Big Data Science in Finance

This course is based on 3 hours per week, 15 week semester.

Course Benefits for Students:

The proposed course is based on a brand-new book "Big Data Science in Finance" by Irene Aldridge and Marco Avellaneda (Wiley, 2021).

The knowledge and the projects are designed to enable the students to learn the principles of data science, gain a deeper understanding of the cutting-edge developments in Finance and acquire practical understanding of the topics.

Reading:

Required Text: Aldridge, Irene and Marco Avellaneda, 2021. Big Data Science in Finance. Hoboken: Wiley & Sons. ISBN: 978-1119602989

Course Syllabus

Session 1: Why Big Data? The changes in market structure, technology and mathematical innovation driving the trend.

Session 2: Neural networks in Finance.

Session 3: Supervised Learning and Applications:

- Ridge Regression, LASSO, Elastic Nets
- K Nearest Neighbors (K-NN)

Session 4: Supervised Learning and Applications, continued:

- Decision Trees, Random Decision Forests and Extra Trees
- Support Vector Machines (SVMs)

Session 5: Semi-Supervised Learning and Applications:

- Performance Evaluation via Cross-Validation

- Generative Models

Session 6: Semi-Supervised Learning, continued:

- Generative Models, continued
- Discriminative Models
- Graph-based Models

Session 7: Letting the Data Speak with Unsupervised Learning

- Dimensionality Reduction in Finance
- Dimensionality Reduction with Unsupervised Learning
- Singular Value Decomposition
- Deconstructing Financial Returns

Session 8: Letting the Data Speak with Unsupervised Learning, continued:

- Singular Vectors as Portfolio Weights
- Principal Component Regression
- Key Big Data Tools: SVD and PCA in Detail

Session 9: Midterm Exam.

Session 10. Big Data Factor Models

- Optimal Factorization
- Eigenportfolios
- Factor Discovery

Session 11: Big Data Factor Models, continued:

- Approximate Factor Models
- Unknown Factors (POET)
- Instrumented PCA
- The Three Pass Model
- Risk-Premium PCA
- Nonlinear Factorization
- Projected PCA

Session 12: Data as a Signal vs. Noise

- Random data show in Eigenvalue distribution
- What's in a data bag?
- The Marcenko-Pastur Theorem
- Spike Model
- Dealing with Highly Correlated Data
- The Karhunen-Loeve Transform
- Data Imputation

- Missing Eigenvalues
- The Tracy-Widom Distribution
- Identifying Missing Streaming Data (the Johnson-Lindenstrauss Lemma)

Session 13: Applications: Unsupervised Learning in Option Pricing and Stochastic Modeling

Session 14: Data Clustering

Session 15: Final Exam