Statistics 5525: Homework 1

For each homework assignment, turn in at the beginning of class on the indicated due date. Late assignments will only be accepted with special permission. Write each problem up *very* neatly (LaTeX is preferred). Show all of your work.

Problem 1

From chapter 2 in TESL, the 2-dimensional dataset is created using the following procedure:

1. sample $m_i \sim N([1,0]', I)$ for $i = 1, \ldots, 5$.

2. sample $m_i \sim N([0,1]', I)$ for $i = 6, \ldots, 10$.

3. sample m from $(m_1, \ldots, m_{10})$,
   each with probability $1/10$ and sample $x_j \sim N(m, I/5)$.

4. if $m \in \{m_1, \ldots, m_5\}$, let $x_j \in \{\text{class } 1\}$ (otherwise $x_j \in \{\text{class } 2\}$).
   (Note: save the subclass information (subclasses $1, \ldots, 10$). You will use this later.)

5. repeat steps 3 and 4 for $j = 1, \ldots, 100$.

Part a

Simulate from the above procedure (Gaussian mixture model), and plot the data. Use 2 separate colors to denote the two classes.

Part b

Use the least squares method to classify the data. Show a plot denoting your linear separating boundary. State both false positive and false negative rates.

Problem 2

Given knowledge of all the parameters in the model $(\{m_1, \ldots m_{10}\})$, the mixture weights ($1/10$ for each subclass), and the covariance functions, derive the optimal separating boundary.

Problem 3

Show a plot denoting this boundary. (Note: This may look a little different than what is in the book since your $\{m_1, \ldots m_{10}\}$ are different from what they've simulated). State both false positive and false negative rates.

Problem 4

If you aren't given knowledge of the m_is, but are given the subclass labels, show how to construct the separating boundary. (Note: many methods exist, and some are better than others. You may use any method you deem reasonable). Show a plot of your results, and state both false positive and false negative rates.

Problem 5

If you weren't provided the $\{m_1, \ldots m_{10}\}$ and subclass labels, but were given the super class labels (class 1,2), what would you do? You don't have to derive anything, or code anything up. Simply describe in plain english.