Assignment 3

Note: Show all your work as you have on previous assignments.

Problem 1 (25 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 (25 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 (25 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 (25 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 using Weka, JMP Pro, R, or Python:

- 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.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. Make sure the filename includes your name.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If a problem requires you to work out formulas or calculations, you should provide this detailed work either in an Appendix in your Word/PDF or as a separate file. All separate files should include your name in the filename.**

- **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**