DATA 221
Homework 2 (rev 1)

W. Trimble

Due: 11:59pm Thursday 2023-01-19

1. **Naive Bayesian Spam Classifier**

   Using the Kaggle ”SMS Spam Collection Dataset,” a collection of 5000 text messages, 13% of which are labeled as spam, count the word usage for the spam messages and the word usages for the ham messages.

   `https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset/`

   Construct a function that scores new text messages by estimating $\frac{P(spam)}{P(ham)}$. Find the empirical word frequencies in the dataset $P(word|ham)$ and $P(word|spam)$ for words that occur more than five times total in the dataset. (This is somewhat ill-posed, since we have to assign a number to this probability ratio (even if it is 1) even for words that occur 0 times in the training data.)

   As an ad-hoc data regularization approach, let us cap the maximum absolute value of the score that we will give to any word at 20; an utterance of 4 words that only appear in the spam corpus will get a score of 160000:1, and an utterance of 2 words that appear only in the ham corpus would get 1:400.

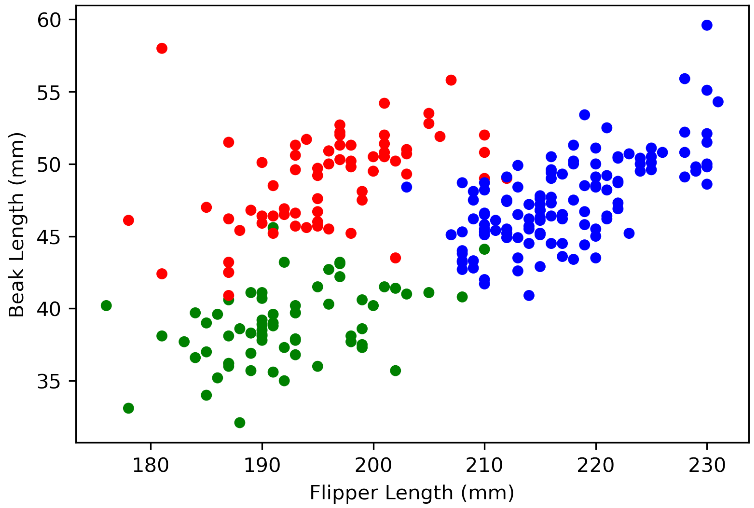
   (a) Use this to score all of the messages in the corpus described in Daniel R. O’Day and Ricardo A. Calix. ”Text Message Corpus: Applying Natural Language Processing to Mobile Device Forensics,” Proceedings of the Socio-Mobile Media Computing Workshop at IEEE International Conference on Multimedia and Expo, San Jose, USA. July 15 - 19, 2013, which is distributed at

   `https://web.archive.org/web/20130921042957/`

   `http://cybersecurity.cit.purduecal.edu/content/dl/master_corpus.txt`

   (b) Plot the histogram of log-odds-scores for all the messages in the training set (disaggregated by Spam /not spam) and for all the log-odds scores for the messages in the American text-message corpus disaggregated by Incoming / Outgoing.

   Example code for tokenization will be provided. This is a version of the loaded dice problem.



2. **Linear Regression, Logistic Regression, and Linear Discriminant Analysis** The Palmer penguins dataset has been split into a training set and a testing set. (n=265 and n=97). This dataset has two ”label” variables (sex and species) and four numerical ”features.” We will try to classify penguins by species using three techniques that look at linear combinations of the feature vectors.

   (a) Find linear regression coefficents for the indicator variables for species identity against the four-dimensional X. Plot the decision boundaries between the classes implied by the regression coefficients on top of the scatter plot.

   (b) Find logistic regression coefficents for the indicator variables for species identity against the four-dimensional X. Plot the decision boundaries between the classes implied by the regression coefficients on top of the scatter plot.

   (c) Find the class-conditional multivariate normal densities in four dimensions for each of the three penguin species using the training subset. Plot the (quadratic) decision boundaries between penguin species on the scatter plot of flipper length vs. beak depth.

   (d) Classify the test set by species and report the confusion matrix for one of the three classification methods above.

   Here you can either plot the boundaries by finding the equations for the boundary or, if you find it easier, evaluate a classifier at a few hundred points on a 2d grid and plot a symbol on the graph indicating which regions of X get which classification; you can solve this with math or you can solve it numerically.