CISC5352 Final Project: Examine AI tools in Fintech [1]

[1] Each group should do at least two projects

1 Credit Risk worthiness evaluation (200 points)

- Credit risk analytics is key in personal loan decision making for banks. Using credit risk analytics, banks are able to analyze previous lending data, along with associated default rates, to create an effective predictive model in loan decision making.
- Two credit risk datasets: credit_risk_small_data_0.02.csv and credit_sim_data.csv both are imbalanced data. Why the former one fails almost all machine learning methods, but the latter has a very good results?
- Can you find a method to detect their built-in difference even before running a learning machine?
- Compare the performance of deep learning models for two datasets
  - Run an LSTM model for the two datasets and explain their performance (extra credits)
  - Design a multilayer perceptron (MLP) with at least 2 hidden layers (DNN) to predict your test data ,and calculate at least six classification metrics including diagnostic index
    1. Compare three training methods of MLP: *adam*, *SGD* and lbfgs.
    2. Turn DNN by gridSearch using d-index and check its performance
- Credit data in credit_risk_small_data_0.02.csv has good separations under t-SNE /UMAP (you may need to use minmax normalization)
  - Partition data as 80% training and 20% test.
  - Do t-SNE/UMAP for the training data.
  - Project the test data to the t-SNE/UMAP subspaces so that can you do credit risk prediction in the subspace
  - (Note:
    * you may need to think hard how to project the test data to the t-sne/umap subspace

- You may also think hard how to predict test data's label (you can also use unsupervised learning approach)

Hierarchical learning in P2P leading

- Read paper: *Credit risk evaluation in peer-to-peer lending with linguistic data transformation and supervised learning*
- Clean their dataset: Loan.csv so that it will be ready for credit risk prediction (You may need to remove some features)
- Apply hierarchical learning to the cl;eaned dataset and compare its performance with that in paper
  - Note: Pay more attendtion to your learnning machine selection in hierarchical learning

2. Implied volatility Pricing (200 points)[2]

1. Determine European options from Option2017_2_Clean.csv by using Brent or bisection method and write all European options as EuropeanOption2017.csv (I already did this step for you. If you need more accuracy, you can redo it by yourself)
2. Apply the following methods to estimate the implied volatility for the put and call option data you get by using tolerance 10^{-12} in iteration and interest rate 0.03.
   (a) the classic Bisection method (*you need to code it,* but you can refer to the bisection method in scipy)
   (b) Brent method (you don't need to code it, you can only use the implementation from scipy)
   (c) Muller-Bisection method (*you need to code it*)
   (d) Newton method (*you need to code it*, but you can refer to the Newton method in scipy)
   (e) New newton: use brent as the fill-in method to compute initial points
   (f) New Halley: use brent as the fill-in method to compute initial points for Halley
   (g) Halley's irrational formula (Note: you can only pick plus sign in your implementation)

   You need to use

$$vomma = \frac{\partial^2 f}{\partial \sigma^2} = vega \times \frac{d_1 d_2}{\sigma}$$

$$d_1 = \frac{\ln(S/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$$

[2] QF students MUST do this one)

$$d_2 = d_1 - \sigma\sqrt{T}$$

Note: σ is actually unknown, which is just the x_n in your iteration scheme.

3. Evaluate your implied volatility calculation by using the following measure for each method:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\sigma_{imp,i} - \sigma^*_{imp,i})^2$$

Where $\sigma_{imp,i}$ is the implied volatility you calculate for the i^{th} option and $\sigma^*_{imp,i}$ is the true implied volatility.

- You need to compare MSE plots for each method
- You also need to calculate the average iteration number to converge for each method

4. Compare the efficiency of the methods in implied volatility pricing.

- Efficiency aims to answer the query: how efficient the method Θ can be when it is convergent? It is defined as the following ratio for a given method Θ,

$$\eta = \frac{1}{(1 + mse)\log_2(1 + E(t))} \tag{1}$$

where mse represents the mean square error (MSE) of the method for all convergent cases and $E(t)$ is its corresponding time expectation (average iteration number to converge).

- A few average iteration steps and small MSE values under convergence both contribute to a good efficiency. The closer the efficiency to 1, the more efficient the method. In fact, $\eta = 1$ means the method achieves 100% efficiency, the ideal state of pricing. That is, $mse = 0$ and the method only takes one iteration step averagely to reach the true implied volatility value.

5. Draw your conclusion about accuracy and time for the four methods in implied volatility prediction.

- 6. Extra credits (50 points): Find America options in Option2017_2_Clean.csv by using Barone-Adesi and Whaley Option Pricing Model and write them in a file called AmericaOption2017.csv. Apply selective learning to this dataset and draw your conclusion

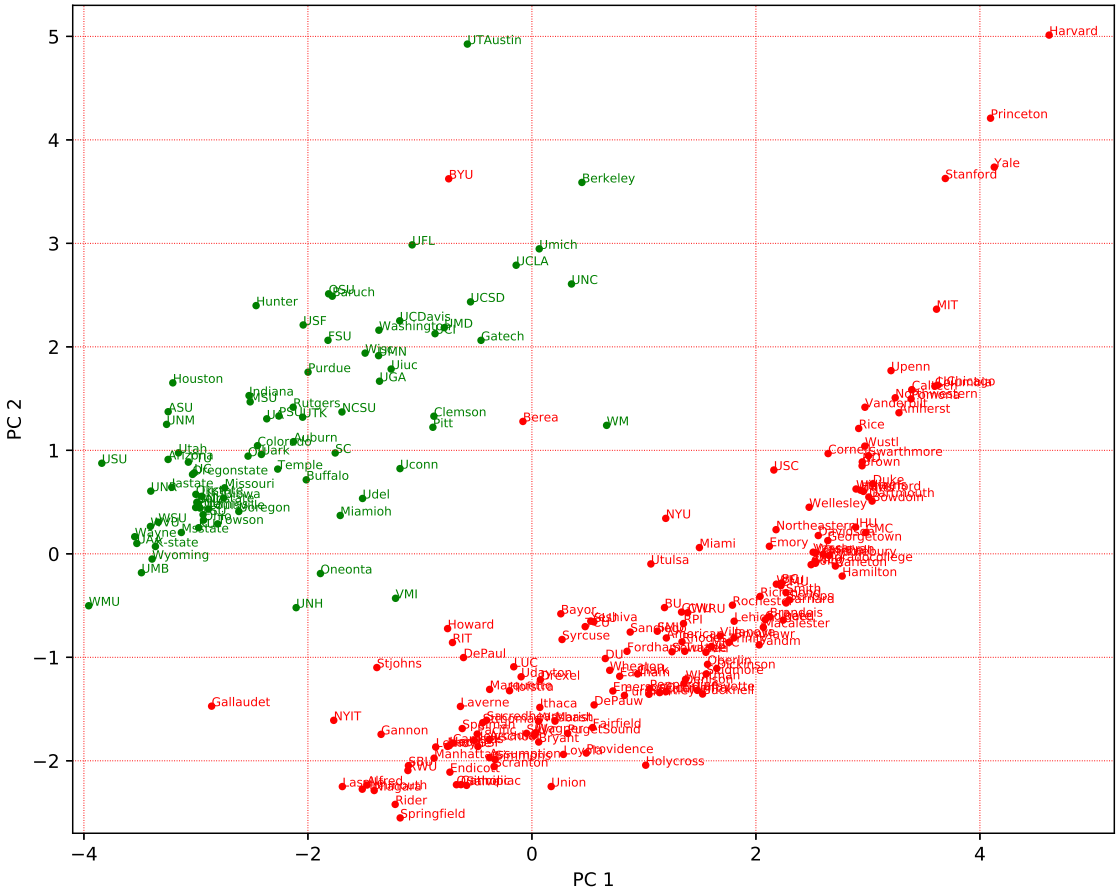
3) Endowment analytics (200 points)

Basic data information

- The dataset in the file: *endowments_2018_with_init.csv* consists of a total of 232 institutions and their initials
- The dataset in the file: *endowments_2018_with_twitter_info.csv* consists of a total of 232 institutions and their twitter information
  - It includes the top-100 richest universities and 132 institutions with at least $50 million in endowments.
- There are the core nine variables
  - high school GPA, average SAT scores, and graduation and acceptance ratios (*academic characteristic variables)*
  - public/private, enrollments (undergraduate enrollment): *(Institutional characteristic variables)*
  - endowment, tuition, and loan rate (*Financial characteristic variables)*
- Money used is the label for these universities: 0: poorly used, 1: fairly used, 2 well-used.

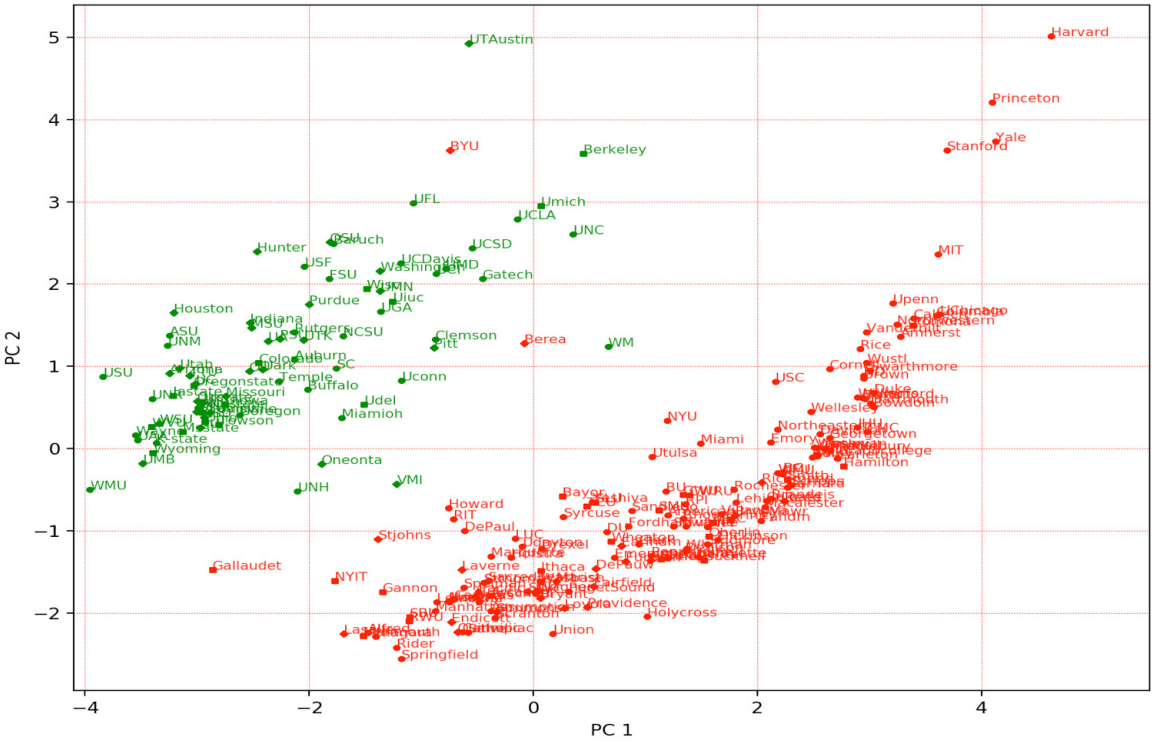
Dimension reduction analysis

- Do PCA, SPCA, t-SNE, LLE analysis for two datasets to generate biplots as follows and interpret its meaning and outliers (Note: you don't need to use the label info when doing such analysis; The twitter names or inits are good for you to label data)



- Write an algorithm to find the 5 nearest neighbors (peers) in the PCA/SPCA/t-SNE/LLE space for each institution by using at least two components)
  - Show the 5 neighbors (peers) of Fordham University, Michigan state, Yale, Smith college)
  - If there is no 5 neighbors available, your algorithm should give an reason.
- Visualize the two datases data similar to the following plot (label data via their money usage) via PCA, SPCA, t-SNE, LLE analysis for two datasets

  - Diamond: poorly used; square: fairly used; dot: well-used
  - Green: public school, red: private school



- Apply hiearchical learning to the endowment data in *endowments_ 2018_ with_ init.csv* by doing 70% and 30% train/test partition (you need to write your training and test daat as train.csv and test.csv)
  - Calculate the classification measures and d-index (Note you need to write a routine to calculate the multi-class d-index)
- Apply hiearchical learning to the new test data: newData.csv and Calculate the classification measures and d-index
- Draw your conclusion

What should you turn in?

- 1. A folder that contains
  - A ppt to show details of your analytics (at LEAST 100 pages)
  - your data
  - source files
  - corresponding related output.
  - A link to your presentation video
- 2. Send the zipped file (.zip instead of ,rar) of your folder to Blackboard before the DUE