Statistical Data Mining Example 1

1. Data were generated from a mixture of Gaussian described on page 16-17 and recorded in the files ‘trainred.txt’, ‘traingreen.txt’, ‘testred.txt’, and ‘testgreen.txt’. Load the data into your statistical software R. Each file has two columns representing the first and second coordinates, respectively, of the points.

2. Plot the training data, using red to indicate the red points and green to indicate the green points.

3. Use linear regression to fit these data as described on page 13 and find the training error rate for this method.

4. Create a plot similar to Figure 2.1 on page 13.

5. Use the linear regression fit obtained from the training data to predict the color at each input value in the test set (without using the output values in the test set).

6. Use the k-nearest neighbor algorithm with $k = 1, 3, 7$ to predict the color at each input value in the training set, and find the training error rate for each k.

7. Use the k-nearest neighbor algorithm fits obtained from the training data to predict the color at each input value in the test set (without using the output values in the test set).