CS488/588 Homework 3 -70 pts

Due: March, 24 (Upload soft copy on Canvas as a single file with .docx or .pdf format)

For each of the below questions provide your analysis/inference of the results.

1a. Write a python program for dimensionality reduction using PCA on the Iris and Indian Pines dataset that implements the following: (25 points)

i) For PCA, plot the explained variance for all the PC's in the dataset.(5 points)

ii) Reduced data visualization using PCA to 2 dimensions – display the new transformed data which is reduced to two dimensions for visualization. – display the first two PC's (directions of projections) with respect to color-coded class separability.

(10 points: 5 points per dataset for PCA plots, i.e. 5pts per plots)

iii) Reduced data visualization using LDA to 2 dimensions – display the new transformed data which is reduced to two dimensions for visualization. – display the first two directions in LDA (directions of projections) with respect to color-coded class separability.

(10 points: 5 points per dataset for LDA plots, i.e. 5pts per plots)

*Note only provide data visualizations here.

The number of dimensions chosen for analysis from i) can be anything – provide your justification in 1b for your choice. For visualization in ii) and iii) only plot the first two of dimensions chosen. Also, number of PCs or LDs chosen must be same for all analysis.

1b. Discuss the analysis of results from 1a) for each dataset in terms of:

Role of dimensionality reduction /feature extraction on data analysis, inferences about data separability and your choice of 'K' PC's and LDA features to be retained in each dataset –which method worked best on each dataset and why?

(5 points: i.e. 2.5pts per dataset inferences)

* Use data visualization to draw analysis

2a. Write a python program to perform supervised classification on the Iris and Indian Pines datasets using Naïve Bayes, and Support vector machines (with RBF and Poly kernel) classifiers for training sizes ={10%, 20%, 30%, 40%, 50%} for each of the below cases:

i) with dimensionality reduction – Reduce data based on your choice of 'K' dimensions from 1a) using each of the dimensionality reduction methods (PCA, LDA) followed by supervised classification by the listed classifiers.

ii) without dimensionality reduction – data is followed by supervised classification using the listed classifiers.

iii) Provide the plots for overall training accuracy, and overall classification accuracy vs. the training size for all methods (classification schemes). Tabulate the classwise classification accuracies (i.e. extension of the sensitivity and specificity values) only for 30% training size over all methods for each dataset for case i) i.e. with dimensionality reduction PCA and LDA for Indian pines dataset only.

(30 pts: Total 4 plots, i.e. 2 plots for each dataset [case i and ii, overall training, and testing accuracy] + 2 tables (classwise accuracy), i.e, 1 table per (PCA, LDA) = 30 pts total, i.e. 5 pts per plots/table)

*Note only provide data visualizations here – label figures

Provide appropriate legends in all figures to denote the methods. For figures follow the below nomenclature: Figure 1: what it does the figure denote and for which dataset results and label all figures in the homework with a caption. (Eg. Figure 1. Classification accuracy with/without dimensionality reduction for Iris dataset.)

You can present PCA+ classification, LDA+ classification plots and without dimensionality reduction + classification plots separately.

2b. Discuss the analysis of results from 2a) for each dataset in terms of:

i) Role of dimensionality reduction /feature extraction on data analysis and classification performance based on data separability, classification accuracy, sensitivity and specificity parameters.—which supervised classification method worked best on each dataset either with or without dimensionality reduction casesand why?

(10 points: i.e. 5pts per dataset inferences)

* Use data visualization to draw analysis

Note: Clearly label each section of code and figures. For Indian Pines dataset, read the indianR.mat data given in the homework folder, where X-data, gth- groundtruth labels