SML

Assignment 2

February 2024

1 Instructions

- You can use inbuilt libraries for Math, plotting, and handling the data (eg. NumPy, Pandas, Matplotlib).
- Usage instructions for other libraries can be found in the question.
- Only (*.py) files should be submitted for code.
- Create a (*.pdf) report explaining your assumptions, approach, results, and any further detail asked in the question.
- You should be able to replicate your results during demo.

---

2 Question-1

Use https://storage.googleapis.com/tensorflow/tf-keras-datasets/mnist.npz MNIST dataset for this question and perform following tasks.

- It has all in all 60K train samples from 10 classes and 10K test samples. 10 classes are digits from 0-9. Labels or classes for all samples in train and test set is available.
- Visualize 5 samples from each class in the train set in form of images.
- Images are of size 28×28. Vectorize them to make it 784 dimensional. Apply QDA on the given dataset. For each of the 10 classes you need to compute its mean vector and covariance vector. Use the QDA expression derived in the lecture. Your code should clearly have this expression. Note mean and covariance will be computed from train set only. Test set is not seen at this stage.
- Find the class of all samples in test set. Report accuracy and class-wise accuracy for testing dataset. Accuracy is ratio of total number of samples correctly classified to the total number of samples tested. Total number of samples tested is 10K. Similarly, for each class report the accuracy. Note the labels or classes for each sample is given in the dataset.

3 Question-2

Use same downloaded dataset from Question 1 and perform following tasks.

- Choose 100 samples from each class and create a 784×1000 data matrix. Let this be X.
- Remove mean from X.
- Apply PCA on the centralized X. You need to compute covariance $S = XX^\top/999$. The find its eigenvectors and eigenvalues. You can use any library for this. Sort them in descending order and create matrix U.
- Perform $Y = U^\top X$ and reconstruct $X_{recon} = UY$. Check the MSE between X and X_{recon}. This should be close to 0. MSE $= \sum_{i,j}(X(i,j) - X_{recon}(i,j))^2$.
- Now chose $p = 5, 10, 20$ eigenvectors from U. For each p, obtain U_pY, add mean that was removed from X, reshape each column to 28×28, and plot the image. You should see that as p increase the reconstructed images look more like their original counterparts. Plot 5 images from each class.
- Let test set be X_{test}. Find $Y = U_p^\top X_{test}$. For each value of p find Y, and apply QDA from Q1 on Y. Obtain accuracy on test set as well as per class accuracy. As p inreases, accuracy shall increase.