CSP 304: Machine Learning Lab (Spring 2023)

Indian Institute of Information Technology (IIIT), Kota
Instructor: Ankit Sharma

Posted in week: April 10- 14, 2023
Due in week: April 18-21, 2023

Homework 2

Dimensionality Reduction

Topic — This homework illustrates the working of PCA for dimensionality reduction.

Data — A pair of training data and test data are given in the data.zip file attached.

Problem

In this problem, you will apply dimension reduction and classification on the Optdigits dataset provided in optdigits train.txt and optdigits test.txt. Each row is an image vector and the last column is the digit *class*.

Q1 — **KNN Classifier without PCA (10 points):** Use the implementation (developed in *Homework 1*) for k-Nearest Neighbor (KNN) to classify the Optdigits dataset with $k = \{1, 3, 5, 7\}$. Print out the error rate on the test set for each value of k.

Q2 — **PCA Implementation (60 points):** Implement your own version of Principal Component Analysis (PCA) and apply it the Optdigits training data.

Q3 — **KNN Classifier with PCA (50 points):** Generate a plot of proportion of variance (see Figure $6.4(b)$ in the main textbook), and select the minimum number (K) of eigenvectors that explain at least 90% of the variance. Show both the plot and K in the report. Project the training and test data to the K principal components and run KNN on the projected data for $k = \{1, 3, 5, 7\}$. Print out the error rate on the test set for each value of k

Q4 — **Component Plotting (20 points):** Next, project both the training and test data to R^2 using only the first two principal components to plot all samples in the projected space and label some data points with the corresponding digit in 10 different colors for the 10 types of digits for a good visualization (similar to Figure 6.5 of Textbook).

Notes — For Q3 and Q4 you can use a ready-made implementation of PCA as well i.e. it is not mandatory (although *strongly encouraged*) to use your own implementation (from Q2). However, in case of using ready-made implementation of PCA, you will not get any points for Q2.

Deliverable

Report — A report in PDF format describing the various experimental tasks mentioned above. Description should include the details of your experiments (process & setup), results and discussions.

Code — Properly commented code file(s) or notebook(s) or any other setup and/or read-me files. All programming questions must be written in MATLAB/Python, no other programming languages will be accepted. And for Python only numpy and matplotlib can be relied on to implement the algorithm. Similarly, for MATLAB you have to provide your own implementation using basic functionalities

and no ready made libraries. For Q2 you can *only* use the *eigh* function in the *linalg* module of *numpy* to calculate eigenvalues and eigenvectors. To obtain distance between each pair of samples in KNN, you might consider to use *cdist* in the spatial.distance module of *scipy*. To visualize the projected data, you can use the *scatter* function in the *pyplot* module of *matplotlib*, and for adding text to corresponding point, you can use either *text* or *annotate* function in the *pyplot* module of *matplotlib*.

Your main function for Q2 should be: **myPCA**(*data*, *num_principal_components*). The function returns the principal components and the corresponding eigenvalues.

Bind all the files (report & code files) into a single ZIP file and upload as a response to Google classroom. Although you have worked in groups, but each member of the group should return the Google classroom assignment by uploading same material.

Notes Although not mandatory, the report is encouraged to be written in LaTeX preferably via www.overleaf.com using the NIPS format available here: https://neurips.cc/Conferences/2021/PaperInformation/StyleFiles. Also, an example overleaf NIPS template can be readily found here: https://www.overleaf.com/latex/templates/neurips-2021/bfjnthbqvhgs.

Eval. Demonstrate the various experiments you have conducted to the instructor during lab sessions.
