Modeling
Fall 2009

Exam 2: Take Home

This is the take-home portion of Exam 2. You will be tested on your use of Matlab; you may use your class notes and the text. You are expected to work by yourself. You should type the results in a script file that you will publish to turn in. If you prefer it, you may turn in one file per question. Be sure to include comments that finish answering each question!

Due: Thursday. If you require it, you may work on it Thursday evening. Email me the PDF file of your finished work.

1. This question asks you to verify some of the statements we made in constructing the best basis. In particular, we said (p. 95, Ch 6 notes) that using p vectors in $\mathbb{R}^n$, $\mathbf{x}^{(i)}$, $i = 1..p$, and an arbitrary (fixed) orthonormal basis for $\mathbb{R}^n$, $\phi_1, \ldots, \phi_n$ (which are vectors), then the error in using k basis vectors is given by:

$$E_{mse} = \frac{1}{p} \sum_{i=1}^{p} \|\mathbf{x}_{err}^{(i)}\|^2 = \sum_{j=k+1}^{n} \phi_j^T C \phi_j$$

   where C is the covariance matrix (with $1/p$ rather than $1/(p-1)$). Furthermore, this error is minimized when $\phi_i = \mathbf{v}_i$, where $\mathbf{v}_i$ are the eigenvectors of C (ordered by the magnitude of the eigenvalues, largest to smallest). In Question 1, you will verify these statements in Matlab using arbitrary data, and fix $k = 1$. That is, you will:

   - Construct 10 (arbitrary) points in $\mathbb{R}^{50}$ using `randn`.
   - Mean subtract the data (necessary for the equation above).
   - Compute the covariance, C. Verify that Matlab's `cov` function gives you the same answer as our definition (use the help to find out how to divide by p instead of $p-1$).
   - Now compute the mean square error two ways:
     - One by using the definition of $\mathbf{x}_{err}^{(i)}$ on page 94.
     - Verify that $\mathbf{v}_1$ as an eigenvector of C is the first vector from the SVD of your data (It could be U or V depending on how you arranged your data). Hint: It could be multiplied by -1 (but that is OK).
     - Compare this to $\mathbf{v}_1^T C \mathbf{v}_1$, which should be the same as λ_1 (and verify that).

   Hint: A problem using `eig` for the eigenvalues/vectors is that they come unordered. Better to use the SVD- For example, `[U,S,V]=svd(C)` will return the ordered eigenvalues of C in S and the corresponding eigenvectors in U (or V, they are the same).

2. Pattern classification: Construct a pattern classifier using two methods- (i) Batch using the pseudo-inverse function we wrote in class, and (ii) Using Widrow-Hoff (or equivalently modified Hebb). Compare your answers by constructing and plotting the decision boundaries you get for each method. Download the data and a script file for plotting the data from our class website (at the bottom of the page). Running the script file will produce your targets. Try several values of α to determine a reasonable value (that means to try a few, you don't need to spend a lot of time here, but look at your results and pick the best).

   (NOTE: Do not use `wid_hoff1.m`, write your own code!).

3. A Best Basis for the Space of Faces.

   In this question, you will be given a set of faces (much like the movie data, and your solution to that HW question should be what you use here).

   The data represents 30 photographs of undergraduate students from Whitman. Each photograph is an array of 294×262 pixels each, and they are stored as vectors in $\mathbb{R}^{77028}$. Therefore, when you type `load Faces` you will have a matrix Y that is 77028×30. Be sure to type the following when you start:

```
load Faces.mat
Y=double(Y);
```

   (The Faces.mat file is on our class website).

   You will also see two other vectors, `boys` and `girls`. The vectors contain the indices for the photos of boys and girls, respectively (so that `Y(:,girls)` would contain the data only for the girls, for example).

   We want to find the best basis for the space of faces in this database- That is, the best basis in $\mathbb{R}^{77028}$- Just like we did for the sequence of movie frames. In particular, you should:

   (a) Find the mean face and visualize it in Figure 1. Use the `reshape` and `imagesc` commands- remember, these pictures have slightly different dimensions than the movie frames.

   For extra fun, find the mean boy and the mean girl!

   (b) Mean-subtract your data (your mean is a vector in $\mathbb{R}^{77028}$). Find the best basis vectors via the SVD. Visualize the first four "eigenfaces" using the `subplot`, `reshape`, and `imagesc` commands. Put these in Figure 2.

   (c) Choose a face at random, and construct the 5, 10 and 15 dimensional reconstructions. Plot the corresponding images in Figure 3 (using `subplot`).

   (d) Project your faces to the best two dimensional representation (you should get a matrix that is 30×2 or 2×30). In Figure 4, plot the boys as red asterisks and the girls as blue diamonds. For example, if your matrix `Coords` is 30×2, type:

```
plot(Coords(boys,1),Coords(boys,2),'r*');
hold on;
plot(Coords(girls,1),Coords(girls,2),'b^');
hold off
```