Homework Assignment 4

Due: 10/5

- You must show all intermediate calculations/results.

Problem 1 (10 points). Consider the following dataset:

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

Classify a new object O = <A1 = Cool, A2 = East, A3 = Low> using the Naïve Bayes algorithm we discussed in the class.

problem 2 (10 points). Info gain calculation

Consider the following dataset:

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

(1). Calculate the information gain of attribute A1.
(2). Calculate the information gain of attribute A2.
(3). Which is better as the test attribute at the root level?

Problem 3 (10 points). Consider a dataset that has two predictor variables, *age* and *bp*, and the class attribute *class*. The class attribute values are *Yes* and *No*. Suppose that you ran a logistic regression algorithm and obtained the following coefficients for the class *Yes*:

intercept = -5.5461, age = 0.0418, bp = 0.0517

Calculate the probability that a new object *O* = <age = 68, bp = 145> belongs to the class *Yes* using the method we discussed in the class. You must show all intermediate steps and calculations.

Problem 4 (10 points). This question is about the discriminant analysis method that we discussed in the class. Suppose that you have a dataset with two classes, Class 1 and Class 2, and that you are trying to classify an object *O* using the method and you calculated the distance between *O* and the centroids of the two classes and obtained the following:

Squared distance to Class 1: 1.3871
Squared distance to Class 2: 3.7342

Calculate the probability that the object *O* belongs to Class 1 and the probability that the object *O* belongs to Class 2.

Problem 5 (10 points). Use the *heart_failure.csv* dataset and use R. The dataset was downloaded from the UCI Machine Learning Repository and modified for this assignment. In the dataset:

- Each tuple represents a patient.
- *DEATH_EVENT* is the class attribute; 1 means the patient died and 0 means survived.

(1). Convert the data type of the class attribute to factor.
(2). Split the dataset into training and test sets with the 66%-34% ratio. Make sure that you use a stratified splitting method.
(3). Build a Naïve Bayes model from the training dataset.
(4). Test the model on the test dataset.
(5). In your submission file, include the confusion matrix and the prediction accuracy of each class.
(6). Build a decision tree model using the *rpart* algorithm from the training dataset. Use *information gain* as the purity measure.
(7). Plot the tree of the model.
(8). Test the model on the test dataset.
(9). In your submission file, include the screenshot of the tree, confusion matrix and the prediction accuracy of each class.

Submission:

Name your file *LastName_FirstName*_HW4.doc or *LastName_FirstName*_HW4.pdf. If you have multiple files, then combine all files into a single archive file. Name the archive file as *LastName_FirstName*_HW4.EXT. Here, "EXT" is an appropriate archive file extension (e.g., zip or rar). Upload this archive file to Blackboard.