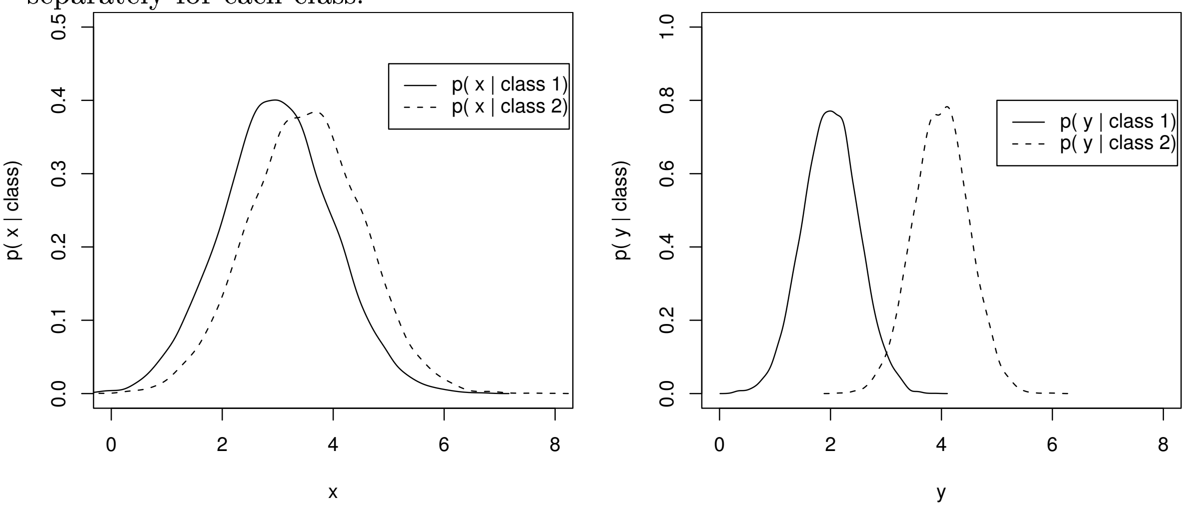
36-350: Data Mining

Homework 3
Date: September 8, 2002 **Due: start of class September 15, 2002**

1. Consider two colors x and y. Each image has a count for x and a count for y, which vary between images. The distribution of the count for x and for y is depicted below, separately for each class:



   Which color is more informative, and why?

2. Suppose that a particular dimension has the same average value in each group. It is still possible for the dimension to be informative about the group. Draw an example where this happens.

3. Consider the following subtable of counts for “suddenly”:

```
> subtable(xp,"auto","suddenly")
          suddenly not suddenly
auto             0          611
not auto         2          694
```

   (a) What is the probability that a random document is “auto”? What is the entropy of $c \in \{\text{“auto”}, \text{not “auto”}\}$?

   (b) What is the probability that a random word drawn from the collection is “suddenly”?

   (c) What is the entropy of c if the word turns out to be “suddenly”?

   (d) What is the expected information in testing for “suddenly”?

4. In lab, you found a word where the expected information and actual information disagreed about the words’ relevance. Based on its table of counts, explain why the measures disagree.

5. In lab, you found a word where the smoothed actual information and raw actual information disagreed about the word's relevance. Based on its table of counts, explain why the measures disagree.