

DSCI 440W IA5 (100 pts)
Due 11:59 pm May 8, 2023

(100 pts) Implementation part: PCA

In this assignment you will work with the USPS handwritten digit dataset. In particular, the training data set contains handwritten digits 4 and 9. Each digit example is an image of by 16 by 16 pixels. Treating the gray-scale value of each pixel as a feature (between 0 and 255), each example has $16 \times 16 = 256$ features. For each class, we have 700 training examples. You can view these images collectively at

http://www.cs.nyu.edu/~rweis/data/usps_4.jpg

and

http://www.cs.nyu.edu/~rweis/data/usps_9.jpg

The data is in csv format and each row corresponds to a handwritten digit image (the first 256 columns) and its corresponding label (last column, 0 for digit 4 and 1 for digit 9). Note that you can use the Python command imshow (from matplotlib.pyplot) to view the image of a particular training example. For example, $X_{train}[0]$ is the 0th row vector of 256 dimensions for a particular digit image, the following code allows you to see the image (I like to display them in blue):

```
# plot training example (example with index 0)
plt.figure(figsize=(16, 16))
hlp = np.reshape(X_train[0], (16, 16))
color_map = plt.imshow(hlp.transpose())
color_map.set_cmap("Blues_r")
```

Here, I use numpy (as np) and matplotlib.pyplot (as plt).

In this assignment you want to apply PCA to reduce dimensions of the training vectors. In particular:

- Load the data. Make sure each training vector has 256 dimensions (features).

- Construct matrix

$$S = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^t,$$

where N is the number of training examples, x_i is the i th training vector (with 256 dimensions), and \bar{x} is a 256 dimensional mean (average) vector (off all training examples).

- Find eigenvalues and eigenvectors of S . Rank eigenvalues in a decreasing order. You may use function stem to plot eigenvalues. Make sure the eigenvectors are in the order that corresponds to the order of the eigenvalues.
- Determine the number of eigenvectors/eigenvalues to choose (i.e. new dimension) if you want to retain 75% of the variance after projection. Report this number.
- Now pick first three eigenvectors (i.e., we will project 256 dimensional data onto 3 dimensional space). Display these three eigenvectors using imshow. These are the "eigendigits" (analogy to "eigenfaces"). Insert the image of those eigendigits into your report. Do they look like 4 ? or 9? or both? Discuss your observations.
- Project each training vector (256 dimensional) onto 3 dimension space defined by first three eigenvectors. Plot new 3 dimensional data using scatter function. Make sure to use different color for each class.
- Discuss your observation. Are the two classes well separated? If yes, explain why. If not, explain why.