EXAMPLES SHEET

David Wong

October 2020

These are some exercises designed to test your understanding of ROC analysis and decision trees.

These exercises are of a similar difficult to what you will encounter in the final assessment. We have chosen to make these exercises *long-form* (rather than multiple choice), to encourage you to work through calculations carefully.

If you have questions before the next session, please enter them on sli.do.

answers are provided in a separate sheet, but try to attempt the questions first before looking at the solutions

Question 1

Shape	Colour	Odour	Edible
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

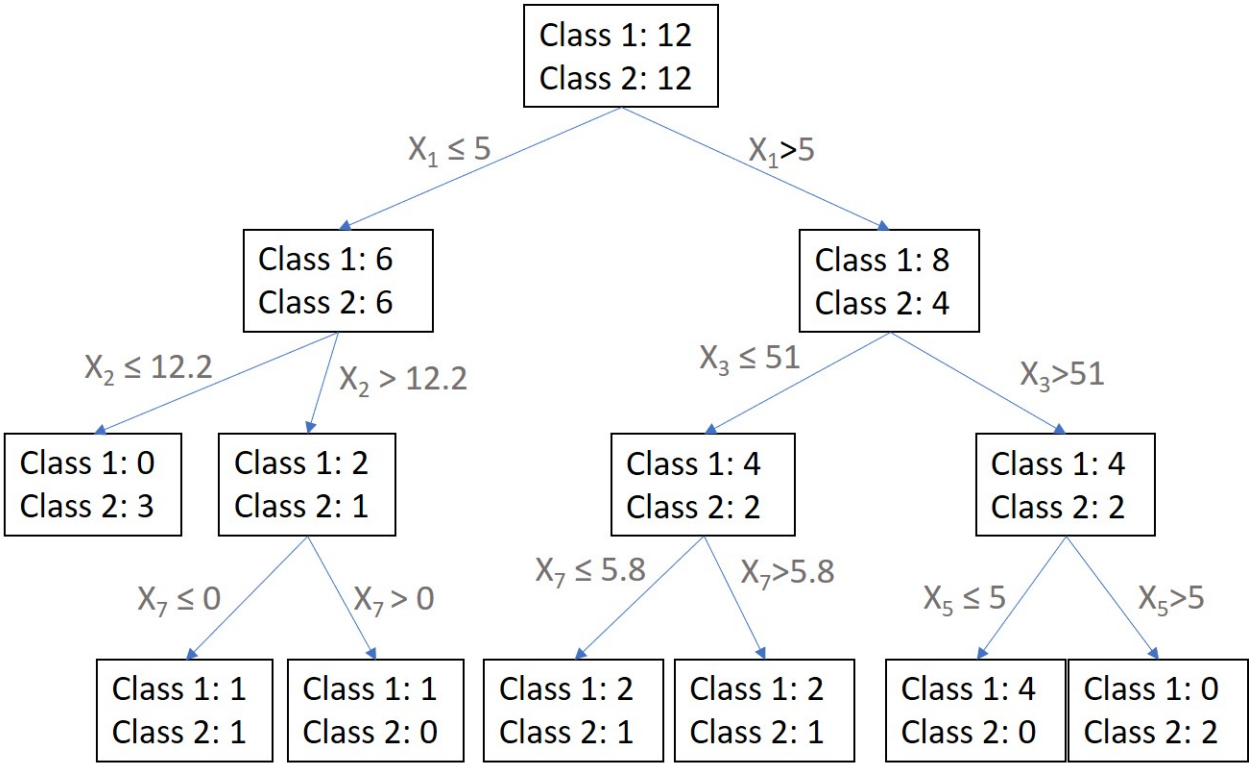
This question uses the data in the table, above. The data will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, colour and odour.

1. Describe the ID3 decision algorithm (briefly - i.e. no more than 1 paragraph)
2. What is the $Entropy(Mushrooms|Odour = 1\ OR\ Odour = 3)$?
3. Which feature would the decision tree algorithm choose for the root (first decision branch)? Justify your answer.
4. Consider a decision tree built from an arbitrary set of data. If the output is discrete and binary, what is the maximum expected training set error that any data set could possibly have (expressed as the fraction of the number of misclassified examples over the total number of training examples)? Justify your answer

1. ID3 trys to select each feature of the dataset and calculates their entropy, then calculate which feature makes it gain most information.
2. E(M|O=1 or O=3) = -(1/2log1/2+1/2*log(1/2)) = 1
3.
H(M|Odor) = 3/11*E(M|Odor=1) + 8/11 * E(M|Odor=2 Or Odor =3) = 0.59

4.

Question 2



1. Briefly, describe (post-)pruning and explain why it is necessary.
2. A validation data set is applied to a trained decision tree (of depth 4), resulting in the classifications depicted in the figure above. Which branches can be pruned without decreasing the validation accuracy?

1. Keep adding decisions until all trees is as accurate as possible. Then remove a branch, calculate validation set error and if error has decreased, remove another.
post-pruning makes it possible to gurantee the 'best' answer.

2. x7 in the middle can be pruned. zero information gain

Question 3

1. You have tested a binary classifier and found that it correctly predicts the negative class 150 times, incorrectly predicts positive when it should have predicted negative 30 times, correctly predicts the positive class 60 times, and incorrectly predicts negative when it should have predicted positive 10 times. Draw a confusion matrix for this classifier and calculate the: sensitivity, specificity, accuracy, F1 score.

	P	N
T	60	30
F	10	150

	P	N
Y	TP	FP
N	FN	TN

sensitivity = tp/P = 60/70 = 6/7
specificity = tn/N = 150/180 = 5/6
accuracy = (60+150) / 250 = 21/25
F1 score = 2/(1/sensitivity+1/precision) = 2/(1/(6/7)+1/(60/90)) = 0.75

Question 4

X1	X2	X3	Y	Model output
0.65	0.03	0.99	0	0.1
0.55	0.86	0.04	0	0.7
0.24	0.71	0.44	0	0.25
0.43	0.55	0.33	0	0.8
0.07	0.89	0.78	0	0.12
0.41	0.93	0.72	0	0.44
0.83	0.46	0.75	1	1
0.22	0.83	0.06	1	0.8
0.87	0.79	0.82	1	0.62
0.24	0.93	0.50	1	0.75
0.46	0.82	0.12	1	0.9
0.57	0.60	0.33	1	0.45

The table above shows a test set with three features $X1, X2$ and $X3$ and the corresponding binary labels, Y. A logistic regression model is used to predict Y, and the logistic model output probability is shown in the last column. Calculate the sensitivity and specificity at decision thresholds of of *model output* $= \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, and create the resulting ROC curve.

	P	N
Y	5	2
N	1	4

sensitivity = 5/6
specificity = 4/6 = 2/3

thres	TPR	FPR	
0	1	1	
0.2	1	2/3	
0.4	1	1/2	
0.6	5/6	1/3	this one
0.8	1/2	1/6	
1	1/6	0	