CS 440 HW 6

Extra Credit Due Date: April 2, 2017 11:59 pm

Due Date: April 4, 2017 11:59 pm

1. We have a dataset of 30 students in which each student has 3 features, namely BMI, Grade, and PlayBall. BMI has 2 possible values: normal (N) and overweight (O). Grade has 3 possible values: Freshmen (F), Sophomore (S), and Junior(J). PlayBall has 2 possible values: True (+) or False (-). Implementation is not required for HW6 and no code should be handed in. However for your own use, you may find it convenient to implement all or part of the procedures required below.

	Grade	BMI	PlayBall
1	S	N	-
2	J	O	+
3	F	O	+
4	F	N	+
5	J	N	-
6	S	O	-
7	J	N	-
8	S	N	-
9	J	N	-
10	J	N	+
11	F	N	+
12	J	N	-
13	S	O	+
14	F	N	+
15	J	O	-
16	J	O	-
17	F	N	+
18	F	N	+
19	J	O	-

20	S	O	+
21	F	N	+
22	F	N	-
23	J	O	+
24	S	O	+
25	J	N	-
26	J	N	-
27	S	O	+
28	F	O	+
29	S	N	+
30	S	O	-

a) Build a Decision Tree to predict PlayBall using a training set of students 1-20. Use the learning algorithm in class (greedy information gain based on entropy). Grow the tree until homogeneity or until data runs out. Heterogeneous leafs labels are estimated probabilistically from the distribution of their training examples. Give the estimated information gain for each test in the final tree. Specify the training examples at each leaf.
b) Using your decision tree, give the estimated classifications of students 21-30.
c) Using your decision tree, give the estimated classifications of students 11-20.
d) Explain the difference between (b) and (c). Which is likely to be a better estimate of your classifier's accuracy? Why?
e) Discuss what relevant differences you would expect between the following two hypothetical scenarios:
    i. The Decision Tree is constructed using students 1-10 and evaluated against students 11-30.
    ii. The Decision Tree is constructed using students 1-25 and evaluated against students 26-30.
f) Build and describe a Decision Tree as in (a) but using all the data to predict Grade. (Now, Grade is the classification attribute instead of PlayBall.)

2. Overfitting in decision trees
    a) In your own words describe three approaches to avoid overfitting in decision trees.
    b) In your opinion, which of the approaches would work best for predicting PlayBall from the data?
    c) Describe and justify the steps you would take and say why it is likely to be better than the other alternatives.
    d) Does the algorithm always produce the shortest decision tree? Why or Why not?