also submit your R or Python scripts.

