## Assignment 3

**Note: Show all your work.**

**Problem 1 (20 points)**. This problem is about the decision tree algorithm we discussed. 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 |

Calculate the information gain of A2 and A3 and determine which is better as the test attribute at the root. You must show all calculations, including the calculation of *info* and *information gain*.

**Problem 2 (20 points)**. Consider the following dataset, which is a part of the *iris* dataset:

| OID | petallength | petalwidth | class |
|---|---|---|---|
| 1 | 1.4 | 0.2 | Iris-setosa |
| 2 | 1.3 | 0.2 | Iris-setosa |
| 3 | 4.8 | 1.8 | Iris-virginica |
| 4 | 4.5 | 1.3 | Iris-versicolor |
| 5 | 4.7 | 1.6 | Iris-versicolor |
| 6 | 1.6 | 0.2 | Iris-setosa |
| 7 | 1.4 | 0.1 | Iris-setosa |
| 8 | 4.6 | 1.5 | Iris-versicolor |
| 9 | 6.7 | 2.2 | Iris-virginica |
| 10 | 6.9 | 2.3 | Iris-virginica |

Suppose you want to classify an unseen object X: <petallength = 4.2, petalwidth = 1.3> using the KNN method we discussed in the class.

(1). Calculate the distance between X and all 10 objects. Use the Euclidean distance.
(2). Classify X using five nearest neighbors.

**Problem 3 (20 points)**. This problem is about the logistic regression we discussed in the class. Consider a dataset that has two independent variables A1 and A2 and a class attribute, which takes on either *yes* or *no*. Suppose you ran a logistic regression algorithm on the dataset and obtained the following coefficients for class *yes*:

Coefficient of A1 = 0.045
Coefficient of A2 = 0.003
Intercept = -3.485

Classify the following two unseen objects using the above model:

O1: <A1 = 47, A2 = 213>
O2: <A1 = 65, A2 = 276>

Assume that the classification threshold is 0.5. You must not use any software except for calculation and you must show all calculations.

**Problem 4 (20 points).** Study *discriminant analysis* classification method, and

(1). Write a brief, one-page description of the method.
(2). Run a *linear discriminant analysis* method on Accidents1000 dataset:

- Weka: Use *Accidents1000.arff*. Choose *Percentage split* and set 66%. Submit the screenshot of output window.
- JMP Pro: use *Accidents1000.jmp*. Validation column is already created. Submit the screenshot of output window.
- Other tools: Use *Accidents1000.csv*. Split the dataset with 66-34 ratio. Submit the script file and the screenshot of your output.

Include the prediction accuracy on the test (or validation) dataset in your submission.

**Problem 5 (20 points).** Using Weka**,** run six classifier algorithms on the *echocardiogram-cs699.arff* dataset, which was downloaded from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/echocardiogram) and modified for this assignment.

For each classifier algorithm, perform:

(1) Run the classifier algorithm alone
(2) Run Bagging with the classifier
(3) Run AdaBoostM1 with the classifier

(1). Collect all accuracies (% correctly classified instances) and enter these accuracies in the following table:

| | Classifier alone | Bagging with classifier | AdaBoostM1 with classifier |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Naïve Bayes | | | |
| Logistic | | | |
| MultilayerPerceptron | | | |
| J48 | | | |
| RandomForest | | | |
| IBk (with k = 3) | | | |

(2). Include in your submission screenshots of all Weka's classification output, which includes all performance measures and a confusion matrix.

(3). Do Bagging and AdaBoostM1 increase accuracies?

**Submission:**

Submit the solutions 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 necessary, you may submit an additional file that shows how you obtained your answers. Make sure that this additional file also has your last name and first name as part of the file name. If you have multiple files, then combine them into a single archive file, name it *LastName_FirstName_hw3.EXT*, where *EXT* is an appropriate file extension (such as zip or rar), and upload it to Blackboard.