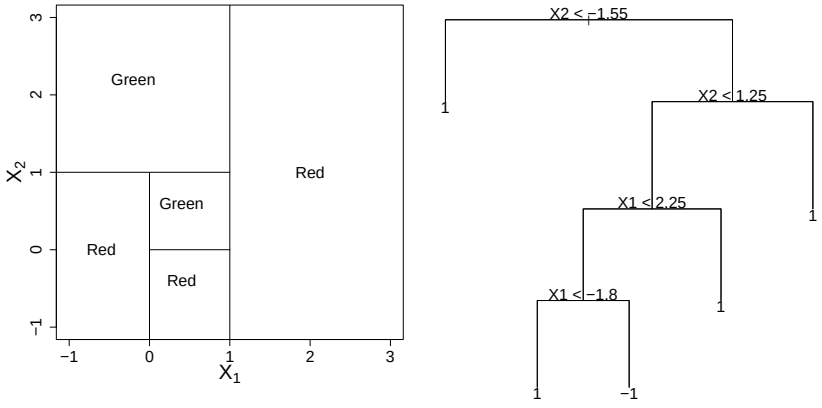
PH240C Homework 2

Due: October 19th at 7pm to your GSI

October 6, 2021

1. (20pt) This question concerns classification trees.

   (a) Sketch the tree corresponding to the partition of the predictor space shown. The class labels inside the boxes on the diagram indicate the majority of Y within each region. Label each split on your tree and indicate class label assigned to each terminal node.



   (b) Sketch a diagram similar to the one in part (a) for the tree shown below. Divide the predictor space into regions, and indicate the class label (1 or -1) for each region.

   (c) Use the tree in (b) to classify the observation $(X_1, X_2) = (2, 1)$. Trace and present the corresponding path through the tree above and write down your prediction below.

2. (35pt) This question uses the `crabs` data, available through the R package `MASS`. The data contain five size-related measurements on two different species of crabs, blue and orange, with 50 male and 50 female crabs of each species measured. Set the random seed to 6789 and randomly select 80% of the data as your training data. Make sure you select the same number of observations from each species/sex combination. Set the remaining 20% aside to use as test data.

   (a) Train a classification tree to predict Species from the five numerical measurements and sex, selecting the optimal size by cross-validation but using no more than 10 splits. Plot the tree. Comment on which variables are used by the tree. Compute training and test errors.

(b) Train random forests on the data, using $m = 5$ randomly selected predictors at each split, and 1000 trees total. Make a variable importance plot and compare with your results for a single tree. Compute training and test errors.

(c) Fit AdaBoost to the data. Plot the training and test errors as a function of the number of trees M constructed by boosting, for a range of values of M up to 1000. Report training ansd test errors for a value of M of your choice, and explain why you chose that M.

(d) Comment on which method appears to perform best for this dataset, and whether the results are consistent across methods.

3. (20 pt) Load the dataset in "dataHW2.Rda", build a classifier based on learnt supervised learning methods. Report your findings. The score you get for this question is 20 pt times the prediction accuracy of your classifier.