BERNARD BARUCH COLLEGE - CUNY

MTH 9898 BIG DATA IN FINANCE

1UWA - Spring 2019

- Lectures: Wednesday 6:05PM-21:00PM in B9-140 VC
- Instructor: Giulio Trigila
- Office: B6-286 VC
- Office hours: Monday 5-7pm
- Email: giulio.trigila@baruch.cuny.edu [1]

1. Course Description

This course will focus on specific Machine Learning algorithms that are well suited for the analysis of large quantities of structured and unstructured financial data.

We start the course with a brief review of supervised and unsupervised learning underlying the main challenges in data science: the curse of dimensionality, model selection and algorithmic scalability. We then move to more specific techniques used in the world of finance related to big data like Stochastic gradient descent and dimensionality reduction.

Along the course we will make use of two very popular machine learning libraries written in Python: Scikit-learn and Pandas. Practice in the use of these two libraries will serve as an introduction to the Apache Spark library MLlib specifically developed for the analysis of large datasets.

2. List of topics

(1) Basics of Machine Learning:
   - Supervised and Unsupervised Learning
   - Regression, Clustering and Classification
   - Bias-Variance trade-off
   - Model selection and Cross validation
(2) Stochastic gradient descent
(3) Decision tree, Random forest, Gradient boosted tree
(4) Clustering
   - Expectation maximization
   - K-means
   - Gaussian mixture models
(5) Dimensionality reduction

Date: December 15, 2018.

- PCA, Incremental PCA
- SVD

(6) Large covariance matrix estimation
(7) Large scale hypothesis testing and False Discovery Rates

3. Learning goals

(1) Understand the major challenges related to the analysis of large quantities of data defined in high dimensional spaces.
(2) Be familiar with the Python libraries specifically developed for machine learning and with the theory at the base of the algorithms analyzed in class.
(3) Be familiar with the Apache-Spark cluster computing framework.

4. Suggested texts

- Pattern Recognition and Machine Learning, C. M. Bishop. Springer
- Trevor J. Hastie and Robert John Tibshirani and Jerome H. Friedman, The Elements Of Statistical Learning : Data Mining, Inference, and Prediction, 2nd Ed., Springer Series in Statistics
- Wes McKinney, Python for Data Analysis, 2nd Ed., O'Reilly.

5. Software

(1) Python 3: Pandas, Scikit-learn.
(2) Apache Spark: PySpark, MLlib.

6. Grading

There will be two midterms, weekly homework, and a final project. The class will be split in groups of 4 or 5 people. Each group is assigned a project that will be discussed in class. The project consists in applying the algorithms explained in class (by means of the Python libraries) to analyze real world data sets. Each group will give a short presentation of the project during the last lecture.

7. Class Policy

- Regular attendance is essential and expected.
- All cellphones should be turned off. I always take a 15min break in each class where you can always check your cellphone.

8. Academic Honesty

Please check the Academic Honesty webpage.