Course Code
Machine Learnings and Algorithms

Course Information

Instructor: Prof. Zhen Zhang
Office: PHBS Building
Phone: 86-755-8801-8753
Email: zhangz@sustech.edu.cn
Office Hour: TBA

Teaching Assistant:
Phone:
Email:

Classes:
Lectures: Day, Time
Venue: PHBS Building, Room

Course Website:
http://www.tomsargent.com

1. Course Description

1.1 Context

Course overview: This course teaches students both foundational and frontier algorithms and models used to analyze economic and financial data. Analytical goals are data description, data reduction, and detection of relationships among variables, and ways to interpret those relationships in terms of underlying economic and social forces. The course teaches how estimated (i.e., "fit") models can be used for prediction, forecasting, and possibly inference about cause and effect.

Prerequisites:
We assume that students are comfortable writing programs in Python using the core data-analytics packages `numpy` and `pandas`. We also assume that students have a solid mathematical background that includes at least one course in calculus, one course in linear algebra, and exposure to probability theory and statistics.

Motivated students without a strong mathematical background can still succeed in this course, provided they are willing to work hard. This course is practical and application oriented. We aim to arm students with understandings of how the methods and algorithms work. Some of these understandings are grounded in the mathematics, but they are substantially strengthened through exploration and practice. To explore data, it always helps to have some theory as a guide.

1.2 Textbooks and Reading Materials

- **(IntroSLA) An Introduction to Statistical Learning with Applications in Python** by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor, Springer, 2023. https://www.statlearning.com/
- **(ElemSML) The Elements of Statistical Machine Learning: Data mining, Inference and Prediction** by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Springer, 2009. https://hastie.su.domains/ElemStatLearn/
- **Statistical Rethinking** by Richard McElreath.
- **(FoundationML) Foundations of Machine Learning** by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, MIT Press, 2018.
- **(DL) Deep Learning** by Ian Goodfellow, Yoshua Bengio and Aaron Courville, MIT Press, 2016. https://www.deeplearningbook.org/
- **(FinML) Financial Machine Learning** by Bryan Kelly and Dacheng Xiu, in: Foundations and Trends in Finance, Vol. 13, No. 3-4, (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4501707

2. Learning Outcomes

2.1 Intended Learning Outcomes

Learning Goals	Objectives	Assessment (YES with details or NO)
1. Our graduates will be effective communicators.	1.1. Our students will produce quality business and research-oriented documents.	NO
	1.2. Students are able to professionally present their ideas and also logically explain and defend their argument.	NO
2. Our graduates will be skilled in team work and leadership.	2.1. Students will be able to lead and participate in group for projects, discussion, and presentation.	YES
	2.2. Students will be able to apply leadership theories and related skills.	NO
3. Our graduates will be trained in ethics.	3.1. In a case setting, students will use appropriate techniques to analyze business problems and identify the ethical aspects, provide a solution and defend it.	NO
	3.2. Our students will practice ethics in the duration of the program.	NO
4. Our graduates will have a global perspective.	4.1. Students will have an international exposure.	YES
5. Our graduates will be skilled in problem-solving and critical thinking.	5.1. Our students will have a good understanding of fundamental theories in their fields.	YES
	5.2. Our students will be prepared to face problems in various business settings and find solutions.	YES
	5.3. Our students will demonstrate competency in critical thinking.	YES

2.2 Course specific objectives

This course provides students with a hands-on introduction to widely used algorithms and models used for understanding social science data. It is expected that students will be able to apply these algorithms to new datasets and problems as well explain (at a high level) how the algorithms arrive at the answers they produce.

Successful students will be able to do the following after completing this course

- Apply *exploratory data analysis* (EDA) techniques to understand a new dataset
- Communicate with subject matter experts about important relationships among variables in a dataset and key aspects of the data collection process

- Determine what class of algorithms should be applied to new problems
- Implement cross-validation procedures for model evaluation and selection
- Visualize and interpret model outputs with the aim of understanding how they relate to model inputs
- Understand the concept of overfitting and how to alter model structure or apply regularization techniques to control its effects
- Gain insight into various model evaluation techniques applicable to different models
- Utilize high-performance libraries such as `pandas` and `scikit-learn` to implement machine learning models
- Master the mathematical foundations of machine learning algorithms
- Gain exposure to deep learning techniques and the use of large language model

2.3 Assessment/Grading Details

Assessments

This course will use a mixture of homework assignments, in-class quizzes, a project, and a final exam to evaluate students.

Homework: Homework will be assigned almost every two week. There are totally 4 homework assignments.

In-class quizzes: There will be 4 in-class quizzes during the semester/module. These quizzes will be open book and in some sense play the role of check-in.

Project: There will be a class project aimed at helping you apply the tools that you have learned to a "real-world problem." Typically this can be done in teamwork.

Exams: There will be 1 final exam whose difficulty is on the average of homework and quizzes.

Other than for quizzes and exams, we highly encourage students to work together. We have found that groups of 3-4 seem to work best. We believe that collaborative work is the best way to learn the type of material that we cover. We advise students not to rely on others to do work that you do not understand.

Grading Policy

The assignments just described will be the main inputs to the grade for the course. Assignments will be weighted evenly within groups and overall according to the following decision rule:

- Homework assignments: 30%
- In-class quizzes: 15%
- Project: 30%
- Exam: 25%

This weighting reflects our opinion that the most important skills to be acquired in this class are communicated by one's ability successfully to apply the tools that you learn to an interesting question in Economics and Finance.

2.4 Academic Honesty and Plagiarism

It is important for a student's effort and credit to be recognized through class assessment. Credits earned for a student work due to efforts done by others are clearly unfair. Deliberate dishonesty is considered academic misconducts, which include plagiarism; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; or altering, forging, or misusing a University academic record; or fabricating or falsifying of data, research procedures, or data analysis.

All assessments are subject to academic misconduct check. Misconduct check may include reproducing the assessment, providing a copy to another member of faculty, and/or communicate a copy of this assignment to the PHBS Discipline Committee. A suspected plagiarized document/assignment submitted to a plagiarism checking service may be kept in its database for future reference purpose.

Where violation is suspected, penalties will be implemented. The penalties for academic misconduct may include: deduction of honour points, a mark of zero on the assessment, a fail grade for the whole course, and reference of the matter to the Peking University Registrar.

For more information of plagiarism, please refer to *PHBS Student Handbook*.

3. Topics, Teaching and Assessment Schedule

The schedule is tentative and subject to change so that we can adapt material to incorporate new developments in the fast-moving fields of AI and machine learning. Several of the modules below will occupy more than one week. The learning goals target key concepts to be mastered after each module. Later modules build on earlier ones.

Week 1 Introduction to Machine Learning
Sources and tools:

- Class notes
- `https://datascience.quantecon.org/applications/ml_in_economics.html`

Topics to be mastered:

- Definition of Machine Learning (ML)
- Relation to econometrics
- Supervised learning
- Unsupervised learning
- Mathematical and statistical foundations

Mathematical Preliminary (self-learning)

Sources and tools:

- Class notes
- Online notes (to be provided)

Topics to be mastered:

- Linear algebra
- Numerical methods
- Probability
- Statistics and information theory

Week 2 Data Preprocessing
Sources and tools:

- Class notes
- `https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing`

Topics to be mastered:

- Data statistics
- Metrics and distances
- Standardization and discretization
- Missing value treatment
- Outlier detection

Python programming (self-learning)
Sources and tools:

- Class notes
- `https://scikit-learn.org/`

Topics to be mastered:

- Syntax
- Basic packages: `numpy, scipy, matplotlib`
- ML packages: `pandas, seaborn, scikit-learn, pytorch`

Week 3 Linear Models, I
Sources and tools:

- Class notes
- Chapters 3 and 4 of the book IntroSLA (or ElemSML)
- `https://python.quantecon.org/ols.html`
- `https://www.statsmodels.org/stable/index.html`
- `https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model`

Topics to be mastered:

- Linear regression for regression tasks
- Logistic regression for classification tasks
- Regression with `statsmodels`
- Regression with `scikit-learn`

Linear Models, II
Sources and tools:

- Class notes
- Chapters 3 and 4 of the book IntroSLA (or ElemSML)
- https://datascience.quantecon.org/applications/classification.html
- https://datascience.quantecon.org/applications/regression.html

Topics to be mastered:

- Loss functions
- Maximum Likelihood Estimation
- Goodness of fit (evaluation metrics)
- Dealing with imbalanced datasets
- Handling categorical, sparse, temporal, or discrete data

Week 4 Overfitting and model selection, I
Sources and tools:

- Class notes
- https://scikit-learn.org/stable/model_selection.html
- Chapter 6 of IntroSLA (or Chapter 7 of ElemSML)
- Chapter 5 of DL

Topics to be mastered:

- Diagnose overfitting
- Bias-variance tradeoff
- VC dimension
- Regularization via L1 and/or L2 penalty (Lasso, Ridge, ElasticNet regression)
- Bayesian interpretation of Lasso and Ridge regression

Overfitting and model selection, II
Sources and tools:

- Class notes
- https://scikit-learn.org/stable/model_selection.html
- Chapters 5 and 8 of IntroSLA (or Chapters 8 and 16 of ElemSML)
- Chapter 5 of DL

Topics to be mastered:

- Cross validation procedure, grid search
- Cross validation for time series
- Model ensembles, bagging and boosting
- Additive models, gradient descent, AdaBoost, GBDT
- Hyperparameter tuning

Week 5
Non-linear models
Sources and tools:

- Class notes
- https://python.quantecon.org/mle.html

- `https://scikit-learn.org/stable/`
- Chapters 8 and 9 of IntroSLA
- Chapters 9, 12, 13 and 15 of ElemSML

Topics to be mastered:

- Decision trees
- Random forest
- Naïve Bayes
- k-nearest neighbors
- Support vector machines
- Non-linear models in `scikit-learn`

Week 6 Unsupervised Learning: Clustering
Sources and tools:

- Class notes
- `https://scikit-learn.org/stable/unsupervised_learning.html`
- Chapter 12 of IntroSLA (or Chapter 14 of ElemSML)

Topics to be mastered:

- K-Means
- Hierarchical clustering
- DBSCAN
- Expectation-Maximization (EM)
- Spectral clustering

Week 7 Unsupervised Learning: Dimensionality reduction
Sources and tools:

- Class notes
- `https://scikit-learn.org/stable/unsupervised_learning.html`
- `https://datascience.quantecon.org/applications/working_with_text.html`
- Chapter 12 of IntroSLA (or Chapter 14 of ElemSML)

Topics to be mastered:

- Principal Component Analysis
- Nonlinear methods: kernel trick
- Manifold learning
- t-SNE

Week 8 Deep learning, I
Sources and tools:

- Class notes
- `https://pytorch.org/`
- Chapter 10 of IntroSLA
- Chapter 6 and 7 of Deep Learning

Topics to be mastered:

- Manifold hypothesis
- Multilayer perceptron as nested linear regression
- Feedforward neural network and functional composition
- Backpropagation and (stochastic) gradient descent
- Regularization for Deep Learning

Deep learning, II
Sources and tools:

- Class notes
- `https://pytorch.org/`
- Chapter 10 of IntroSLA
- Chapter 9 of Deep Learning

Topics to be mastered:

- Convolutional neural networks (CNN)
- Convolution and pooling
- Neuroscientific basis for convolutional networks
- Residual neural network (ResNet)

4. Miscellaneous

Professional Behaviour

Attend class. They say “eighty percent of success is just showing up.” We have found that those who show up perform systematically better.

Arrive to class on time and stay until the end of class. Chronically arriving late or leaving class early is unprofessional and disruptive to the rest of the class.

We understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course.