CS273a Homework #5

Introduction to Machine Learning: Fall 2012

Due: Friday December 7th, 2012

Write neatly (or type) and show all your work!

Problem 1: EigenFaces

In class I mentioned that PCA has been applied to faces, and showed some of the results. Here, you'll explore this representation yourself. First, load the data and display a few faces to make sure you understand the data format:

```
X = load('data/faces.txt');          % load face dataset
img = reshape(X(i,:),[24 24]);  % convert vectorized datum to 24x24 image patch
imagesc(img); axis square; colormap gray;    % display an image patch; you may have to squint
```

(a) Subtract the mean of the face images ($X_0 = X - \mu$) to make your data zero-mean.

(b) Take the SVD of the data, so that

$$X_0 \approx U \cdot S \cdot V^T$$

Note that since the number of data is larger than the number of dimensions, S will be zero on its lower part; I suggest taking $X_0 \approx W \cdot V$ where $W = U \cdot S$. You may also prefer to use `svds`, which can return only the top K singular values and their associated columns of U and V.

(c) For $K = 1 \ldots 10$, compute the approximation to X_0 given by $\hat{X}_0 = W(:, 1:K) \cdot V(:, 1:K)^T$, and compute the mean squared error in the SVD's approximation, `mean(mean( (`$X_0 - \hat{X}_0$`).^2 ))`. Plot these values as a function of K.

(d) Display the first few principal directions of the data, by computing $\mu + \alpha$ V(:,j)' and $\mu - \alpha$ V(:,j)', where α is a scale factor (I suggest, for example, `2*median(abs(W(:,j)))`, to get a sense of the scale found in the data). These should be vectors of length 24^2, so you can reshape them and view them as "face images" similar to the original data.

(e) These are often called "latent space" methods, as the coefficients can be interpreted as a new geometric space in which the data are being described. To visualize this, choose a few faces at random (say, about 15–25), and display them as images with the coordinates given by their coefficients on the first two principal components:

```
idx = ...      % pick some data at random or otherwise
figure; hold on; axis ij; colormap(gray);
range = max(W(idx,1:2)) - min(W(idx,1:2)); % find range of coordinates to be plotted
scale = [200 200]./range;                  % want 24x24 to be visible but not large on new scale
for i=idx, imagesc(W(i,1)*scale(1),W(i,2)*scale(2), reshape(X(i,:),24,24)); end;
```

This can often help you get a "feel" for what the latent representation is capturing.

(f) Choose two faces and reconstruct them using only K principal directions, for $K = 5, 10, 50$.

Problem 2: Latent Semantic Indexing

In this problem we'll use the SVD to describe the text data we used in Homework 4. Again, load and normalize the data:

```
cd text
% Read in vocabulary and data (word counts per document)
[vocab] = textread('vocab.txt','%s');
[did,wid,cnt] = textread('docword.txt','%d%d%d','headerlines',3);

X = sparse(did,wid,cnt);  % convert to a matlab sparse matrix
D = max(did);             % number of docs
W = max(wid);             % size of vocab
N = sum(cnt);             % total number of words

% It is often helpful to normalize by the document length:
Xn= X ./ repmat(sum(X,2),[1,W]) ; % divide word counts by doc length
```

As in Problem 1, you should find a singular value decomposition of a matrix

$$Xn \approx U \cdot S \cdot V^T$$

where again U is $D \times T$ and V is $W \times T$, giving a low-rank (rank T) approximation to X.

Normally (in most problems) we would subtract the mean from X before performing PCA, but in text this is not usually done. (Don't do it for this problem.)

(a) Find the SVD of Xn using `svd` or `svds`, keeping or computing $T = 8$ components (corresponding to the top T singular values).

(b) The vectors of V indicate the words that define each direction in the latent space, along with their sign. Find the 10 "most positive" and 10 "most negative" words in each topic – you can do this using `sort`'s second return value – and print them out, similarly to printing the clusters in Homework 4. Can you interpret some or all of these "topics"?

(c) Choose a topic and sign (positive/negative) that you feel is relatively interpretable, and find the index of the three documents that have the largest magnitude coefficient in this direction. Print out several lines of each document (again similarly to Homework 4). Are the "topics" a good description of the documents?

As a side note, the non-interpretability of "negative" coefficents in explaining word probabilities is one reason many text models use non-negative matrix factorization techniques, which explain the (strictly non-negative) entries of Xn with a produce $W \cdot V$ whose entries are also strictly non-negative.

Problem 3: Go work on your project