6.891 Fall 1999: Problem Set #5

Due: Thursday, 11/18/99
***** Final Version *****

Problem 1: Continue preparing for the final project.

By now you should have a pretty good idea what you will be working on for the final project.

Note: this problem set is a bit easier than the others. Hopefully this will make more time available for the final project.

Problem 2: Principal Components Analysis

Many data sets appear to be very high dimensional but are in reality quite redundant, like images. Intuitively, images are redundant because their pixels are highly correlated (e.g. if one pixel is bright the next pixel is very often bright, and vice versa).

One way to observe this redundancy is to look at the eigenvalues of the covariance matrix. If many of the eigenvalues are small, then the data is linearly redundant.

A: Show that if one of the eigenvalues of the covariance matrix is zero, then all of the data lies in a hyperplane that is perpendicular to the associated eigenvector.

B: Let D be the dimension of the data, and N the number of data points in a training set. Show that if D is larger than N the covariance matrix must have some zero eigenvectors.

In practice the eigenvalues of the covariance matrix of real data are rarely zero (except in the case above). What you often see is a rapid decrease in the size of the eigenvalues – the largest are big, but size rapidly drops off and approaches zero. In general you would like to reduce the number of dimensions, but what is the right number?

Images are the classic example of high dimensional data. We have made a dataset containing 16x16 digits digitized from zip codes available in the file `all_chars.mat` (see file `demo_pca.m` for a tip on how to display the data).

Load the data and compute the eigenvectors of the covariance matrix (Hint, use svd to compute the eigenvectors... it makes things easier). Using the 60 largest eigenvectors, reduce the dimensionality of the data to 60. From this representation, reconstruct the data and display it. Compute the sum of the squared error in reconstruction across the entire training set.

C: Look at the quality of the reconstructions for different numbers of dimensions. Based on the appearance of the characters what is a good compromise between dimensionality and quality?

D: Explore the relationship between the dimensionality of the training set and classifications performance. Choose your favorite classifier from problem set 2 or 3 (or your favorite from somewhere else – be sure to describe it clearly). For various numbers of dimensions build a classifier for the digit 2 (i.e. that distinguishes 2's from non-2's). You can use cross validation to pick the optimal number if you are adventurous. Did this number of dimensions agree with your intuitive selection?

Problem 3: Independent Components Analysis

I did not have enough time to present the ICA approach in complete detail. The basic idea is that in many cases a signal can be thought of as sum of independent components. Examples of this phenomena abound: microphones pick up the sum of the sounds in a room, EEG (electro-encephalography) picks up a sum of the activity in the brain, and cameras can pick up a sum of the light from different sources. ICA attempts to pull out these independent components. Note that this is not the same thing as decorrelation, which removes linear dependency in the data.

The amazing thing about the ICA algorithm of Bell and Sejnowski is that it is so very simple. A nice reference for the ICA algorithm is Bell's original paper which is available from related-info.

The goal of this problem set problem is to expose you to the ideas behind ICA. Here we will use it to separate images. The files `store1.bmp` and `store2.bmp` are both images of a department store display case with a reflection taken through a polarized lens. They were kindly provided by Hany Farid who has written a paper on how to use mutual information and entropy to separate such images into the reflection and the object behind the reflection (see related info page for a reference to this paper which contains a nice description of the problem).

We have provided two (hopefully) useful functions: `loadimage` and `drawimage`. The first returns a column vector representing the image and the former accepts a column vector and displays it. You may also find `saveimage` helpful (it saves instead of displaying a column-vector representation of an image).

We have provided you with code which implements ICA specifically for the image separation task: `ica_batch.m`.

A: Run this code on the images provided. Show the resulting images. Note there is a noticeable problem with the result even after finding the optimal unmixing matrix. What is it? Can you modify the original images so that the final unmixed images are improved?

B: For extra credit: we have included the sounds from the demo played in class (which involves 4 voices and 4 microphones). The files are called `mix1.wav`, `mix2.wav`, `mix3.wav`, and `mix4.wav`. See if you can unmix these using similar code. The matlab function to read these sounds is `[Y,FS,BITS]=WAVREAD(FILE)`. The sounds can be played using the `sound` function – though it may not work on all machines!!

Problem 4: Bayes Net Warmup

One of the great things about Bayes Nets is that they provide a visual tool for looking at and understanding probabilistic models. The visual language of Bayes Nets is a bit confusing at first, but it sure beats trying to figure out independence and dependence by multiplying out probabilities.

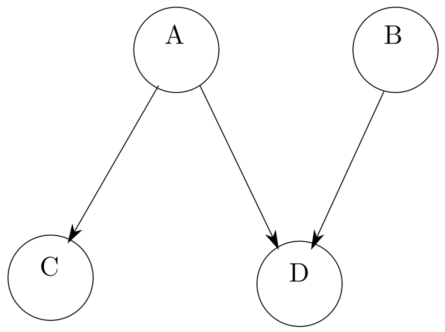

Figure 1: A Simple Bayes Net.

A: After looking at Figure 1, we can already tell that the joint distribution $P(A, B, C, D)$ has a simplified form. Write out that simplified form as a product of marginal and conditional distributions. It is possible to express this joint distribution using only *marginal distributions*. What is this expression?

B: Prove that nodes A and B are independent if we have no knowledge of either C or D.

C: Under what conditions are A and B dependent? Why?

D: Give two conditions under which C and D are independent.

E: Let G be independent of E given F. Given P(E,F) and P(F,G) what is P(E,F,G)?

Problem 5: Bayes Nets for Text

In this problem you will setup a Bayes Net that models the occurrence of words in two classes of documents (much like we saw early in class when we discussed document retrieval).

In our little world there are only two types of documents: sports and politics. As we saw before, each document will be treated as a set of words... order of the words will not matter nor will the number of times a word occurs.

The table below shows the percentage of time that political and sports documents contain one of five words: national, football, contest, votes, and battle. Note that we will ignore all other words.

Word	Politics	Sports
national	35%	55%
football	3%	40%
contest	30%	25%
votes	45%	2%
battle	15%	25%

Political documents are twice as common as sports documents.

A: Setup this problem up as a Bayes net in either JavaBayes or the Bayes Net Toolkit. Either take a screen dump of JavaBayes or show us your code from the Bayes Net Toolkit.

JavaBayes can be run in several ways: going to `/mit/6.891/ps5_code/JavaBayes/Classes` directory and typing `java JavaBayes`, by running the code over the web at `http://www.ai.mit.edu/courses/6.891/ps/5/matlab/JavaBayes/demo.html` or by downloading the entire distribution (see `JavaBayes-0.341.tar`) and then running `java JavaBayes`. If you run JavaBayes through your browser, you will not be able to save out your networks.

The Bayes Net Toolkit is far more powerful and flexible, but it has no user interface. It is certainly worth a look at this code... especially if you are interested in Bayes Nets (see the file `BNT.tar` or the directory `BNT`). But, it will be harder to setup the network.

Note for some of the problems below it might be easier to reason out the answer than to use the Bayes Net software. Try to do both.

B: What is the most likely set of words to occur in a political document? What about sports?

C: What set of words is most likely to be a sports document? What is the probability of this document being a sports document?

D: Given that the words national and football appear in a document, what is the probability that the word votes will appear?