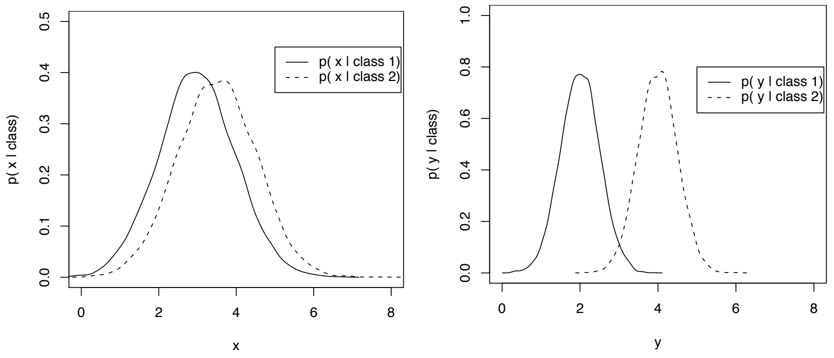
Homework Assignment 3

36-350, Data Mining

Due at the start of class, 3 October 2008

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 of two histograms where this happens.

3. Consider the following table of counts for "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 the class C (which can be "auto" or "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. A dataset is currently divided into three clusters. The first cluster has 100 measurements and mean 0. The second cluster has 5 measurements and mean 2. The third cluster has 3 measurements and mean 5. We want to merge two clusters, while minimizing the sum of squares. Which two should we merge?

5. A web search engine keeps track of the popularity of each web site it indexes. Popularity is measured by the number of hits per day. The organizers want to improve the result of a search by displaying "popular" results separately from "unpopular" results, in two different columns. The problem is that the definition of "popular" versus "unpopular" has to be defined relative to the set of results, e.g. 100 hits per day might be "popular" when most of the other results have 10 hits per day, but would be considered "unpopular" when the other results have 1000 hits per day. They need a solution to this which is automatic—no human intervention. Describe how the organizers can do this using the methods in class.

6. A data miner runs k-means to divide a dataset into 3 clusters. The cluster means turn out to be evenly spaced across the range of the data. Can the miner conclude that the data consists of three separate subgroups? Explain.

7. Below is the PCA projection matrix for the car data shown in class. All of the variables were standardized to have zero mean and unit variance.

```
               h1    h2
Price        0.29  0.43
MPG.highway -0.30 -0.05
EngineSize   0.35  0.06
Horsepower   0.30  0.49
Passengers   0.21 -0.68
Length       0.33 -0.10
Wheelbase    0.33 -0.26
Width        0.34 -0.07
Turn.circle  0.32 -0.08
Weight       0.37  0.01
```

   (a) If a car is one standard deviation above average in `Price`, one standard deviation below average in `MPG.highway`, and average on all other variables, what are its coordinates in the `(h1,h2)` projection?

   (b) If a car is one standard deviation above average in `Passengers`, one standard deviation above average in `Weight`, and average on all other variables, what are its coordinates in the `(h1,h2)` projection?