CSE 881: Data Mining (Fall 2016 Homework 2)

Due date: Sept 20, 2016 (before midnight).

A soft copy of your homework must be submitted via handin. All submitted homework must be your own work.

1. Consider a dataset that has 3 attributes (x_1, x_2, and x_3). The distribution of each attribute is as follows and shown in Figure

   - x_1 has a uniform distribution in the range between 0 and 1.
   - x_2 is generated from a mixture of 3 Gaussian distributions centered at 0.1, 0.5, and 0.9, respectively. The standard deviation of the distributions are 0.02, 0.1, and 0.02, respectively.
   - x_3 is generated from an exponential distribution with mean 0.1.

   (a) Which attribute is likely to produce the same bins regardless of whether you use equal width or equal frequency approaches (assuming the number of bins is not too large).

   (b) Which attribute is more suitable for equal frequency than equal width discretization approaches.

   (c) Which attribute is not appropriate for both equal width and equal frequency discretization approaches.

   (d) If all 3 are initially ratio attributes, what are their attribute types after discretization?

2. The purpose of this exercise is to illustrate the relationship between PCA and SVD. Let $\mathbf{A}$ be an $N \times d$ rectangular data matrix and $\mathbf{C}$ be its $d \times d$ covariance matrix.

   (a) Suppose $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an $N \times N$ matrix whose elements are equal to 1, i.e., $\forall i, j : (\mathbf{1})_{ij} = 1$. Show that the covariance matrix $\mathbf{C}$ can be expressed into the following form:

   $$\mathbf{C} = \frac{1}{N-1}\mathbf{A}^T\left[\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right]\mathbf{A}$$

   .

   (b) Using singular value decomposition, the matrix $\mathbf{A}$ can be factorized as follows: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is the $N \times N$ left singular matrix, $\mathbf{\Sigma}$ is the $N \times d$ matrix containing the singular values, and $\mathbf{V}$ is the $d \times d$ right singular matrix. Similarly, using eigenvalue decomposition, the covariance matrix can be factorized as follows: $\mathbf{C} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T$. Show the relationship between SVD and PCA is given by the following equation:

   $$\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T - \frac{1}{N}\mathbf{A}^T\mathbf{1}_N\mathbf{A} = (N-1)\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T.$$

(c) Find the relationship between the right singular matrix $\mathbf{V}$ and the matrix of principal components $\mathbf{X}$ if the data matrix $\mathbf{A}$ has been column-centered (i.e., every column of $\mathbf{A}$ has been subtracted by the column mean) before applying SVD.

3. Download a collection of handwritten digit images from `http://www.cs.nyu.edu/~roweis/data/mnist_all.mat`. The dimensionality of each image is 28×28 pixels. In this exercise, you will apply principal component analysis to reduce the rank of the images. You will use only the training images for digits 0, 1, 2, and 3 (ignore the images for other digits). You should also ignore the test images in the data set.

   (a) Create a data matrix $\mathbf{X}$ of size 200×784 by choosing the first 50 rows of train0, train1, train2, and train3 and append them together. Note that the matrix train0 contains images for digit 0, train1 contains images for digit 1, and so on. Each column in the matrix represents one of the 784 pixels (28×28) of the corresponding images.

   (b) Plot the resulting images using the Matlab script given below:

```
N = 50;         % number of images associated with each digit
numCols = 10;
numRows = ceil(4*N/numCols);
d = sqrt(size(X,2));

figure;
set(gcf,'color','white');
set(gcf,'Position',[520 85 1020 720]);  % This command will resize the plot
for i=1:size(X,1);
    subplot(numRows,numCols,i);
    img = reshape(X(i,:),d,d)';     % convert each row into 28 x 28 matrix
    imagesc(img);                   % plot the image
    set(gca,'xtick',[]);
    set(gca,'ytick',[]);
end;
colormap(gray);         % convert the images into gray scale
```

   Save the images into a jpg file as follows:

```
matlab> saveas(gcf,'digit_image.jpg','jpeg');
```

   Append the resulting figure with your homework solution.

   (c) Use the `pca` command to generate the principal components of the matrix $\mathbf{X}$.

```
[U,Z,S] = pca(X);
```

   Note that the matrix $\mathbf{U}$ contains the eigenvectors (i.e., principal components) of the covariance matrix for $\mathbf{X}$, $\mathbf{Z}$ represents the projection of each row in $\mathbf{X}$ onto the subspace spanned by the principal components, and $\mathbf{S}$ is a vector containing the variance explained by each principal component. Plot the images associated with the first two

principal components. Hint: use the `reshape` and `imagesc` commands shown in part (b) above. Each image should be plotted in a separate figure. Save the figures as jpeg images and attach them with your homework solution.

(d) Reduce the dimensionality of each data point from 784 to 2 by projecting the data to its first two principal components. This is given by the first two columns of the matrix **Z**. Draw a scatter plot of the data points, using different markers to represent each digit.

```
figure;
set(gcf,'color','white');
plot(Z(1:50,1),Z(1:50,2),'r*');         % images for digit 0 is shown as *
hold on
plot(Z(51:100,1),Z(51:100,2),'b+');     % images for digit 1 is shown as +
plot(Z(101:150,1),Z(101:150,2),'ko');   % images for digit 2 is shown as o
plot(Z(151:200,1),Z(151:200,2),'gv');   % images for digit 3 is shown as triangles
hold off
```

Save the resulting plot into a jpeg image and append it to your solution file. Based on the plot, answer the following question: which classes are easier to be discerned by the first two components and which are harder to be discerned?

(e) Using the script in part (b), plot the resulting digit images when the data is reduced to a matrix of rank 2. To create the reduced-rank matrix **W**, you need to do the following:

```
rank = 2;
W = Z(:,1:rank)*diag(S(1:rank))*U(:,1:rank)';
```

Save the resulting images as a jpeg file and attach it to your solution. Which digits can be more easily discerned and which are harder? Is it consistent with your answer in part (d)?

(f) Repeat part (e) to re-create the digit images using a matrix of rank 50. Can you visually discern more digit images correctly with the increasing rank of the matrix **W**?

Make sure you insert all the figures to your pdf solution file (and label them appropriately so we know which figure is for which question) instead of submitting each figure as a separate jpeg file to handin. You should put all your Matlab code in a single file named `q3.m`. Make sure you add comments to the different parts of your code (using 'represent a comment). Submit the Matlab code to handin as a separate file from the rest of the homework.

4. Kernel PCA

In this exercise, you will apply kernel PCA to the handwritten digit image data created in the previous question. Extract the first two nonlinear principal components and draw the corresponding scatter plot (similar to question 3(d)). You should center the kernel matrix first before performing an eigendecomposition of the matrix.

```
1. Compute the distance between every pair of points and its median
    dist = pdist(X);
    med = median(dist);
    dist = squareform(dist);
2. Set kernel parameter, gamma, based on median value of the distances
    gamma = 1/med^2;
3. Create the Gaussian radial basis function kernel matrix:
    K = exp(-gamma * dist.^2);
4. Sparsify the kernel matrix by keeping only its nearest neighbors
    numNeighbors = 20;
    numZeros = size(K,2) - numNeighbors - 1;
    M = size(K,1);
    [temp,I] = sort(K,2);
    J = repmat([1:M]',1,numZeros);
    I = sub2ind(size(K),J,I(:,1:numZeros));
    K(I) = 0;
5. Center the kernel matrix (read the supplementary notes).
    Kc = ...
6. Extract a matrix W that contains the first two eigenvectors of the centered kernel matrix
7. Project each data point to its first two components
    V = Kc * W
8. Draw a scatter plot of the projected values.
```

Make sure you insert the figure to your pdf solution file. Compare the scatter plots obtained using PCA and kernel PCA. Are there any digits for which kernel PCA can better discriminate their images than PCA? You should put all your code in a single Matlab program named `q4.m`. Submit the Matlab code to handin as a separate file from the rest of the homework.