Miniproject 4
BMI 555 IEE 520
Fall 2020

Due Date: November 6, 2020

1) Build a decision tree classifier on the following data **without** a decision tree software package. You might use *Excel*.

   a) Grow the tree with the data labeled *Training Data*. Use the data labelled *Validation Data* for *reduced error* pruning.
   - Model the categorical attribute *y* as the target.
   - Use entropy as the impurity measure with log base 2.
   - Use binary splits.
   - Handle *x*1 as a numerical attribute and *x*2 and *x*3 as categorical attributes.
   - Grow an unpruned tree with at most two levels (at most 4 terminal nodes), which might be pruned back based on the validation data.

   b) Construct the confusion matrix of your final model. Assume that *y* = 2 is the positive class and construct a ROC curve for the model based on the class probability estimates.

Training Data

x1	x2	x3	y
num	cat	cat	cat
70	1	4	2
67	0	3	1
57	1	2	2
64	1	4	1
74	0	2	1
65	1	4	1
56	1	3	2
59	1	4	2
60	1	4	2
63	0	4	2
59	1	4	1
53	1	4	1
44	1	3	1
61	1	1	2
57	0	4	1
71	0	4	1
46	1	4	2
53	1	4	2
64	1	1	1

Validation Data

x1	x2	x3	y
num	cat	cat	cat
40	1	1	1
67	1	4	2
48	1	2	1
43	1	4	1
47	1	4	1
54	0	2	1
48	0	3	1
46	0	4	1
51	0	3	1
58	1	3	2

2) Assume that *y* is a numerical measure (not true in general, but as a simple example here) and start to build a regression tree for the *Training Data* **without** a decision tree software package.

a) Calculate the *first split only* for a regression tree to model *y*.

b) Use your simple tree and consider the instance below. What is the predicted value of *y* for this instance?

*x*1	*x*2	*x*3	
53	1	2	

3) Use the *diabetes* dataset. Use the *AdaBoostClassifer* package in SKLEARN ONLY with only a decision tree as a base learner. Design the best boosted decision tree for me to use on similar data in the future.
Also, I want to know about the quality of your model. Provide an estimate the generalization error of the model you recommend.
Submit
- Your final code, your output, and your estimate of generalization error
- A brief description of what models you explored and how you estimated generalization error.
- A clear statement of parameters changed from the SKLEARN defaults. It is important to learn to communicate your models clearly.

You are writing for me, so you can summarize based on my knowledge of the topic.