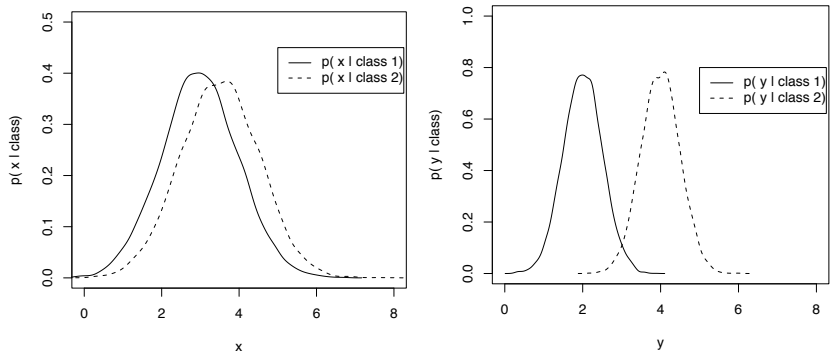
Homework Assignment 3

36-350, Data Mining

Due at the start of class, 18 September 2009

IMPORTANT: The last two problems are much easier if you use the code accompanying lectures 5 and 6.

1. Consider classifying images using their bag-of-colors representations. Each image has a count for each color. There are two classes. Our two favorite colors are x and y; the figures below show the distribution of pixel-counts for x (on the left) and y (on the right), with solid or dashed lines indicating the different classes.



   Which color gives us more information about the image's class? Why?

2. Explain how a feature can provide information that lets us discriminate between classes, even though it has the same average value in each class. (You may want to draw some histograms.)

3. There are two types of widgets, foos and bars. Some widgets contain baz and some do not. Consider the following contingency table.

	baz	
type	absent	present
foo	7	611
bar	250	694

(a) How many widgets are foos? How many are bars? How many contain baz? What is the probability that a random widget is foo? What is the probability that a random widget contains baz?

(b) What is the entropy of widget type?

(c) What is the entropy of widget type conditional on the presence of baz?

(d) What is the mutual information between widget type and baz presence?

4. Refer to the handout for lecture 6, where the greedy feature-selection is used to pick the seven most informative words in the *Times* corpus.

   (a) How much information does each of those words provide about the class, given the other six words?

   (b) Which words (if any) have positive interactions with the other six, and which (if any) have negative interactions?

   (c) Last time, you used leave-one-out cross-validation to evaluate the accuracy of the nearest neighbor classifier and the prototype classifier on the news stories. Re-run this *only* the seven selected features. What happens to the accuracies? (If you couldn't get your code to work properly for this part of the last problem set, use the code from the solutions.)

5. The code for lecture 6 contains a function to pick the q most informative features in a data-frame by greedy search.

   (a) How much information does the trio of words ("art", "painting", "evening") give about the story class?

   (b) Write a function which will pick the q most informative features to add to a given set of starting features. The function should take as inputs: a data frame, the column in the data frame which is to be predicted, the vector of features to start from, and q, the number of features to add. For full credit, the user should be able to specify features either by column numbers or column names.

   Test your function by checking that it gives the same results as in Lecture 6 when started with "art" and ("art","youre"), and $q = 1$ or $q = 2$.

   (c) What are the three most informative words to add to ("art", "painting", "evening")? How much information do they add? How much information do they provide on their own?