CS488 Homework 3 -100 pts

Due: November, 8 (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 pre-clustering analysis using Elbow method on the Iris and Indian Pines dataset to determine the optimum number of 'k' clusters present in the data. Display the Elbow method plot for each dataset.

i) Iris dataset: select the range k={2,3,…6}
ii) Indian pines dataset: select the range k={2,3,…16} (20 points)

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

i) What was your choice of 'k' clusters present in each dataset and why? -provide reasoning in terms of Elbow method inferences. (10 points)

2. Write a python program to perform unsupervised learning on the Iris and Indian Pines datasets using K-Means, and Hierarchical clustering (with Euclidean, and Cosine distances) using 'k' clusters selected from 1a. for each of the below cases:

i) with dimensionality reduction – Reduce data to 2 dimensions using each of the dimensionality reduction methods (PCA, LDA) followed by unsupervised learning using the listed clustering methods.
ii) without dimensionality reduction – data is followed by unsupervised learning using the listed clustering methods.
iii) Provide the 2D data visualization plots for all clustering methods for each dataset. (30 pts)

3a. Write a python program to compute the cluster validity indices (Davies Bouldin and Silhouette index) and output the cluster validity indices for each dataset:

i) Iris dataset: output cluster validity indices as single table with columns = clusters k={2,3,…6} and rows = cluster validation indices.
ii) Indian pines dataset: output cluster validity indices as single table with columns = clusters k={2,3,…16} and rows = cluster validation indices..
iii) What were the number of clusters validated by each of the indices for each dataset and why ?- explain in terms of what each index conveys and why certain 'k' cluster values were chosen. (30 points)

3b. Discuss the analysis of results from 1, 2 and 3a for each dataset in terms of:

i) Role of dimensionality reduction on data separability, clustering performance and cluster validity.

ii) which unsupervised learning method worked best on each dataset- with or without dimensionality reduction and why? (10 points)

Note: Clearly label each section of code and figures. For figures follow the below nomenclature: Figure 1: Description of what it does and for which dataset. For Indian Pines dataset, read the indianR.mat data given in the homework folder, where X-data, gth-groundtruth labels.

In Indian pines data the zero's are the background class and have to be dropped. You have to extract the class samples. I have shown it in the video. You can find the hyperspectral dataset description at the link here (from week 1 lecture): http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (Links to an external site.)

Remove all the data description outputs and other irrelevant outputs from the report before you submit. You may base your analysis on the extensive outputs but they need not be part of your report.