CS304 - Homework 3

Self Study Question (You do not need to submit this): Exercise 9 at the end of Chapter 8 from the textbook "Hands-on Machine Learning", 3E, A. Geron.

Question 1:

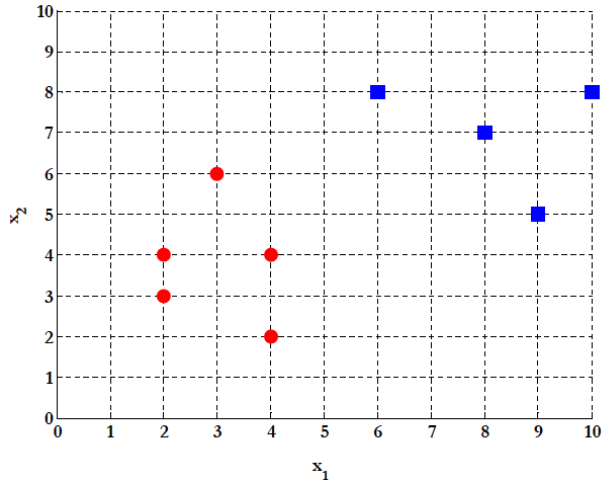
(a) Compute the PCA and LDA 1D projections for the following 2D dataset:

Samples for class 1 (ω_1): $\mathbf{X_1} = (x_1, x_2) = \{(4,2), (2,4), (2,3), (3,6), (4,4)\}$

Samples for class 2 (ω_2): $\mathbf{X_2} = (x_1, x_2) = \{(6,8), (9,5), (8,7), (10,8)\}$

Draw the two principal components on the plot for PCA as vectors, as well as the projected blue and red points in 1D.

Draw the projected 1D data using LDA.



(b) Repeat (a) for the following dataset:

Samples for class 1 (ω_1): $\mathbf{X_1} = (x_1, x_2) = \{(6,8), (2,4), (2,3), (3,6)\}$
Samples for class 2 (ω_2): $\mathbf{X_2} = (x_1, x_2) = \{(9,5), (8,7), (10,8), (4,2), (4,4)\}$

(c) Comment and compare how good the 1D projected data in (a) and (b) can be classified after PCA and LDA.

Note: Submit an .ipynb file for this question containing all of your code, plots and comments.

Question 2:

(a) Download the wine dataset from the official link given below. Inspect the features and the target values. https://archive.ics.uci.edu/dataset/109/wine

Separate dataset into 70% training and 30% test sets. Finally preprocess the data for normalization.

Remember that normalized data usually (not always!) performs better than the raw data.

(b) Find the principal components of the wine dataset.

Observe the explained variance of each components.

Plot the Explained variance ratio vs principal components using bar graph (matplotlib's library bar method.)

(c) Reduce the dimensionality to 2 and plot the scatter diagram. (**Hint:** Using matplotlib scatter method is an option.)

(d) Train a logistic regression classifier using the first 2 principal components and plot the decision regions.

(e) Classify the test data and plot the decision regions.

Visualization helps you to understand how accurately the classifier performs on the test data.

(f) Apply steps b – f using Linear Discriminant Analysis (LDA) and observe how the data is transformed using LDA and how well the test data is classified.

(g) Comment on your the results using PCA and LDA. Which dimensionality reduction method performs better? Why?

(h) The final step is visualization of the training dataset using t-SEN in 2D. Observe how t-SNE visualization performs on the dataset.

Note: Use the attached HW3_Q2_stud.ipynb file for this question and insert all of your code, plots and comments.