CS 383 - Machine Learning

Assignment 1 - Dimensionality Reduction

Introduction

In this assignment, in addition to related theory/math questions, you'll work on visualizing data and reducing its dimensionality.

You may not use any functions from machine learning library in your code, however you may use statistical functions. For example, if available you **MAY NOT** use functions like

- pca
- entropy

however you **MAY** use basic statistical functions like:

- std
- mean
- cov
- svd
- eig

Grading

Although all assignments will be weighed equally in computing your homework grade, below is the grading rubric we will use for this assignment:

Part 1 (Theory)	25pts
Part 2 (PCA)	30pts
Part 3 (Eigenfaces)	30pts
Report	15pts
TOTAL	100pts

Table 1: Grading Rubric

DataSets

Yale Faces Datasaet This dataset consists of 154 images (each of which is 243x320 pixels) taken from 14 people at 11 different viewing conditions (for our purposes, the first person was removed from the official dataset so person ID=2 is the first person).

The filename of each images encode class information:

$$\text{subject} < ID > . < condition >$$

Data obtained from: http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html

1 (25pts) Theory Questions

1. Consider the following supervised dataset consisting of 10 observations, each with two features (observable data is in X) and an associated label in Y:

$$\mathrm{X} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 2 & 0 \\ 2 & 1 \end{bmatrix} \quad \mathrm{Y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

   (a) Compute the average weighted entropy of the class label in the subsets created by splitting the dataset based on the value of the first feature. You may assume that the features are *categoricial* (5pts).

   (b) Now make the same computation, but as if we created subsets using the second feature! (5pts).

   (c) Which feature is more discriminating based on results in Part (a) (2pts)?

   (d) What are the principle components of the observed data X? For this (and the next) part you may assume that the features are *continuous* and therefore should zscore them. Make sure your final principle components are all unit length. You **MAY** use a utility function like *eig* or *svd* to determine these (5pts).

   (e) In your own words, describe these axis in terms of a conventional 2D Cartesian Coordinate system (3pts).

   (f) If we were to project our data down to 1-D using the principle component, what would the new data matrix X be (5pts).

2 (30pts) Dimensionality Reduction via PCA

Download and extract the dataset *yalefaces.zip* from Blackboard. This dataset has 154 images ($N = 154$) each of which is a 243x320 image ($D = 77760$). In order to process this data your script will need to:

1. Read in the list of files
2. Create a 154x1600 data matrix such that for each image file
   (a) Read in the image as a 2D array (243x320 pixels)
   (b) Subsample/resize the image to become a 40x40 pixel image (for processing speed). I suggest you use your image processing library to do this for you.
   (c) *Flatten* the image to a 1D array (1x1600)
   (d) Concatenate this as a row of your data matrix.

Once you have your data matrix, your script should:

1. Standardizes the data
2. Reduces the data to 2D using PCA (using the two most relevant eigenvectors).
3. Plots the data as points in 2D space for visualization

Recall that although you may not use any package ML functions like *pca*, you **may** use statistical functions like *svd, eig*.

Your graph should end up looking similar to Figure 1 (although it may be rotated differently, depending how you ordered things and/or your statistical library).

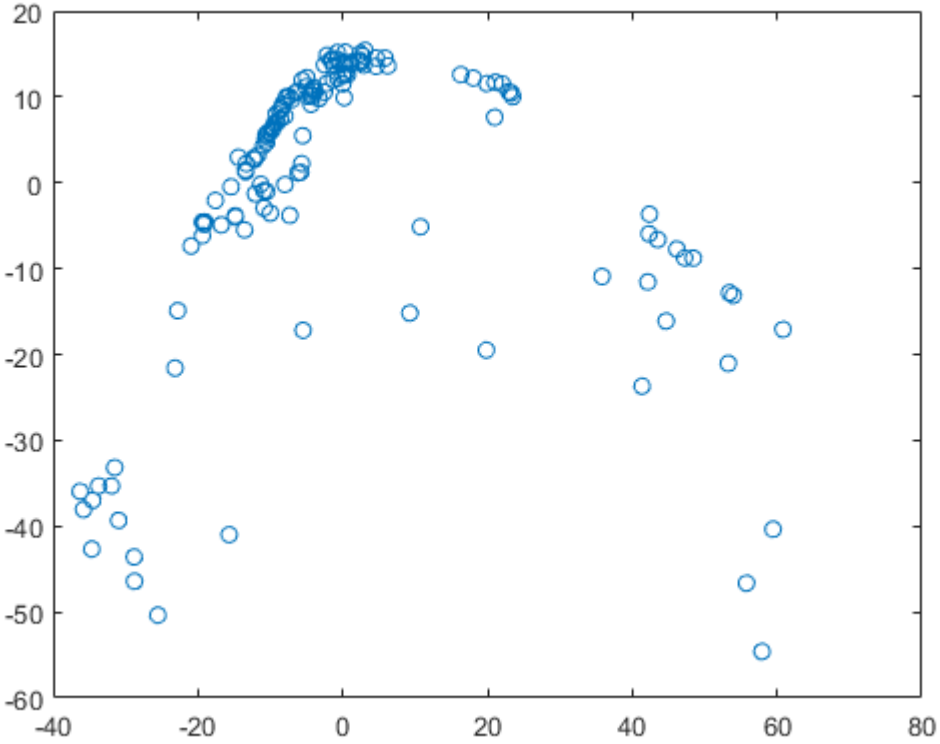

Figure 1: 2D PCA Projection of data

NOTE: Depending on your linear algebra package, the eigenvectors may have the opposite direction. This is fine since technically an eigenvector multiplied by any scalar are equivalent.

3 (30 points) Eigenfaces

One (cool?) application of PCA is lossy-compression. We can project our data down to k dimenions (where $k < D$), then when we need to reconstruct our data, we can just multiply the k dimensional data by (the transpose of) its k associated principle components in order to return to our original feature space!

In this part of the assignment you'll be making a *video* showing how the reconstruction looks as you varying k, from $k = 1$ to $k = D$. affects the reconstruction, visualizing this as a video. If you're working in Matlab I'd suggest using the *VideoWriter* class. If you're working in Python you might want to use Python's opencv module: *pip3 install -user opencv-python*.

Write a script that:

1. Takes your original 154×1600 standardized data matrix from the previous problem and
2. Performs PCA on the data (again, although you may not use any package ML functions like *pca*, you **may** use statistical functions like *svd, eig*).
3. For $k = 1, ..., D$:
   (a) Projects *subject02.centerlight* onto the k most important principle components, resulting in a feature vector of length k.
   (b) Reconstructs this person, again using the k most significant principle components (your feature vector should now once again be 1×1600).
   (c) Un-does the standardization (multiplies back in the std and adds back in the mean feature vector).
   (d) Reshapes this feature vector to be a 40×40 image/matrix.
   (e) Adds this image as the next frame in your video, superimposing on it the current value k.
4. Saves the video.

Note: When $k = D$ you should be able to perfectly reconstruct the original image.

Submission

For your submission, upload to Blackboard a single zip file containing:

1. PDF Writeup
2. Source Code
3. readme.txt file
4. You **do not** need to include the images (nor the video, we will regenate it all with your code). HOWEVER, it should be clear in your script (and/or readme) where your code expects the dataset to reside.

The readme.txt file should contain information on how to run your code to reproduce results for each part of the assignment. In particular for this assignment, it should also indicate where the *yalefaces* directory should be in order to run. Do not include spaces or special characters (other than the underscore character) in your file and directory names. Doing so may break our grading scripts.

The PDF document should contain the following:

1. Part 1: Your answers to the theory questions.
2. Part 2: The visualization of the PCA result
3. Part 3: (Nothing)