Homework 1: Data

Matlab comes with a lot of data, and we'll look at one of them today- This is a set of data used a lot since its introduction by a very famous statistician, Sir Ronald Fisher, who has been called the single most important figure in 20th century statistics.

When you save your workspace in Matlab, you'll find that the software uses a suffix of .mat. For example, if you type: `save Workspace`, you'll see `Workspace.mat` in your directory. To load the workspace back in, in the command window, you would type: `load Workspace.mat`.

In the case of Fisher's iris data, we type in `load fisheriris.mat`, or simply `load fisheriris`. When we do that, we should see that two data sets have been loaded. The first, *meas* is a 150×4 matrix, representing 150 different flowers, where each flower had four measurements taken. In particular, they are:

- Sepal length (in cm)
- Sepal width (in cm)
- Petal length (in cm)
- Petal width (in cm)

Also, we see *species*, which is a cell structure holding the names of each flower type- there are 3 possibilities: setosa, versicolor, or virginica. Google those to see what they look like!

Our goal in this lab is to explore the data statistically. As a hint to work with these exercises, the first 50 samples are setosa, the next 50 are versicolor, and the last 50 are virginica.

Feel free to look over the help files for each command! For example, for the `hist` function, try typing: `doc hist`. You'll see an explanation plus some examples.

1. Data Exploration

   (a) Use the `size` command to have Matlab tell you the how many flowers, and how many features there are in the data.

   (b) For each feature, plot a histogram of the data values using Matlab's `hist` function.

   (c) Compute the mean and variance of the 4 features using Matlab's `mean` and `var` functions.

   (d) "Normalize" the data by subtracting the mean from each feature, and then dividing by the standard deviation (which is the square root of the variance). Show your code, and try not to make a loop.

   (e) For each pair of features, $(1, 2)$, $(1, 3)$ and $(1, 4)$, plot a scatterplot of the of data colored by which flower it is. Try to use different symbols for each flower as well.

Side Note: Sometimes it is useful to explore the data "live" rather than use the command line. For example, you can select data from the **Workspace**, *select a row or column, then select the* **Plots** *tab at the top of the menu. From there, you can interactively select different plots...*

2. Data Classification using Built-in App

   (a) First, we need to prepare the data a bit. Feel free to clear out the old version of the data- We'll need to put it in a different format. We'll talk about what the data types mean a little later, but for now, use the following command, after which you will see a data structure in the workspace that is 150×5.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

   (b) Press the `APPS` tab at the top of the Matlab window, and we'll open up the "Classification Learner".

   (c) See if you can load the data in- We want to use the four measurements as "Predictor", and the class labels as "Response".

   (d) We want to train a `Linear Discriminant` on the data, with **no validation.**

   (e) Once the data has trained, take a look at and record the **confusion matrix**.

   (f) Save the model as `Fisher01`.