Notebook

April 26, 2020

0.0.1 Question 6c

Provide brief explanations of the results from 6a and 6b. Explain why the number of false positives, number of false negatives, accuracy, and recall all turned out the way they did.

From 6a:

The zero predictor is always going to be 0 and always will not be a positive but will always be a false positive of 0. To continue with the zero predictor, this will never predict a true positive and will only predict a false negative which is the total number of spam emails.

From 6b:

In order to get the accuracy we need to divide the total number of ham emails by the total number of emails and zero predictor is 0 because there is no way that a positive be ever predicted.

False Positive: we want to know the number of ham emails that are flagged as spam and filtered out of the inbox, FN, accuracy, recall

False Negative: we want to know the number of spam emails that are mislabeled as hams and ends up in the inbox.

accuracy: we want to know how many hams are labeled correctly out of the total number of emails

recall: want to know how many spams are labeled as spam, since it's zero predictor, so no spam is labeled.

0.0.2 Question 6e

Are there more false positives or false negatives when using the logistic regression classifier from Part A?

```
In [34]: FP,FN
```

```
Out[34]: (122, 1699)
```

0.0.3 Question 6f

1. Our logistic regression classifier got 75.8% prediction accuracy (number of correct predictions / total). How does this compare with predicting 0 for every email?
2. Given the word features we gave you above, name one reason this classifier is performing poorly. Hint: Think about how prevalent these words are in the email set.
3. Which of these two classifiers would you prefer for a spam filter and why? Describe your reasoning and relate it to at least one of the evaluation metrics you have computed so far.

1. By predicting 0 for every email our yields 74.47%, which is less than 75.8%, therefore our logistic regression classifier is better.

2. The words are common in both spam and ham emails, such as the words memo and private for example.

3. I prefer the logistic regression classifier for a spam filter, since it has higher prediction accuracy.

0.0.4 Question 7: Feature/Model Selection Process

In the following cell, describe the process of improving your model. You should use at least 2-3 sentences each to address the follow questions:

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

1. The way I found my better features for my model was to find all of the suggested feature recommendation which involved me creating functions for each of the suggested feature recommendation
2. During this process I tried to include more functions that didn't seem to work or were too complicated to code, so I stayed with the function that worked. Another thing I tried was I looked through some of my spam emails to see which words were common in order to get a better grasp as to what words appeared frequently with in those emails
3. In my search through my spam emails I found that the words that appeared frequently within my emails also appeared frequently within the dataset as well. This was pretty surprising since it seems that these are pretty common words that appear in a lot spam emails.

Generate your visualization in the cell below and provide your description in a comment.

In [37]:
```
# Write your description (2-3 sentences) as a comment here:
# There is a alot of frequency between body and html occuring at the same time considering the
# both main types of html tags. In the ham section we ca see that there is some variantion aft
# also see that within these categories, that spam has more content (aka body) compared to ham
# that there is alot of non-sense within these emails but this could also be because there are
# lead to longer bodys/content. Lastly ham seems to end around the mid 60 mark which would mea
# then ham

# Write the code to generate your visualization here:
hams = train[train['spam'] == 0]
spams = train[train['spam'] == 1]

def TagCt(df, name):
    return df['email'].str.findall(name).str.len()
ax = sns.regplot(x=TagCt(spams,'html'),y=TagCt(spams,'body'),x_jitter=0.5, y_jitter=0.5, label
ax = sns.regplot(x=TagCt(hams,'html'),y=TagCt(hams,'body'),x_jitter=0.5, y_jitter=0.5, label =
ax.legend()
plt.title("Spam VS. Ham")
plt.xlabel('html')
plt.ylabel('body')
ax.set_xlim(0, 80)
ax.set_ylim(0, 80)
# Note: if your plot doesn't appear in the PDF, you should try uncommenting the following line
# plt.show()
```

Out[37]: (0, 80)