Artificial Intelligence in Fintech Quiz (4)

Credit Risk Analytics by SVM (50 points)

- We have the following data set for credit ranking for 12 different industry sections (it is a simulated data):
  - *credit_sim_data.csv*, where the first **1540** samples (rows) are labeled as 'good credit' (label type: '1'), i.e., whose credit rankings are 'AAA', 'AA', or 'A'
  - and the remaining **130** samples are labeled as ' bad credit', (label type: '0') whose credit ranks are 'CCC'.
- There are six variables (columns) in this data set:

```
variable 1: Working capital / Total Assets (WC_TA)
variable 2: Retained Earnings / Total Assets (RE_TA)
variable 3: Earnings Before Interests and Taxes / Total Assets (EBIT_TA)
variable 4: Market Value of Equity / Book Value of Total Debt (MVE_BVTD)
variable 5: Sales / Total Assets (S_TA)
variable 6: Industry sector labels from 1-12
```

- Complete the following problems
  - Visualize this dataset by c-NE, t-SNE, UMAP, and PCA
  - Conduct k-fold (k=10) cross validation for the data and use the following prediction to conduct classifications and compare their results
    - SVM with 'linear', 'rbf', 'poly', 'cosine', and 'sigmoid' and a kernel you design respectively
    - Compare the support vectors under different kernels.
    - Compare the top 100 eigenvalues of kernel matrices under different kernels
  - Visualize support vectors in 2D by combining PCA in SVM.
  - All classification measures including d-index should be used in evaluating your learning results.

Extend d-index to multiclass data and apply it to Iris data classification (50 points)

- Prove the range of d-index in $k - class$ classification is $[2log_2(\frac{k+1}{k}), 2]$ if there is no underfitting
- The d-index has the range: $[log_2(\frac{3(\gamma+1)}{2}) - \varepsilon, 2]$ when the input training dataset has a majority ratio γ under a ML model Θ, where ε is a small enough ratio (e.g., 0.01).
- Make iris data to an imbalanced version: 50, 20, 50 and do d-index based parameter tuning using the grid search

What should you turn in?

- 1. A folder that contains
  - A report to show details of your analytics (at least 40 pages)
  - your data
  - source files
  - corresponding related output.
- 2. Submt your zipped folder to BB