BA476 assignment 3 (55) Due date: 19 April at 23:55

This is an individual assignment, you may discuss it with your colleagues but do not share solutions with each other. If you get stuck on something you are always welcome to stop by office hours to talk it through with us.

Answer each question on a new page and submit a pdf on gradescope. When submitting you will be prompted to indicate on which page you answered each question. Failing to do so may result in parts of your submission not being graded.

1. Consider the following logistic regression model with $\beta_0 = 0.6, \beta_1 = 3, \beta_2 = -0.1$. A sample of the data appears below, with one binary predictor, one continuous predictor and binary outcome.

X1	X2	y
0	14	0
1	35	0
0	11	0
1	55	0
1	23	1

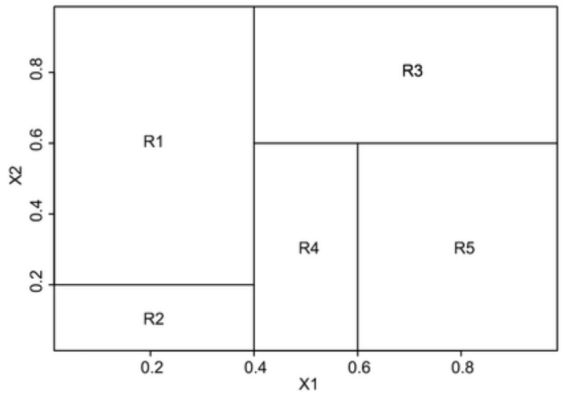
   (a) (3 points) For threshold t = 0.5, what is the accuracy of the model?

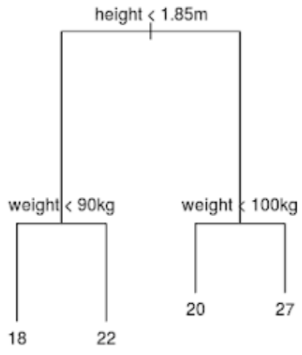
   (b) (1 point) Provide an alternative threshold that would maximize accuracy.

   (c) (1 point) Assume an instance is predicted to have p(x) = 0.7. What is the odds of observing label 1?

   (d) (2 points) True or False: According to the model having $X1 = 1$ increases your odds of having label 1 by $\beta_1 = 3$.

2. (Decision trees) Consider the feature space and decision tree below.



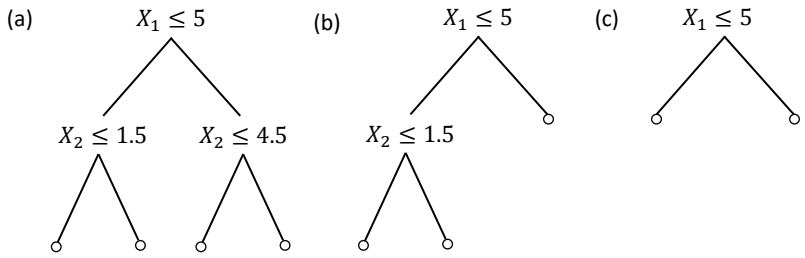


   (a) (3 points) Draw a decision tree corresponding to the partition above.

   (b) (3 points) Draw a partition of the predictor space corresponding to the decision tree above.

3. (Pruning a classification tree) Consider the dataset and decision trees given below for some classification problem. Suppose we use a loss function of $E + \alpha \cdot |T|$, where E is the number of instances misclassified and $|T|$ the number of splits in the tree.

X_1	X_2	Y
7	4	0
4	1	1
2	2	0
10	5	1
4	6	0



(a) (3 points) Which initial split makes fewest mistakes? Show that no better split exists.

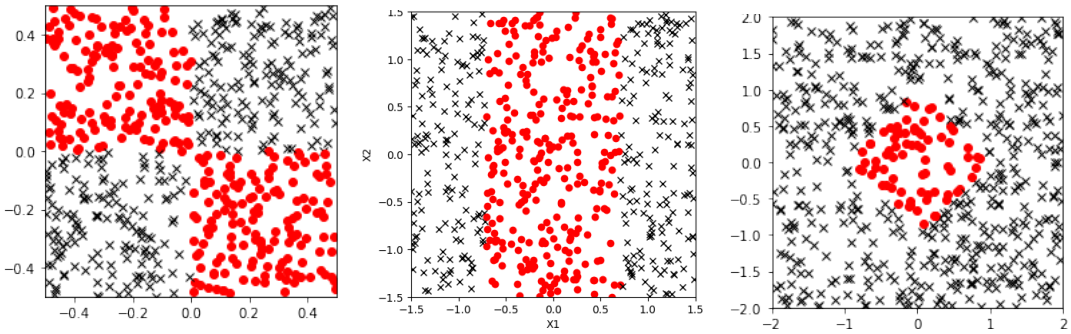
(b) (3 points) How many instances are incorrectly classified in each of the three trees above?

(c) (1 point) What is $|T|$, the number of splits in the tree, for each tree (a)-(c)?

(d) (2 points) Which tree is selected by cost-complexity pruning when $\alpha = 0.5$?

(e) (1 point) Which tree is selected by cost-complexity pruning when $\alpha = 2$?

4. (6 points) (Feature engineering) In each of the following settings we want to use a decision tree to classify the instances shown below. Come up with a new feature X_3 that we can add, based on the existing features X_1 and X_2, so that a simple decision stump which splits on X_3 can achieve perfect training accuracy.



5. (Training a random forest) The outcome variable in the dataset below is whether or not a patient is suffering from cardiac arrest, and the predictors are symptoms.

We will train a random forest by drawing bootstrapped samples of size 6. Each tree should consist of one split, sample two predictors at every split. You should use the number of instances misclassified to determine the best split, and you do not have to prune.

Random forests rely on randomness for (1) the bootstrap, and (2) selecting which predictors to consider at every split. We will use the sequences below to mimic randomness.

1) 5 1 6 5 2 3 5 4 2 4 1 1 3 5 5 4 4 1 3 2 2 4 5 3 4 5 5 3 6

2) 3 2 1 2 6 5 3 6 5 3 2 5 2 5 6 6 2 5 1 2 5 3 6 2 5 4 6 6 1

3) 2 2 4 1 1 5 1 2 1 4 6 1 2 5 3 2 1 3 4 5 4 6 2 3 3 2 6 4 4

Bp change	Tightness in chest	ECG	Age	Cardiac arrest?
Yes	No	Normal	51	False
No	Yes	Hypertrophy	61	True
No	No	Abnormal	48	False
No	Yes	Abnormal	57	True
Yes	No	Hypertrophy	79	True
No	Yes	Normal	77	False

If a number in the sequence is not applicable or doesn't make sense, skip it. For example, the first thing we do for tree 1 is draw a bootstrap sample according to the first six random numbers 516523 - these are the instances we select. Next, we have to randomly sample two predictors for the first split, so we get 5 and 4. Since we only have four predictors 5 is not valid, skip it and use the next number (2). Our sampled predictors are therefore 4 and 2 – age and tightness in chest. Now we proceed to find the best split as usual (by evaluating all possible thresholds for predictors 4, 2). If asked to form a second split we use the next two numbers (41), etc. Use sequence 2 for tree 2 and sequence 3 for tree 3.

(a) (3 points) We will now train the first tree in the forest. The example above shows how to take a bootstrap sample and sample predictors for the first split. Try all possible splits and find the best first split on the bootstrapped training set using the sampled predictors.

(b) (1 point) We now have our first tree $\hat{f}_1$, consisting of a single split. Indicate its predictions on each of the instances in the training set.

(c) (4 points) Train the second and third trees the same way. Indicate the predictions of $\hat{f}_2$ and $\hat{f}_3$ on each of the training instances.

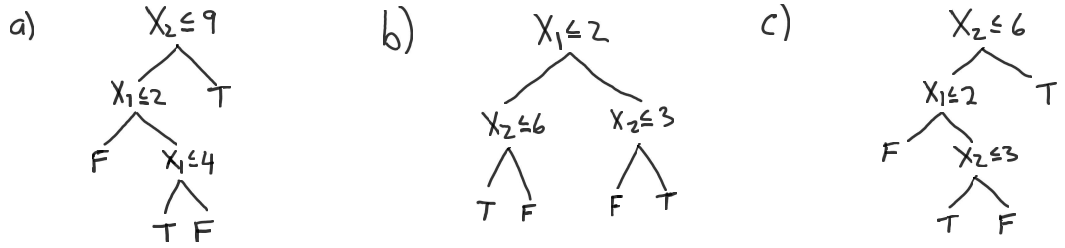
(d) (2 points) Calculate the predictions of the aggregated model on each of the training instances. Then construct a single decision tree that makes the same predictions on every instance as this random forest.

6. (Improving ensemble interpretability) After some tuning, the random forest from the previous question performs extremely well, and you want to deploy the model in a local hospital's emergency room. However, the hospital's lawyers are worried about the interpretability of the model: it is important that they can clearly explain why the model makes each prediction. Inspired by your observation in part (d), you suggest the following: train a random forest, then construct a single decision tree that makes the same predictions as the random forest everywhere in feature space. This tree can then be used to explain the random forest's predictions.

   The smallest decision tree that matches the predictions of the random forest everywhere in the feature space is called a *born-again tree*. Consider the random forest (for a binary classification task) consisting of three trees (a-c) below.

   (a) (3 points) Plot the feature space partitions corresponding to each of the trees in the forest.

   (b) (5 points) Show the random forest's predictions everywhere in the space. You may assume that the random forest makes use of majority voting (so an instance's predicted label is the label predicted most often by the trees in the forest).



Then, construct the born-again tree which matches the random forests predictions everywhere in feature space. (Hint: you should be able to find a tree with depth 3 and at most 4 splits).

7. (Fairness/bias) You are asked by ABC Healthcare to investigate if their system that recommends special program to improve care for patients with complex medical needs exhibits bias.

   The table below shows patients selected last year. You have access to 2 predictors: lifetime healthcare cost and a health risk score (higher risk scores imply a higher probability of future medical needs). The 'group' column shows membership of a protected class. The 'outcome' column represents, in hindsight, whether the patient should have been enrolled in the special program. Column 'P(x)' represents the output of a new algorithm the organization is considering for deployment. You may assume a threshold of 0.5 is used, so candidates predicted to have probability at least 50% of benefiting from the special program will be enrolled.

group	X1_healthcare_cost	X2_risk_score	outcome	P(x)
A	154966	81	0	0.58
A	129090	94	1	0.61
A	194500	84	1	0.74
A	160276	76	1	0.66
A	100011	79	0	0.48
A	125670	88	0	0.88
B	129100	94	1	0.11
B	93460	80	1	0.52
B	72200	91	0	0.67
B	74790	93	0	0.22

   (a) (3 points) Does the current algorithm (with threshold $t = 0.5$) satisfy demographic parity, equality of opportunity, or individual fairness? Support your answers.

   (b) (3 points) Can you find a threshold that does not lead to trivial decisions (accepting or rejecting everyone is seen as a trivial outcome) and satisfies at least one of demographic parity or equality of opportunity? Give the threshold and motivate why the relevant property is satisfied.

   (c) (2 points) As seen in the previous question, it can be hard to satisfy any fairness properties even when you trust that your data is unbiased. Based on the information given, discuss at least one way in which biased could have been introduced to you training process.