CISC 5352 Financial Data Analytics Project 2: [1]

[1] Each group must at least one project.

An ensemble system for implied volatility (100 points)

- Using at least k-NN, SVM, Random Forest (RF), and Gradient Boost(GB) to build an ensemble learning system to predict implied volatility, in which k-NN, SVM, RF and GB are treated as 'weak' learners[2]. The goal of this ensemble learning system is to provide better prediction results.

- You need to at least build the following four ensemble learning systems such that, given a test sample x, its prediction function is defined as

  - $\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}_i(x)$
  - $\hat{f}(x) = median\{\hat{f}_1(x), \hat{f}_2(x) \cdots \hat{f}_N(x)\}$
  - $\hat{f}(x) = \sum_{i=1}^{N} w_i \hat{f}_i(x)$, $w_i = 1 - \frac{mse_i}{\sum_{i=1}^{N} mse_i}$
  - $\hat{f}(x) = \sum_{i=1}^{N} w_i \hat{f}_i(x)$, $w_i = \frac{mse_i}{\sum_{i=1}^{N} mse_i}$
  -
$$mse_i = \frac{1}{n}\sum_{k=1}^{n} |f(x_k) - \hat{f}_i(x_k)|^2$$
  - $f(x)$ represents the true prediction function for an input variable x, which is an option contract in our context, and $f(x)$ is a true implied volatility for the option contract x.
  - $\hat{f}_i(x)$ is the prediction function estimated by i^{th} weak learner. For example, $\hat{f}_1(x), \hat{f}_2(x), \hat{f}_3(x), \hat{f}_4(x)$ represents the prediction functions estimated by learning machines k-NN, SVM, RF and GB respectively.
  - $\hat{f}(x)$ represents the prediction function estimated by the ensemble system

[2]

- Our ensemble learning system is actually a 'strong' learner based ensemble system.

  - $mse_i = \frac{1}{n}\sum_{k=1}^{n} |f(x_k) - \hat{f}_i(x_k)|^2$ is the MSE for the i^{th} learning machine for total n test samples. It is actually same as the previous definition I gave in project 1: $MSE = \frac{1}{n}\sum_{k=1}^{n} |predictedIV_i - IV_i|^2$

- **Data sets**
  - You are required to use previous two datasets in project 1. For the consistency, please still use 80% data for training and 20% for test.

- **Comparison peers (Each group at least implements one of them)**
  - Implement a neural net based model to predict implied volatility
  - implement a logistic regression model to predict implied volatility

- Complete the following assignments to evaluate the ensemble systems for implied volatility prediction, before draw your conclusion.
  - Compare the MSE values of each ensemble system, each learner, and their comparison peers.
  - Compare the sample deviation values for the ensemble systems, learners and comparison peers (you need to draw them in a same plot).
  - Given test sample x_k , the sample deviation δ_{ik} is defined with respect to i^{th} learner as
    - ∗ $\delta_{ik} = |f(x_k) - \hat{f}_i(x_k)|$

- **Extra credits (100 points):**
  - **Collect more data to use your** OptionDataWebGleaner to get a large dataset and clean it.
  - Use at least 8,000 samples as training and 2000 samples as test and use spark to do previous machine learning assignments

B. Examine richest universities (100 points)[3]

- Go to the link and download and process the richest universities 's data
  - https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_the_United_States_by_endowment
  - Note: you need to remove possible outliers (e.g. University of Illinois system endowment is hard to count as one IL school's endowment)
- Also download and process US news ranking data
  - http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities
  - http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-liberal-arts-colleges
- Rank each rich university to three types. We assume each university would use their money to increase their ranking.
  - 1. *Money-well-used-for-ranking*: it means the university/college achieves at least top 30 percentile ranking among their peers (e.g. Harvard University belongs to this type because it is ranked No.2 in national universities)
  - 2. *Money-fairly-used-for-ranking*: it means the university/college achieves its rank between top 31 percentile to 70 percentile among their peers
  - 3. *Money-poorly-used-for-ranking*:it means the university/college achieves its rank lower than 70 percentile among their peers
- Build your data based on information you get and write it in a csv file.
  - Do your feature engineering!
    - ∗ You can add more information you think necessary such as, tuition & fee, acceptance ratio, enrollment, public/private type, location (rural (0), city (1), suburb(2))

[3] Use python's Scrapy package or other web-crawling toolx can speedup your data collection and process a lot!

- Employ kNN,SVM, GB, and RF to classify your data under 5-fold cross-validation and compare their performance.
- **Extra credits (50 points):**
  - Investigate the professors' salary level in these richest universities: are they corralated with their universities' endowment?
  - You need to crawl `http://faculty-salaries.startclass.com/` to get faculty salary data.

What should you turn in?

- 1. A folder that contains
  - A ppt to show details of your analytics (at least 30 pages)
  - your data
  - source files
  - corresponding related output.
- 2. Please name your folder last_name1_last-name2_CISC5352_project_2. For example, Brown_Smith_CISC5352_project_1 if your group members with last names: Brown and Smith.
- 3. Send the zipped file (.zip instead of ,rar) of your folder to Blackboard before 11:59 pm Dec 06, 2016