Assignment 3

Due: 2/13

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 (**20 points)** Consider the following dataset (sorted in non-decreasing order): <10, 12, 16, 16, 29, 32, 51, 60, 60, 66, 70, 72, 87, 96, 120>

(1) Perform the equal width binning on the above data with 3 bins. Note that the bin boundaries are integers in the textbook and in the online lecture module (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, 1] using Min-max normalization, what is the new value of 51?
(4) If you transform the dataset using z-score normalization using the standard deviation, what is the new value of 51?
(5) If you transform the dataset using z-score normalization using the mean absolute deviation, what is the new value of 51?

Problem 2 (10 points) This problem is a practice of calculating correlations between input attributes (or predictive attributes) and the output attribute (or predictable attribute) in the *a3-p2.csv* dataset. This dataset has 5 attributes and 100 tuples. The first 4 attributes are input attributes and the last attribute, *A5*, is the output attribute. Your task is to calculate the correlation between each input attribute and the output attribute. In other words, you are required to calculate the following four correlations:

correl(*A1*, *A5*)
correl(*A2*, *A5*)
correl(*A3*, *A5*)
correl(*A4*, *A5*)

Here, *correl*(*X*, *Y*) denotes the correlation between *X* and *Y*.

In your submission, include all four correlations, and indicate the attribute that has the strongest correlation with *A5*.

Problem 3 (10 points) This problem is a practice of data preprocessing. Section 2.13 of online lecture Module 2 illustrates a series of data preprocessing steps. Study this section

and perform two preprocessing tasks on *automobile-cs699.arff* dataset, as described below. **Note that the screenshots shown in the online lecture module may not be exactly the same as what you will see on your screen.**

(1). Open the *automobile-cs699.arff* file on Weka explorer. Note that the dataset has missing values in some of the attributes. Run the *ReplaceMissingValues* filter on the dataset. Save the resulting dataset as *automobile-cs699-missing-values.arff*. Do not exit the explorer and continue to the next step.

(2). Make sure that the *automobile-cs699-missing-values.arff* dataset is still open in Explorer. This time, discretize the last attribute, *price*. Use the equal-width method and 3 bins. You must set *IgnoreClass* to *True*. Save the resulting dataset as *automobile-cs699-missing-values-discretize.arff*.

Submission:

The submission of this assignment consists of three files (plus some additional files if needed). The first file includes the answers to Problem 1 and Problem 2, and it should be named as *lastName_firstName*_HW3_1_2.doc (or *lastName_firstName*_HW3_1_2.pdf). The second and the third files are answers to Problem 3. If needed you may submit additional files that show how you obtained your answers.

Then, combine all files into a single archive file. Name the archive file as *lastName_firstName*_HW3.EXT. Here, "EXT" is an appropriate archive file extension (e.g., zip or rar). This is an example archive file name: *Smith_John_HW3.zip*. Upload this archive file to Blackboard.