Proj2 Rubric

May 6, 2018

Question 1

Written Description (2 pts)

No partial credit. Full credit for any answer that points out any difference that might relate to the identification of spam.

Our answer is "'It looks like the spam email has HTML tags. If many spam emails have HTML tags, we can use them to predict whether an email is spam or ham."

Question 3

Part a (2 pts)

Subtract off $\frac{1}{2}$ **points** for any of the following reasons:

- They used the exact same set of words as in the example plot.
- They labelled the legend something other than "Spam" and "Ham" (capitalized)
- They labelled the y-axis something other than "Proportion of Emails"
- Each pair of bars (i.e. the proportion of spam and ham emails containing a certain word) is within 1%. This most likely indicates that they did not reset the index for the variable `train`, but it could also mean all the words they picked were not discriminative.

Part b (2 pts)

Subtract off $\frac{1}{2}$ **points** for any of the following reasons:

- They labelled the legend something other than "Spam" and "Ham" (capitalized)
- They did not provide a human-readable and formatted label for the x-axis.

Additionally, if they did use the "Fraction of Uppercase Letters" as their variable, subtract off 1 **point** if their y limits look significantly different from the example plot (e.g. some students had density plots up to $y = 25$ because they divided by the total number of characters instead of the total number of letters).

Question 6

Written Description (3 pts)

This question is at the very end of Question 6 right before **Part II - Moving Forward**. For the written responses, parts (d), (e), and (f) are each worth 1 point. They should say something along these lines for each question:

(d) Our logistic regression classifier does virtually no better than labelling 'ham' for every email.

(e) The words we've chosen as our features aren't actually present in many of the emails so the classifier can't use them to distinguish between ham/spam emails.

(f) Any justification for which classifier they'd prefer that accounts for their findings in this question.

Question 7 (6 pts)

2 **points** for a reasonable answer to each of the following:

- How did you find better features for your model?
- What did you try that worked / didn't work?
- What was surprising in your search for good features?

Question 8 (6 pts)

Plot (3 pts)

Subtract off 1 **point** for any of the following reasons:

- No labeled axes, title or legend. It should be clear on the plot itself what feature you are studying and which points are spam or ham.
- Not comparing spam and ham in the plot (or another comparison that's helpful for feature selection).
- Plot doesn't show compelling difference, plot doesn't show the feature, or any plotting problems such as overplotting that make it challenging to appreciate the plot.

Commentary (3 pts)

No points if there is no description of findings and implications or if their findings are completely wrong. Subtract off **1 point** if their description of findings and implications was only partially correct.

Question 9 (3 pts)

Subtract off 3 **points** if the plot was not created or if the shape was incorrect.