CIS*4020 Assignment 2

Classification

Total Marks: 36

Handed out: November 2nd

Due date: December 1st

The objectives of this assignment include:

i. Gaining familiarity with programming in Python
ii. Becoming comfortable with classification
iii. Learning to work with data, choose appropriate variables
iv. Explore challenges associated with labeled data

Requirements:

You are expected to hand in a completed version of the Jupyter notebook provided:

1. Details of what steps are required, and associated marks are included in the Jupyter notebook itself.

Note that your implementation for any given part of the problem may be simple (or complex) depending on design decisions made. I am most interested in your analysis of the problem, and design decisions that were made, and also presenting the opportunity to think critically about the problems presented. There are placeholders for you to include discussion, and the discussion you include is important to show that you have understood conceptual aspects of the problem.

Please rename the Jupyter notebook provided to lastname_firstname_A2.ipynb for your submission.

Getting started:

See the notes regarding getting started in the description for Assignment 2.

Summary of Assignment 2 elements (Detailed requirements are in the notebook markdown itself)

Part I

a. Sample a subset of the data based on class label
b. Apply PCA and LDA and visualize the results
c. Perform classification using naïve Bayes and interpret the results
d. Perform classification using Logistic Regression and interpret the results
e. Perform classification using SVMs and interpret the results

f. Comment on which approaches might allow you to best identify the confidence of class assignments (e.g. how happy, or how sad rather than just a label)
g. Plot results based on the approach in f.

Part II

a. Apply PCA and LDA. For PCA, plot the incremental gain in capturing the variance in the data for each successive principal component that is added and compare this with the subset of data in step 1.
b. Perform classification using naïve Bayes and interpret and visualize the results
c. Perform classification using Logistic Regression and interpret and visualize the results
d. Perform classification using SVMs and interpret and visualize the results
e. Discuss which classes are most similar / different with some justification given
f. Explain how you might identify faces that are mislabeled in the dataset and provide an implementation.

Challenge (required)

a. Based on a 7-dimensional vector of confidence values for face identity, plot the position of these in a 2D embedding. The “points” that are shown should be the actual face images within the plot (i.e. not just dots, but face pictures).