CSE455/CSE552 – Machine Learning (Spring 2018)

Homework #3

Handed out: 8:00pm Monday April 30, 2018.

Due: 11:55pm Sunday May 13, 2018.

Hand-in Policy: Via Moodle. No late submissions will be accepted.
Collaboration Policy: No collaboration is permitted.
Grading: This homework will be graded on the scale 100.

Description: The aim of this homework is to explore feature reduction techniques using PCA (principle component analysis). Use the following data for testing your implementation: (MNIST Digit Recogntion Data – available through `mnist.load_data()` in Keras).

Part I: Implementing PCA

Write the function **`pca(X)`** that takes an $n \times n$ matrix and returns **`mean`**, **`weights`** and **`vectors`**. The mean is the mean of the columns of X. The principle components of X are in **`vectors`**. The corresponding eigenvalues are in **`weights`**. You should use only a function performing SVD and nothing else from any Python libraries.

Part II: Using PCA before Classification

Using only a portion of the data (e.g., about 1000 images randomly chosen from the training set) perform PCA and train a classifier.

- Using the MNIST data, do a series of PCA-based reductions on the data. This should test at least four different values for the number of components chosen.
- Plot the class locations on the test data on a 2D map with horizontal axis as the first principal and with vertical axis as the second principal component (like the one discussed in class). Do the same for the first and third principal components. This should show you some clustering of the labels (better than if you just chose any two pixels).
- Feed the reduced features to a Random Forest Decision tree and show classification results using cross-validation. You should use all the data in training. This should be repeated for a few number of components extracted by PCA.

Part III: Comparing Linear and Non-linear Versions of PCA

Use an existing implementation of a non-linear PCA and repeat Part II.

What to hand in: You are expected to hand in one of the following

HW3_lastname_firstname_studentnumber_code.ipynb (the Python notebook file containing the code and report output).

Your notebook should include something like the following:

Part I: Code

Results:

Conclusions:

Part II: Code

Results:

Conclusions:

Part III: Code

Results:

Conclusions: