Data 100, Spring 2022 **April 22, 2022**

Discussion #13

Decision Trees and Random Forests

1. (a) When creating a decision tree for classification, give two reasons why we might end up having a terminal node that has more than one class.

   (b) Suppose we have a decision tree for classifying the iris data set. Suppose that one terminal decision tree node contains 22 setosas and 13 versicolors. If we're trying to make a prediction and our sequence of yes/no questions leads us to this node, what should we do?

   ○ A. predict that the class is setosa

   ○ B. give a probability of setosa = $\sigma(22/35)$

   ○ C. refuse to make a prediction

   ○ D. other (describe)

   (c) What techniques can we use to avoid overfitting decision trees?

   (d) Suppose we limit the complexity of our decision tree model by setting a maximum possible node depth d, i.e. no new nodes may be created with depth greater than d. What technique should we use to pick d?

   (e) What is the advantage of a random forest over a decision tree?

   □ A. lower bias

   □ B. lower variability

   □ C. lower bias and variability

   □ D. none of these

   (f) Let's say we have a dataset with 3 classes, labeled A, B, and C, from which we want to build a classifier using a decision tree. The following table displays how many points are in each class:

Class	Number of Points
A	4
B	4
C	2

Recall from lecture that we want to minimize the weighted entropy of our splits. What is the weighted entropy of the following split?

Node 1: 4 in class A, 0 in class B, 2 in class C; **Node 2:** 0 in class A, 4 in class B, 0 in class C.

○ A. 0

○ B. 10

○ C. $-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

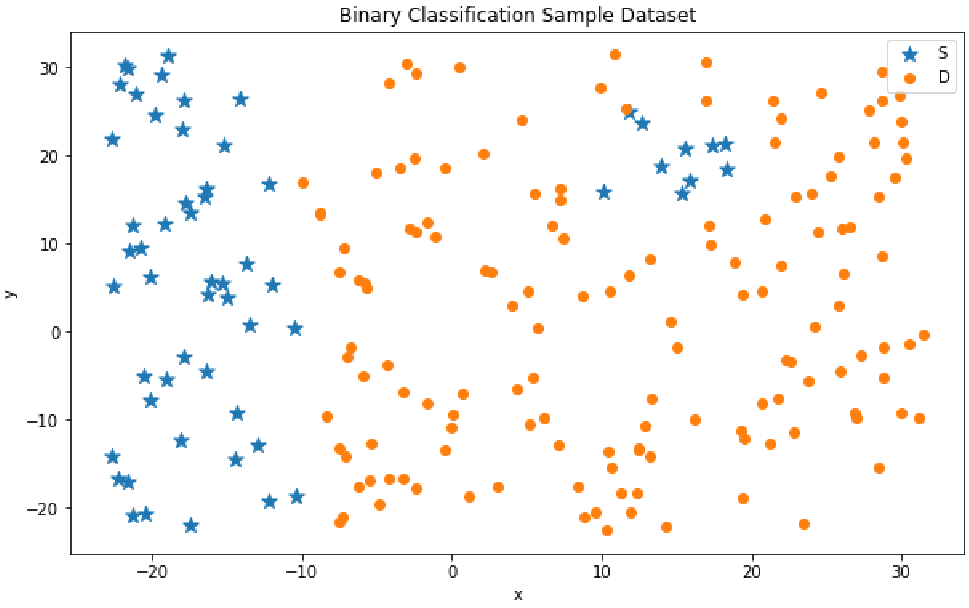
○ D. $-\frac{2}{5}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$

○ E. $-\frac{3}{5}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$

○ F. None of the Above

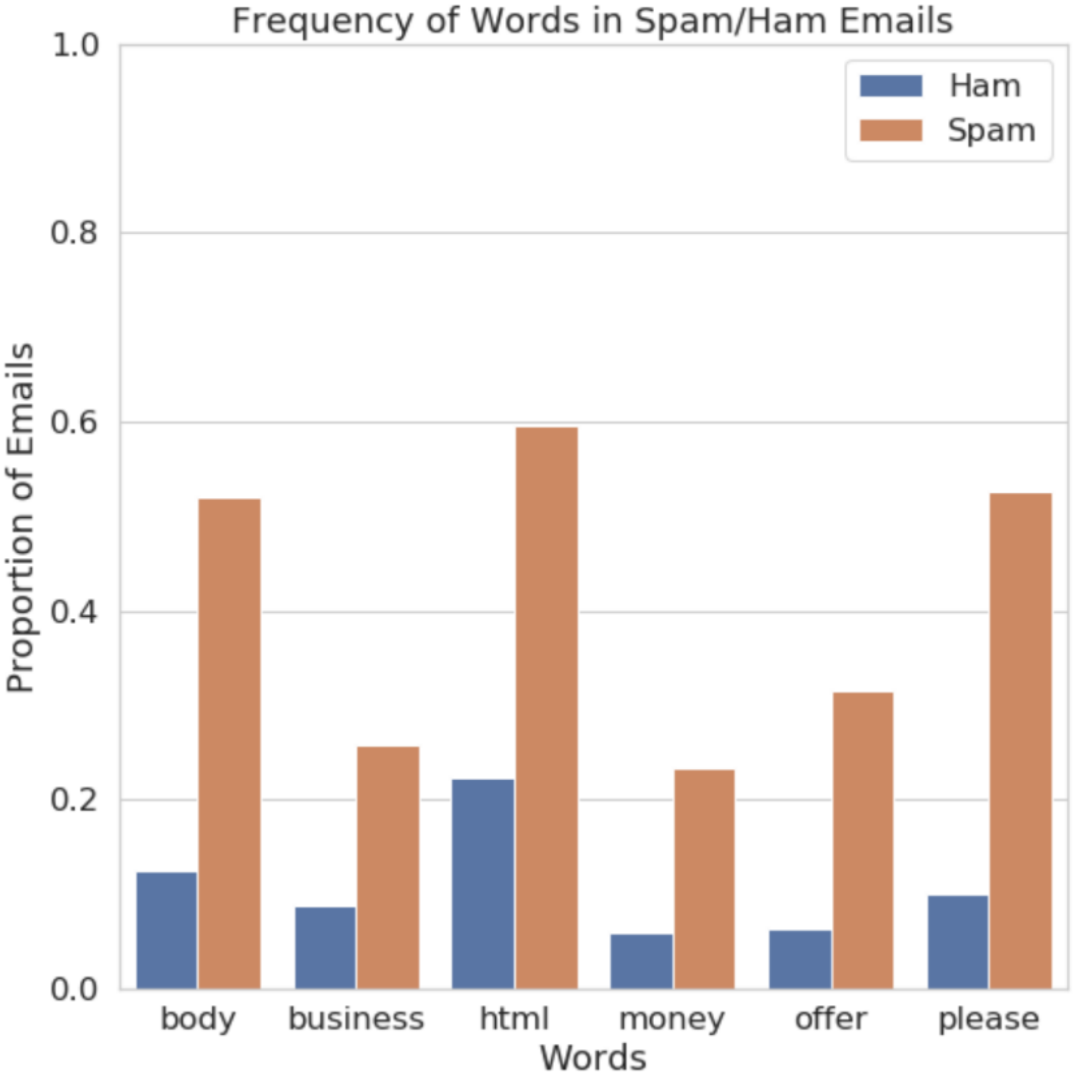
(g) Suppose we wish to train a classifier on a dataset containing classes S and D. The dataset contains two features that are plotted below (x and y), along with their respective classes denoted by stars (S) and dots (D).

Draw the approximate optimal decision boundaries chosen by a decision tree and a logistic regression model trained on the data shown by the figure below.



Spam vs. Ham (Easy/Moderate)

2. We will borrow some techniques from decision trees to build the best possible spam/ham detection classifier possible. Consider the visualization of the words that occur frequently in spam emails but infrequently in ham emails (or perhaps vice versa). These are relevant since it provides the model with word features that differentiate between the classes.



We will study this in-depth in the following parts using some of the concepts that we have learned from our study of decision trees! Assume that our spam/ham dataset contains 20,000 emails, with 10,000 spam emails and 10,000 ham emails (this isn't true - but we will pretend it is to make calculations easier).

(a) Suppose that we are building a decision tree of whether an email is spam or ham, where the decision tree can read the text in emails. Estimate the weighted node entropy of a split in a decision tree, where the left split corresponds to emails containing the word ”html” and the right split corresponds to emails not containing the word ”html”.

(b) What split word among those shown in the figure is the most effective using the same calculations as we performed in the previous subpart?

(c) In general, what kinds of words in the text would it find most useful to differentiate or decide between the two classes? Describe a procedure to select the best words for the spam/ham logistic classifier.