The American University in Cairo

Department of Computer Science and Engineering

CSCE 5269 – Pattern Analysis

Dr. Mohamed Moustafa	Assignment 2 [10%]	Spring 2018

Problem 1 (4%)

Design a classifier for the heart disease dataset: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

Details:

1. Divide the given data set into training set (first 216 data points) and testing set (the remaining 57 of points).
2. Using all 13 features, design a Logistic Regression classifier.
3. Repeat (2) but with an FDA-based classifier.

Deliverables:

a) Source code of your programs/scripts (using your preferred language). **[2 pts]**

b) Document containing:

   i. Plot ROC curves of both classifiers in one figure for easier comparison. Identify your best classifier for this problem. **[2 pts]**

reference results on this (and other) datasets:
http://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf

Problem 2 (6%)

Design a classifier that uses principal component analysis (PCA) for the MNIST database:
http://yann.lecun.com/exdb/mnist/
Input is a 28x28 grayscale image of a digit, and the output is from 0 to 9.

Deliverable:

a) Source code of your programs/scripts (using your preferred language). **[1 pt]**

b) Document containing:

   i. visualization of the first few eigenvectors for your data as images. **[1 pt]**

   ii. Plot of data variance preserved (y-axis) versus number of principal components. **[1 pt]**

   iii. Plot of Average CCR (y-axis) versus number of used principal components. **[1 pt]**

   iv. Your best Average CCR on the 10K testing set. **[2 pts] [Warning: this is a competitive part. The most accurate submission will get the full 2 pts. Others will get partial credit relative to their distances from the most accurate one]**