CSI772/STAT772
Spring 2013
Gentle

Exercise 1.A

Consider a mixture of two bivariate normal distributions, $\mathrm{N}_2(\mu_1, \Sigma)$ and $\mathrm{N}_2(\mu_2, \Sigma)$, where $\mu_1 = (0, 0)$, $\mu_2 = (1, 2)$, and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$. This is a data-generating process that yields (G, X_1, X_2), where G is a group indicator taking the value 1 or 2, and $(X_1, X_2) \sim \mathrm{N}_2(\mu_i, \Sigma)$, where $i = G$.

1. Suppose the mixture is half and half; that is, the probability that any given observation is from the first distribution is 0.5.

   Determine the boundary for a Bayesian classifier, and use R to plot that boundary on a single set of labeled bivariate axes in which the horizontal axis goes from -3 to 5, and the vertical axis goes from -3 to 7. (Your plot does not contain any observations.)

2. Now, suppose that the probability that any given observation is from the first distribution is 0.25. Repeat the previous problem.

3. Under the assumption that an observation has a probability of 0.25 as in the previous problem, use R to generate a random sample of 200 observations from this mixture. (Note that the actual number in the first group may not be exactly 50.)

   For a response variable taking the values 1 and 2, fit a linear regression model of the form

   $$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

   and use the classifier $\widehat{G} = 1$ if $\hat{y} < 1.5$ and $\widehat{G} = 2$ otherwise.

   On a single set of axes, plot your observations color-coded to indicate the group they are in, plot the Bayes classifier boundary (from the previous problem), and plot your linear regression classifier using a different line type. Put a legend on your graph to distinguish the two classifiers (not the observations).