GE 461 Introduction to Data Science
Spring 2023
Project: Dimensionality Reduction and Visualization

Assigned: April 13, 2023
Due: 23:59, May 7, 2023

Character recognition has been a classical problem in pattern recognition. Digit recognition is an important problem in character recognition with many applications such as automatic routing of mails according to zip codes, automatic recognition of signature dates on documents, automatic digitization of distances and other labels on maps and technical drawings, and automatic recognition of license plates of vehicles.

In this assignment, the goal is to develop a handwritten digit recognition system. The input to the system will be a digitized image of a digit and the output will be the type of the digit $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Figure 1 shows example images from the MNIST database of handwritten digits (`http://yann.lecun.com/exdb/mnist/`) that has been heavily used in the literature. The data set that will be used in this assignment contains around 500 samples for each class (for a total of 5,000 samples). The data file (`digits.zip`) that is available on the course home page includes a text file named `digits.txt` that contains a matrix with 5,000 rows where each row corresponds to a sample, and another text file named `labels.txt` that contains a vector where each value shows the class for the corresponding sample. Each sample is a digitized 20×20 image of a digit (represented in column-major order as a 400-dimensional vector).

Download the digit data set and divide it into two subsets by randomly selecting half of the samples from each class for training and the remaining samples for testing. Then, perform the classification experiments described below. Use the same subsets for the rest of the assignment. Do not forget to use separate subsets for training and testing.

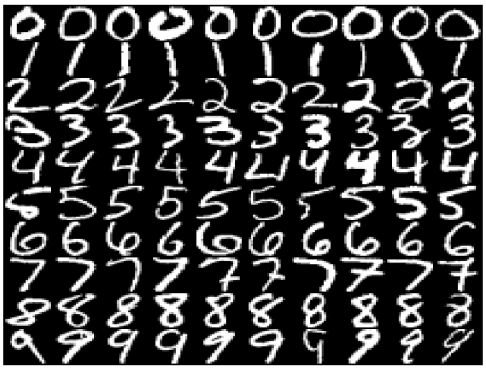

Figure 1: Example digits.

Hint: In Matlab, you can visualize the digits using the following piece of code:

```
load digits.txt
i = 1;  %select the digit to display
I = digits( i, : );
imagesc( reshape( I, 20, 20 ) );
colormap( gray );
axis image;
```

Background

A quadratic Gaussian classifier models each class with a Gaussian distribution. Gaussian can be considered as a model where the feature vectors for a given class are continuous-valued, randomly corrupted versions of a single typical or prototype vector. The Gaussian density for the feature vector $\mathbf{x} \in \mathbb{R}^d$ is defined as

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

where d is the dimension of $\mathbf{x}$, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix, respectively.

The training procedure involves fitting a Gaussian $p_c(\mathbf{x}|\boldsymbol{\mu_c}, \boldsymbol{\Sigma_c})$ to the corresponding subset of the training data for each class $c = 1, \ldots, C$ where C is the number of classes. In this project, fitting should be done by using the maximum likelihood estimates of the mean vector and the covariance matrix

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i}-\hat{\boldsymbol{\mu}})(\mathbf{x_i}-\hat{\boldsymbol{\mu}})^T$$

using the samples for each class. In the formulas above, n represents the number of samples that belong to a particular class. Estimation should be done separately by using the subset of samples for each class.

Once the Gaussian densities are learned for all classes, during the classification process, a sample can be assigned to the class c^* that gives the highest probability for that sample's feature vector $\mathbf{x}$ as

$$c^* = \underset{c=1,\ldots,C}{\arg\max}\; p_c(\mathbf{x}|\boldsymbol{\mu_c}, \boldsymbol{\Sigma_c}).$$

Given a data set with known class labels, the performance of the classifier can be evaluated by comparing the predicted class by the classifier to the true class known for each sample. The quantitative performance can be summarized using the classification error that is computed as the ratio of the number of wrongly predicted samples to the total number of samples.

Question 1 (35 pts)

In this question, you will use principal components analysis (PCA) to project the 400-dimensional data onto lower dimensional subspaces to observe the effect of dimensionality on the performance of the Gaussian classifier.

1. First, center the data by subtracting the mean of the whole data from each sample.

2. Then, use PCA to obtain a new set of bases (use the training data set, i.e., 2,500 samples for PCA). Plot the eigenvalues in descending order. How many components (subspace dimension) would you choose by just looking at this plot?

3. Display the sample mean for the whole training data set as an image (using samples for all classes together, but before centering in Step 1). Also display the bases (eigenvectors) that you chose as images (e.g., like in Figure 1) and discuss the results with respect to your expectations.

4. Choose different subspaces with dimensions between 1 and 200 (choose at least 20 different subspace dimensions, the more the better), and project the data (project both the training data and the test data using the transformation matrix estimated from the training data) onto these subspaces. Train a Gaussian classifier using data in each subspace (do not forget to use half of the data for training and the remaining half for testing).

5. Plot classification error vs. the number of components used for each subspace, and discuss your results. Compute the classification error for both the training set and the test set (training is always done using the training set), and provide two plots.

Question 2 (35 pts)

In this question, you will use Isomap (J. B. Tenenbaum, V. de Silva, J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," Science, vol. 290, pp:2319-2323, 2000) to map the 400-dimensional data onto lower dimensional manifolds.

1. Use Isomap to obtain low-dimensional embeddings of the digits data. Note that you need to use the full data set, i.e., 5,000 patterns, but you may still have several patterns that were not embedded. This is a common observation in many techniques that are based on neighborhood graphs where the embedding implementation only uses the largest connected component of the neighborhood graph and ignores the patterns belonging to other components.

2. Choose dimensions between 1 and 200 (choose at least 20 different dimensions, the more the better) and train a Gaussian classifier for each dimensionality (do not forget to use half of the data for training and the remaining half for testing).

3. Plot classification error vs. dimension, and discuss your results. Compute the classification error for both the training set and the test set (training is always done using the training set), and provide two plots. The discussion should include the setting (particular choice for the parameters) for Isomap, the effect of dimensionality on the classification error, and comparison of the Isomap results with the PCA results.

Question 3 (30 pts)

In this question, you will use dimensionality reduction particularly designed for visualization of high-dimensional data sets. In particular, you are asked to use t-SNE (L. J. P. van der Maaten and G. E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," Journal of Machine Learning Research, vol. 9, pp:2579-2605, November 2008) for mapping the digit data set to two dimensions. Compute the resulting mapping for the whole data set, and present the scatter of the samples together with their class information. Discuss the setup that you used (e.g., parameters needed for initialization, iterations, or stopping, etc), and comment on the resulting visualizations.

Note: You are required to upload your assignment report (as a pdf file) and all code that you wrote in a single archive (in zip format) to Moodle. Your report must include all required details listed for each question. You are free to write your own code or use other tools for the Gaussian classifier, PCA, Isomap, and t-SNE. However, you should still include any other code (e.g., experimentation, error analysis, plots, etc.) that you have written in your submission, and

properly cite all tools that you have used (you do not need to upload the external libraries together with your submission but provide a citation and a link in your report). Make sure that the tools you are using are correct implementations of the particular steps outlined in this assignment. Note that using code from other sources without proper citation will be considered as plagiarism.

Do not forget to write your name in the report that you are submitting. Also do not forget to include your name in the filename of the pdf file so that the filename becomes unique.