kath.eddinger@berkeley.edu ¶

In [8]:

```
output:
```

The spam email contains some html text formatting (ham does not). Spam email has an http link with IP address, while the ham links have website names with format 'blah.something'. The spam email uses language that is more informal and direct, using personal pronouns ("you") and includes statements about preference ("would you rather...", "its up to you"). It is not informative, and it is implied that recipient must take some kind of action ("come in here and see how") inorder to recieve the actual information. The ham email is much more formal and does not use personal pronouns, indicating that the email was not meant for any one person specifically, but for a group of people. It is a generic announcement, meant to be informative. The recipients of this email do not need to be told that they should read the links, because they already know what the content will be (i.e. more information about the bankrupcy and auctioning of a&l daily). It also includes a sign off ("thanks, misha!").

1. *The sensitivity (proportion of spam emails that are classified as spam) will be 0%, the specificity (proportion of ham emails classified as ham) will be 100%.*
2. *If we predict ham for every email, the accuracy on the training set would be 74% (the percentage of ham emails).*
3. *Our logisitic regression classifier is only predicting about 1% better than prediciting ham for every email, so its isn't very good.*
4. *The sensitivity of the logistic regression classifier is 63.5%, the specificity is 76.1%. The classifier is more likely to make false negatives (marking spam as ham).*
5. *The poor performance of our classifier tells us that our current words probably do not appear in many of the spam emails, or they appear in equally many spam emails as they do ham emails.*

1. *I came up with ideas based on the suggestions and what I saw in the first few rows of train. Then I created some features. I tried out each feature on its own to see how well it predicted. If it didn't do very well, I decided not to use that feature. I plotted my feature grouped by its spam or ham category to see if my feature was making any meaningful divisions. Another thing I did was look at model weights (for appropriately scaled features like words). I removed some of the words with zero weights.*
2. *I tried counting the number of uppercase words in the subject, but then I decided that the percentage of uppercase words might be more appropriate. The percentages worked better. I also tried a small number of words that I thought of myself, which wasn't very good and then I found about 460 or so spam words from the internet, which worked quite well.*
3. *When I looked at the weights for my word model alone, I was surprised that some of the words that I thought would be very indicative of spam had low or zero weights. I realized later that this could be due in part to other words or phrases that tend to co-occur in spam emails.*

In [33]:

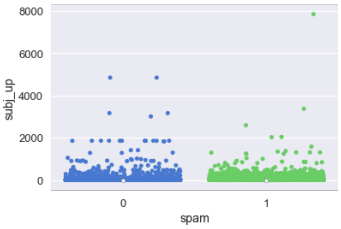
```
output:
```

(note, I accidentally deleted one of the blue cells and could not get the highlighter to re-highlight my first plot)1. This strip-plot shows how many uppercase words are in each email. This plot was useful to me because it shows that this feature does not do much to differentiate between spam and ham, which would be harder to determine using a bar plot.

(2). This violin plot was useful because it shows the difference between distributions of spam and ham based on percentage of uppercase words in email subjects. The spam plot is more evenly distributed, while the ham plot has a wide and concentrated base, with a small secondary distribution that reaches up to about 40%. The small cluster of outliers in spam, which tells me that this feature is doing something to differentiate spam from ham.

Out[33]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x113d39128>
```



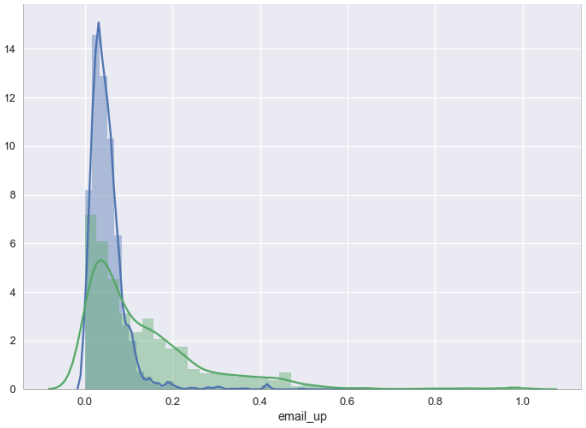
In [34]:

```
output:
```

This distribution plot was helpful to tell whether or not the percentage of uppercase in emails is doing anything to seperate spam from ham. You can see that ham (blue)has a high peak and small right tail which means that ham tends to cluster in a predicable way. Spam (green) has a wider distribution, showing that it is seperable from ham using this classifier.

Out[34]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x113fa2c50>
```



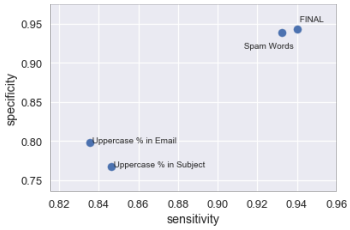
In [35]:

```
output:
```

This plot was useful in being able to tell the degree to which each classifier was affecting sensitivity versus specificity. This plot revealed an interesting different in the uppercase in email versusuppercase in subject classifiers as they achieve similar prediction accuracy, but seem to be explaining different parts of the variance.

Out[35]:

```
<matplotlib.text.Text at 0x113cca470>
```



In [36]:

```
output:
```

Out[36]:

```
<matplotlib.text.Text at 0x1134e2400>
```