Artificial Intelligence in Fintech Quiz (3)[1]

[1] Choose at least three problems

A) Analyze SP500 stock data [2] (30 points)

- USE PCA, SPCA, nSVA, and t-SNE rank the SP500 stock data
    - Visualize the section information of the SP500 data
    - Find 20 top-ranked stocks and find the importance scores for variables. What can you find? Why?
    - Check sections 2 and 3 information of the SEC 8K of the top-10 ranked stocks what can you find?
- Compare PCA biplot and PCA triplot and t-SNE plot for this dataset

[2] You need to go through our class codes

B) HFT feature interpolations (30 points)

- Add at least 4 variables to the HFT datasets in your quiz 1 so that one variable should be Bollinger Bands
- Partition transactions as up down in each interval and visualize your data by t-SNE, SPCA, and PCA
- Use PCA to rank the importance of features
- Rank the importance of the observations and mark your ranking in the price plots
- Use other distances rather than Euclidean distance in t-SNE to visualize HFT data, what can you find? (extra credits 20 points)

C) Vehicle data analysis (30 points)

- Write a python program to conduct the same analysis as we did in R for the vehicle data in our lecture
- You need to use ALL variables rather than only 11 variables we used.
  - PCA biplot
  - PCA triplot
  - t-SNE biplot
  - Rank top 20 outliers using PCA, nSVA, and t-SNE and explain their differences.

D) Locality preservation analysis (50 points)

Let $d(x_i, x_j)$ be the pairwise Euclidean distance between x_i and x_j input data and $d(y_i, y_j)$ be the corresponding pairwise distance between y_i and y_j that are corresponding embedding points of x_i and x_j using a dimension reduction method (e.g., PCA).

- What are the relationships between $d(x_{i,} x_j)$ and $d(y_i, y_j)$ under PCA (with or without dimension reduction) and t-SNE?
- Compare the 10-nearest neighbors ('10-nn-x') for each point of data in the input space and the corresponding 10-nearest neighbor ('10-nn-y') in the embedding space of PCA and t-SNE? How many percents of them will repeat themselves? How about their order information?
- Please use at least three datasets that including Option data, HFT data and Vehicle data

What should you turn in?

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