Assignment 2

Note: Show all your work. You can do manual calculations, use R, or use any software (e.g., Weka, Excel, JMP) to answer the questions, unless otherwise noted. In any case, you need to attach the relevant file(s) or screenshot(s) that shows how you obtained your answers.

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.
(2) Repeat the same with equal depth binning with 3 bins.
(3) If you transform the dataset into the interval of [0, 10] using Min-max normalization, what is the new value of 125?
(4) If you transform the dataset using z-score normalization using the standard deviation, what is the new value of 125?
(5) If you transform the dataset using z-score normalization using the mean absolute deviation, what is the new value of 125?

Note: For Problem 1-(4) and Problem 1-(5), you need to show the mean, standard deviation, mean absolute deviation, and the new, transformed value as well as all calculation steps.

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.

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

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:

ID	A1	A2	A3	Class
1	Medium	Mild	East	Y
2	Low	Mild	East	Y
3	High	Mild	East	N
4	Low	Mild	West	N
5	Low	Cool	East	Y
6	Medium	Hot	West	N
7	High	Hot	East	Y
8	Low	Cool	West	N
9	Medium	Hot	East	Y
10	High	Cool	East	Y
11	Medium	Mild	East	Y
12	Low	Cool	West	N

(1). Derive classification rules using the 1R method which we discussed in the class.
(2). Classify a new instance *X* = (*A1* = Medium, *A2* = Cool, *A3* = East) using the rules.

You must derive rules manually without using a data mining software.

Problem 5 (20 points) Consider the following dataset:

ID	A1	A2	A3	Class
1	Medium	Mild	East	Y
2	Low	Mild	East	Y
3	High	Mild	East	N
4	Low	Mild	West	N
5	Low	Cool	East	Y
6	Medium	Hot	West	N
7	High	Hot	East	Y
8	Low	Cool	West	N
9	Medium	Hot	East	Y
10	High	Cool	East	Y
11	Medium	Mild	East	Y
12	Low	Cool	West	N

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

You must solve this problem manually without using a data mining software.

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.

(1). If you use Weka:

- Run Naïve Bayes on *autism-adult-a2.arff.*
- Choose "Percentage split" for test options and specify 66%.
- Calculate the prediction accuracy on the test dataset.
- Also submit the screenshot of the output, which shows performance measures

(2). If you use JMP Pro

- Run Naïve Bayes on *autism-adult-a2.jmp* (validation column was already created).
- Calculate the prediction accuracy on the validation dataset.
- Also submit the screenshot of the output, which shows performance measures and ROC curves.

(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.

Submission:

Name your file *LastName_FirstName*_HW2.docx (or *LastName_FirstName*_HW2.pdf). If you have multiple files, then combine all files into a single archive file. Name the archive file *LastName_FirstName*_HW2.EXT. Here, “EXT” is an appropriate archive file extension (e.g., zip or rar). Upload this archive file to Blackboard