भारतीय प्रौद्योगिकी संस्थान हैदराबाद
कैंडी - ५०२ २८५, सांगा रेड्डी, तेलंगाण, भारत
फोन : (०४०) २३०१ ६०३३; फेक्स : (०४०) ६००३ / ३२

Indian Institute of Technology Hyderabad
Kandi – 502 285, Sangareddy, Telangana, INDIA
Phone: (040) 2301 6033; Fax: (040) 2301 6003 /32

Assignment on Clustering

Instructor: Dr. Manish Singh TA: Rohan Banerjee

Section A: Know Your Data

- Load the original digits dataset (digits_data).
- Obtain a 2-d representation using PCA and t-SNE (pdigits_data, tdigits_data).
- Visualized the two reduced datasets.

EX 1: Which of the two reduced representations is better for clustering? Explain.

Section B: Data Pre-processing and Cluster Evaluation

- Compare the clusters detected by K-Means (k=10) for the distorted tdigits_data with and without scaling.
- Perform cluster evaluation for the above two clusters using unsupervised and supervised cluster evaluation metrics.

EX2: Which one gives better clustering? How is scaling helpful in this context?

Section C: Overview of Clustering Methods

- Run the clustering algorithms: A, B, C and D on the tdigits_data using the given hyper-parameters.

EX3: Rank the four algorithms in terms of unsupervised cluster evaluation.

EX4: Which supervised evaluation metric seems best to you? Explain.

AEX1: Repeat EX3 and EX4 with pdigits_data. How the clusters compare to the tdigits_data.

Section D: Find Optimal Clustering Parameters

EX5: Find the optimal value of K for K-Means that has highest SC.

EX6: Find the optimal value of EPS and minPTS for the DBSCAN algorithm using SC.

Ex7: Which linkage - single, complete, ward or complete gives the highest SC for agglomerative clustering.

AEX2: Take the optimal values you from EX6, and then run the DBSCAN algorithm on the pdigits_data. How the

clusters compare with the clusters obtained in EX6.

Section E: Application of Clustering

- Find the top-5 most similar digit pairs using the tdigits_data.
- Create 10 clusters using K-Means and Spectral clustering. Label the clusters using the most frequently occurring digit in the cluster. Use these cluster labels to find the top-5 most similar digit pairs.

EX8: Between K-Means and Spectral, which algorithm gives the more similar pairs compared to the ground-truth tdigits_data?