Homework 3

Due: 9/28

- You must show all calculations and important intermediate results. Otherwise, you will lose points even if your answers are correct.
- If you use R, you must submit the R code file.

Problem 1 (10 points). This problem is about the smoothing by binning, which we discussed in the class. Consider the following variable:

<23, 26, 30, 32, 35, 47, 48, 52, 59, 63, 92, 110, 132, 147>

(1). Smooth the variable using the equal-width binning method and bin means. Use three bins.
(2). Smooth the variable using the equal-width binning method and bin boundaries. Use three bins.

Problem 2 (10 points). Use the *hw3_p2.csv* file for this problem. The dataset has 9 attributes. The first 8 attributes are predictor attributes and the last attribute, *area*, is the class attribute.

(1). Which predictor attribute has the highest correlation with the class attribute?
(2). Among the 8 predictor attributes, which pair of attributes has the highest correlation?

You may use any tool, including R, for this problem.

Problem 3 (10 points). Use the *hw3_p3.csv* file for this problem. Determine whether there is a correlation between *housing* and *class* using the chi-square test method that we discussed in the class. You may use any tool to derive a contingency table. However, you must do all calculations yourself after that, including the calculation of expected values and the test statistic. You may use a spreadsheet software or R just for the purpose of calculation.

Problem 4 (10 points). Use the *hw3_p4.csv* file for this problem. This problem is about the PCA that we discussed in the class. Use R for this problem.

(1). Standardize the first 8 attributes (predictors) using the z-score method.
(2). Split the dataset into a training dataset and a test dataset with the ratio of 66:34. Use 31 as the seed (so that I may replicate your code) and you must do stratified splitting.
(3). Apply PCA on the training dataset. If you want to keep 90% of total variability, how many principal components you should keep? If you want to keep 70% of total variability, how many principal components you should keep?
(4). Transform (or project) both the training dataset and the test dataset to new datasets with new attributes (principal components) and show the first 6 tuples of each dataset.

Include all answers in a single Word or PDF document and upload it to Blackboard. Use *LastName_FirstName_hw3.docx* or *LastName_FirstName_hw3.pdf* as the file name. If you have additional files, such as an Excel file or a R code file, then combine all of them into a single archive file and name it *LastName_FirstName_hw*3.*EXT*, where *EXT* is an appropriate archive file extension such as *zip* or *rar*.