Topic 5: Classification

Problems 1 through 3 involve the code for the Adult Height example.

1. In the portion of the code denoted "Simulate the Data," the command normrnd is used to simulate the data. The first input to this command describes the average location of data generated from it and can be any number, positive or negative. The second input describes the amount of variability found in the data and is called the **standard deviation**. It can only take positive values.

   Let's explore how these two inputs impact the distance matrix and classification results by changing the inputs for the height of the women.

   a. Keeping the standard deviation fixed, change the value for the average. Try various values, including numbers less than 64 and those closer to 70. Does changing this have any effect on the classification results? If so, what values improve the results and which make them worse?
   b. Repeat part a, but this time keep the average fixed and change the standard deviation of the data.
   c. Based on your above answers, what combination of average and standard deviation do you think would lead to the best level of performance for classification (ex: low-low, low-high, the original values, etc.) ? The worst?
   d. Now, change both inputs together. How do the classification results compare with your intuition in part c?

2. Let's now look at the command that says "if rand<.5". The number in this command describes the proportion of the data set that represents heights of women. The default value means that roughly half of the sample will have a label of 1. This number can take values between 0 and 1.

   Using the default inputs of 64 and 2 for the average and standard deviation for the women, try changing the proportion of observations that represent women's heights. What effect does the proportion of women have on the classification results?

3. The value *n* represents the number of observations. Let's explore the effect of changing *n* to both small numbers (i.e. 5, 10, 20, 30, 50) and large numbers (i.e. 300, 500, 1000). Repeat the classification procedure a few times for each *n*. What effect, if any, does *n* have on the classification results?

Problems 4 and 5 involve the 2 by 3 black and white images we looked at earlier.

4. Alter the data as follows:

```
A{1}=[1 1 1; 1 1 1; 0 0 0];
A{2}=[1 1 1; 1 1 0; 0 1 1];
A{3}=[1 1 0; 1 1 0; 1 1 0];
A{4}=[1 1 0; 1 0 0; 0 0 1];
A{5}=[1 0 1; 0 1 0; 1 0 1];
A{6}=[0 1 0; 1 0 1; 0 1 0];
A{7}=[1 1 1; 1 0 1; 0 1 0];
A{8}=[0 0 0; 1 0 1; 1 1 1];
A{9}=[0 0 0; 0 0 0; 1 1 1];
A{10}=[0 0 1; 0 0 0; 0 0 0];
A{11}=[0 0 0; 0 0 1; 0 1 1];
A{12}=[0 0 0; 1 0 0; 1 1 0];
```

   What effect does changing the data have on the classification results?

5. Let's explore at the effect of changing the labels on the observations. For each part, let's look at the change in the classification results compared to the original.

   In all parts, the cutoff variables will need to be adjusted in some manner for the average distance classifier. The conditional statements for calculating "avgdistE" and "avgdistM" may also need to be adjusted.

   Alter the labels in the following ways:

   a. `lab=[1 1 1 1 1 2 2 2 2 3 3 3];`
   b. `lab=[1 1 1 1 2 2 2 3 3 3 3 3];`
   c. `lab=[1 1 1 2 2 2 2 2 3 3 3 3];`
   d. `lab=[1 1 1 1 1 1 2 2 2 2 2 2];`
   e. `lab=[1 1 1 2 2 2 3 3 3 4 4 4];`

6. We will refer again to the code with a header of "2 by 3 image Example". This time, though, we will need to alter the code further to accommodate a different data set.

   a. Use the following code to read in the new data set:

```
A{1}=imread('office_1.jpg');
A{2}=imread('office_2.jpg');
A{3}=imread('office_3.jpg');
A{4}=imread('office_4.jpg');
A{5}=imread('office_5.jpg');
A{6}=imread('office_6.jpg');
```

      The first 3 observations belong to group 1 and the rest to group 2.

   b. Alter the remainder of the code accordingly to perform both classification procedures for both types of distances.

   c. Which classification procedure worked the best? Which one performed the worst? Did any perform worse than random classification would?