DATA 221
Homework 2 (rev 3)

W. Trimble

Due: Friday 2022-04-15

1. **Bivariate normal properties** Given a two-dimensional multivariate Gaussian distribution centered on (0,0) :

$$\mathcal{N}(\mathbf{x};\mathbf{\Sigma}) = \frac{1}{(2\pi)}\frac{1}{(\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2)}\exp -\frac{1}{2}(\begin{bmatrix} x_0, & x_1 \end{bmatrix}\begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 \end{bmatrix}^{-1}\begin{bmatrix} x_0 \\ x_1 \end{bmatrix})$$

There are two linear combinations of x_1 and x_2 that maximize (minimize) the variance of the sums and at the same time make the two sums independent of each other.

We can paramaeterize all possible sums of f x_1 and x_2 with θ :

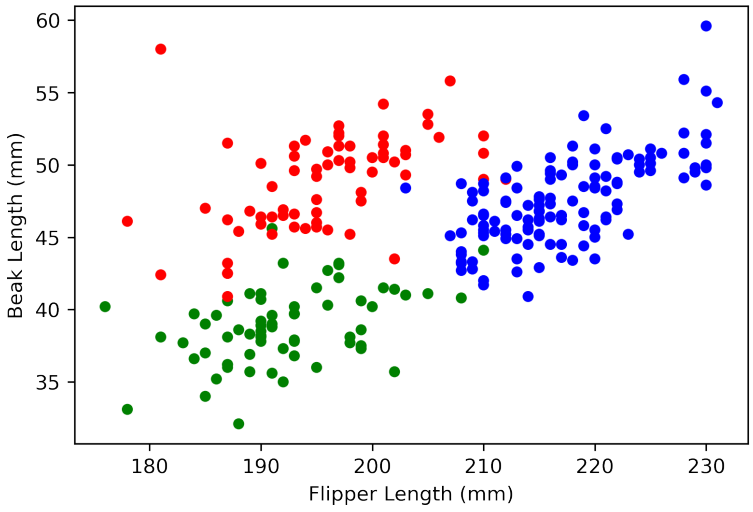
$e_1 = \begin{pmatrix} cos\theta \\ sin\theta \end{pmatrix}$ and $e_2 = \begin{pmatrix} -sin\theta \\ cos\theta \end{pmatrix}$

$$z_1 = e_1 \cdot x$$

and

$$z_2 = e_2 \cdot x$$

Find the value of θ which makes the covariance between z_1 and z_2 vanish. Find the standard deviations of z_1 and z_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.

2. 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.

3. 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.

4. 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.

5. 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, evalutate 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.

   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 $\frac{P(spam)}{P(ham)}$ using the ratios of the probabilities of words that occur more than five times total in the dataset. (This is somewhat ill-posed, since we have to assign a probability ratio (even if it is 1) to words that occur 0 times in ham and to words that occur 0 times in spam.)

   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.

6. Score the following messages: `"Hey there, I am Maya with GP Research. We're surveying IL residents. Can you respond to a few questions?"`
   `"Headed down now."`
   `"The banana chocolate bread is delicious! All that's left is one small heel, which I will dunk in my coffee tomorrow."`
   `"Hurry! For a limited time, add a FREE line to your account. Really, it's on us{no strings attached."`

7. Plot the histogram of log-odds-scores for all the messages in the training set.

   Example code for tokenization will be provided. This is a version of the loaded dice problem with natural-language-type data.