ECE 532, Spring 2019

Homework #5

Due: Tue, Apr 8 @ 2:30 pm

Problem 1.

(a) We will first generate a synthetic data set in two dimensions using a "Gaussian spray can." Let $\mu = \begin{bmatrix} 4 \\ 10 \end{bmatrix}$, $W = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and $\sigma^2 = 4$. Take $z_i = 0$, and generate $M = 10$ i.i.d. data points from a $N(Wz_i + \mu, \sigma^2 I)$ distribution. Provide a scatter plot of the result—this is one "squeeze" of the spray can.

(b) Now repeat a similar procedure as in (a), but this time, first generate a data set $\{z_i\}_{i=1}^N$, for $N = 50$, from a $N(0, 1)$ distribution. For each value of z_i, generate $M = 10$ i.i.d. data points from a $N(Wz_i + \mu, \sigma^2 I)$ distribution. Provide a scatter plot of the resulting data set, which should have a total of $n = MN$ points.

(c) Perform PCA on the data set $\{x_i\}$, by decomposing the matrix $S = \frac{1}{n}\widetilde{X}^T\widetilde{X}$, and plot the first principal component u_1 as a unit eigenvector on the scatter plot in (b).

(d) Suppose you forgot to first recenter the data set before applying PCA. Perform an eigendecomposition of $S = \frac{1}{n}X^T X$ and plot the unit eigenvector v_1 corresponding to the largest eigenvalue of S on the same scatter plot. Explain the result.

(e) Technically, the procedure in (b) is not exactly the same as the generative model for PPCA. What is the difference? Generate $n = 500$ data points according to the probabilistic model for PPCA, and provide a scatter plot of the result.

Problem 2. The file `X_train.csv` contains image data from 360 faces, where each row corresponds to a 112×92 image. You can view the i^{th} image using the commands `colormap(gray); imagesc(reshape(X(i,:),112,92))`, if X is the data matrix. The file `Y_train.csv` contains labels identifying the individuals for the corresponding rows of `X_train.csv`. The files `X_test.csv` and `Y_test.csv` contain a separate set of test image data, with 40 faces.

(a) Perform PCA on the matrix of training data, and extract the top 10 principal components. Use the `imagesc` function to visualize these top 10 "eigenfaces." Roughly speaking, which facial characteristics are captured in each eigenface?

(b) Plot the eigenvalues of the matrix $S = \frac{1}{n}\widetilde{X}^T\widetilde{X}$ in decreasing order. What are the elbow(s) of the plot? What percentage of explained variance is captured by these point(s)?

(c) We will now incorporate the image labels to perform face recognition on the test data set, using the following steps:

   - Perform PCA on the training data to reduce the dimension from $p = 10304$ to $q = 50$.
   - For each image in the test data set, first project the image onto the 50-dimensional space.
   - Identify the closest point among the training data in the projected space (in terms of Euclidean distance). Use the label of this training point as the predicted label of the test point. This method is called 1-nearest neighbor classification.

   Report the fraction of correctly classified images in the test data set based on this classification procedure. For the images that are misclassified in the test set, display both the test image and the nearest neighbor in the training data set using the `imagesc` function.

(d) Now try performing the 1-nearest neighbor classification method in part (c), without first applying PCA. What is the resulting fraction of correctly classified images in the test data set?

Problem 3. We return to the Wisconsin breast cancer data set `bc_wisc.csv` analyzed in HW 3. Recall that the 560 rows in the data matrix correspond to different individuals, with the first column listing the ID number of the individual, the second column recording whether the tumor was benign (0) or malignant (1), and the remaining 30 columns listing measurements related to the tumor.

(a) Try running logistic regression on the 30-dimensional data set. What happens?

(b) Now run PCA to perform dimension reduction with $q = 2, 3, 4, \ldots, 10$. Report the percentage of variance explained for each value of q.

(c) For the values $q = 2$ and $q = 3$, visualize the projected data using 2D and 3D scatter plots. Use the markers o/+ for the categories 0/1. From looking at the scatter plots, do any points look like possible outliers? If so, report the row indices of these points.

(d) Finally, run logistic regression in the reduced q-dimensional space for $q = 2, 3, 4, \ldots, 10$. Report the proportion of correctly classified data points for the logistic regression classifier for each value of q. Are the results what you might have expected?