Big Data Project 1

PCA for Iris data (30 points)

- Do PCA, and sparse PCA, t-SNE, UMAP, locally linear embedding (LLE) visualizations for Iris data under different normalization methods
- Which combinations will get the best visualization performance? why
- Calculate the locality conservation ratio for each dimension reduction method (PCA, sparse PCA, t-SNE, UMAP, LLE).

$$\eta = \frac{1}{n}\sum_{i=1}^{n} |A_i \cap B_i|/10$$

A_i is the set with 10 nearest neighbors of the i^{th} observation

B_i is the set with 10 nearest neighbors of the embedding point of the i^{th} observation under a dimension reductiuon method

What's your conclusion, why?

Drug discovery (20 points)

- GDSC_IC50.csv is a dataset about drug sensitivity. It has 555 cells across 98 drugs, where each row represents an observation (cell) , and each column represents a drug.
  - Preprocess data by removing possible missing data
  - Label cells into binary types (you can determine your labeling methods)
  - Use PCA to visualize data with the binary labels and find the outliers
  - Use t-SNE, UMAP, LLE to visualize data data with the labels, apply DBSCAN to the embeddings of t-SNE/UMAP/LLE and find their clustering information. What can you find?

Big patent data analysis (50 points)

- PatentData.csv is a big dataset with high-value (1) and low-value (0) patents.
- do PCA visualization for this dataset and find special patents. What are they?
- do t-SNE/UMAP for this dataset and then do Kmeans clustering and calculate the accuracy and other clustering indices for the embedding data
- do PHATE[1] visualization for this dataset and then do Kmeans clustering and calculate the accuracy and other clustering indices for the embedding data
- do t-SNE visualization for the new data in the PCA space with 90% explained variance ratio., what can you find? Can you use the t-SNE data do a 80%-20% partition for SVM and check all classification measures? What can you find?

[1] You need to check paper in this project

What should you turn in?

- 1. A folder that contains
  - A ppt to show details of your analytics (at LEAST 40 pages)
  - your data
  - source files
  - corresponding related output.
  - A link to your presentation video
- 2. Send the zipped file (.zip instead of ,rar)