Project 1 Principal Component Analysis[1]

Problem 1 (5 points)

Consider 3 data points in the 2D space: (-1,-1), (0,0), (1,1)
1) What is the first principal component? Write down the actual vector.

2) If we project the original data points into the 1D subspace spanned by the principal component you choose, what are their coordinates in the 1D subspace? What is the variance of the projected data?

3) For the projected data you just obtained above, if we represent them in the original 2D space and consider them as the reconstruction of the original data space, what is the reconstruction error?

Problem 2 (20 points) Computer Assignment

Part 1: You are asked to program the following functions using MATLAB or any other language you like.

- function [Lambda,U,meanX]=MyPCA(X)

```
%%
% Performs the extraction of the PCA components given a dataset
% Input       : X, DxN matrix of N points x of dimension D
% Output      :
%   Lambda  : set of eigenvalues of the covariance matrix ranked in the decreasing order
%   U       : matrix of eigenvectors (ranked in the same order as eigenvalues)
%   meanX   : mean of the data X
%
```

For this function, you can use for instance (and appropriately) the eig or SVD eigenvalue decomposition function from MATLAB. Do not use MATLAB's princomp or any other built-in functions for PCA. Same for other open source software.

- function [Y]=PCAProjection(Z,meanX,P)

```
%%
% Projects a matrix of data points Z on the first M eigenvectors
% Input:
%   Z       : DxN data matrix (N columns of data zi of dimension D)
```

[1] Adopted from UCSD ECE 175A, EPFL EE613, and CMU CS 10-701.

% meanX : mean of data points provided by MyPCA
% P : DxM projection matrix containing the first M eigenvectors obtained from MyPCA
% Output:
% Y : MxN matrix containing the components in PCA subspace for all data points of Z
%

- function [Ztilde]=PCAReconstruction(Y,meanX,P)

%%
% Reconstructs data points given their coordinates Y in the space spanned by the M eigenvectors of P
% Input:
% Y : MxN coordinates of the N points to (re)construct
% meanX : mean of data points provided by MyPCA
% P : DxM projection matrix containing the first M eigenvectors obtained from MyPCA
% Output:
% Ztilde : DxN matrix containing the constructed vectors
%

Part 2: Now use your code to do a PCA analysis on handwritten digits.

Dataset: We will use the USPS (United States Postal Service) digit dataset (usps.mat file). To perform the task, it is simplest to convert all digit images into a column vector, and you can do this by stacking the columns of the digit matrix on top of each other. After this, each image will have a representation as a column vector.

Task 1: Select one digit (avoid one or zero) and extract the image data corresponding to this character.
1) Compute the principal components, and visualize them in 2D images. Comment on what is captured by each eigenvector (if it makes sense). You only need to show the top few, say, top 10 or 20.
2) (I) Plot the eigenvalues of the pooled dataset covariance matrix in the order of decreasing values. (II) Then, for the same plot, relabel the vertical axis to show the percentage that *each individual eigenvalue* contributes to *the total variance* (the total variance is equal to the sum of all eigenvalues). You can plot them in two different figures or provide both labels on the same plot, as you like. (III) Plot of the percentage of explained variance by the *sum* of the k largest eigenvalues for a few chosen k's.
3) Select a few images from the same digit class, and reconstruct them using k=2, 5, 10, 50 or more first eigenvectors. Calculate the average reconstruction error $\left|\left|\tilde{z}-z\right|\right|^2$ for different k's. According to you, what is the appropriate number of

eigenvectors needed to compress the selected digit?
4) Select a few images of a digit from another class. Reconstruct them with the same number of eigenvectors as above (in step 3). Comment on the results.

Task 2: Now, use all the digit dataset.
1) Apply PCA and visualize the top 10 eigenvectors (in 2D images). Comment on the results.
2) Repeat Step 3 in Task 1, and comment on the results. How many eigenvectors do we need to compress the digits?

Note:

In PCA, usually we consider the scenario where the data number N is much larger than the dimensionality D. However, in this example (the handwritten digits), it is the opposite. In this case, there is little point in applying PCA for values of M that are greater than N-1, since there will be many eigenvalues with value zero. Also, it is computationally infeasible to directly apply PCA and finding eigenvectors of a DxD matrix for high-dimensional data. There are more efficient PCA techniques for high-dimensional data in the literature. Please check Section 12.1.4 in the Bishop textbook “Pattern Recognition and Machine Learning” by Springer, or use google to find appropriate references.