Course Syllabus

ISYE 7406
Data Mining and Statistical Learning

Spring 2025

Professor: Dr. Xiaoming Huo

Course Description

An introduction to some commonly used data mining and statistical learning algorithms such as the K-nearest neighbor (KNN) algorithm, linear methods for regression and classification, tree-based methods, ensemble methods, support vector machine, neural networks, and K-means clustering algorithm. This course emphasizes understanding the theoretical and statistical aspects of the above-mentioned data mining algorithms as well as the implementation of these algorithms with data examples using statistical software (e.g., R, Python, etc.)

Prerequisite

- None. ISYE 6414 (regression analysis) or similar is a recommendation from the department. Still, it is not a requirement (as long as a student is willing to spend extra time and effort to pick up what they need in our course on some very basics of regression).

Course Goals

At the end of this course, you will learn to:

- Understand the theoretical and statistical aspects of several widely used data mining or statistical learning methods;
- Apply appropriate data mining and statistical learning methods to analyze real-world datasets

Grading Policy

The course grade is based on the following:

- Active Participation (5%) for course participation and peer grading 8 times in the semester and each time with 3 assignments of fellow classmates. It consists of 5 homework assignments and 3 project products (proposal, presentation, and final written report), including writing constructive peer grading comments and finishing peer assessment timely, i.e., within 5 days after peer grading assignments. This allows you to not only learn from other students but also be ready to provide invaluable critical comments. You might lose points if you are unprofessional during peer grading (e.g., write a very brief comment about the excellent job while it is clearly not).
- Quizzes (20%)

    - Total of 4 quizzes, and 5 points per quiz: 4*5 = 20 points.
    - You have 40 minutes to complete each quiz.
    - The quiz questions are closely related to those in weekly knowledge checks since the previous quiz. Note that you can only take each quiz once, but you can take an unlimited number of attempts on the weekly knowledge checks.
    - There is no make-up or extension for the quizzes.
- Homework (25%)
    - Total of 5 homework assignments, and 5 points per homework: 5*5 = 25 points
    - All students are also expected to conduct peer review assessment
        - You are expected to write constructive comments to three of your classmates so that students can learn from each other.
        - Rubric/solution sets will be provided for the peer review
    - The TAs will assign all grades based on their own assessment. In particular, the TAs might or might not agree with peer reviews and have the right to discard the peer review grades if needed. Thus, there is no need to complain if you disagree with the peer review comments/grades.
    - If you disagree with the TA's grades or comments, you can leave a private piazza message, and the lead TAs will review your case and regrade. If you still disagree with the lead TA's decision, the instructor will make a final decision.
    - Due to the class size, we usually don't grant extensions unless the requestors can demonstrate that something beyond their control (such as illness, emergency, etc.) occurred. We would request the requestors to show adequate documents to convince us. Due to the number of possible circumstances, we don't give specific instructions on which type of documentation we would ask for. The requestors can use common sense when choosing documentation. The decision from the instructor and one of the head TA is considered final.
    - A 10% penalty will be applied every 12 hours for unexcused late submissions. That is, if a submission is between 0 and 12 hours late, we automatically deduct 0.5 out of 5 points; if a submission is between 12 and 24 hours late, we automatically deduct 1 out of 5 points, and so on.
- Group Project (25%)
    - The detailed guidelines for the group project will be released during week #6
    - Group sizes consist of 1 – 5 students, sign up for one of the pre-created project groups via the "people" tab at Canvas
    - Project Proposal – 3 points, due week #9 midnight ET.
    - (All students are expected to receive the proposal score if submitted).
    - Project Presentation File – 10 points, due week #12 midnight ET.
    - Project Written Report – 12 points, due week #14 midnight ET

    - If there are 2 or more teammates, the Peer Evaluation Form within each team needs to be completed (only check, no points - but possibly used to deduct grades for a team member who did not contribute much to the project).
    - The TAs will assign the grades based on their own assessment
    - All students are also expected to conduct peer review assessment
        - You are expected to write constructive comments to 3 assessments of your classmates so that students can learn from each other.
        - Rubric will be provided for the peer review
    - While the same TA will review the group's proposal, presentation, and written report, it is likely that different students will be assigned to provide peer review comments on the proposal, presentation, and written report.
- Take-Home Final (25%)
    - 25% points (20% Auto-evaluation on prediction accuracy + 5% manual grading on the written file by TAs.)
    - An .csv file on the required prediction
    - A (docx or pdf) written file to explain the methods used for the prediction
    - It will be released at 8:00 am on Friday, April 18, and due back at midnight on Sunday, April 27.

Note that the write-up is also part of grading for data analysis for homework, exams, and the final project.

Grades

Active Participation	5 points (5%)
Quizzes (4 x 5 points)	20 points (20%)
Homework (5 x 5 points)	25 points (25%)
Group Project	25 points (25%)
Take Home Final	25 points (25%)
Total	100 points (100%)

Timing Policy

- The Modules follow a logical sequence
- Assignments should be completed by their due dates
- Quizzes must be completed during the time allotted
- You will have access to the course content for the scheduled course duration.

Attendance Policy

- For the OAN students, this is a fully online course.
- Log in regularly to complete your work so that you do not have to spend a lot of time reviewing and refreshing yourself regarding the content.

Student Honor Code

All learners are expected and required to abide by the letter and the spirit of the Georgia Tech honor code.

- Review the Georgia Tech Student Honor Code www.honor.gatech.edu.
- You are responsible for completing your own work.
- Action will be taken against learners violating the Georgia Tech Honor Code.

Plagiarism Policy

- **Plagiarism is considered a serious offense. You are not allowed to copy and paste or submit materials created or published by others as if you created the materials. All materials submitted and posted must be your own. In particular, no collaboration on the take-home final exam.**
- If we find a potential violation of honor codes, we will follow the Academic Misconduct Sanctioning Guidelines Link.

Office Hours

- TAs and the instructor hold office hours every day of the week. The schedules are set at the beginning of the semester and will be posted to Piazza once they are final. They are typically unstructured, so we recommend you bring questions or topics you want to discuss with the instructors.

Office of Disability Services

We will make accommodations for students with documented disabilities if needed. These accommodations must be arranged in advance and in accordance with the Office of Disability Services (http://disabilityservices.gatech.edu).

Communication

- All discussions will take place in Piazza. The general technical questions can be submitted publicly at Piazza, and sensitive/personal requests can be submitted as a private chat at Piazza. This will allow the instructor and TAs to work as a team to handle all students' requests and questions in a timely manner.
- All learners should ask questions and answer their fellow learners' questions on the course discussion forum, Piazza. Often, discussions with fellow learners are the sources of key pieces of learning.

Netiquette

- Netiquette refers to etiquette that is used when communicating on the internet. When you are communicating via discussion forums or synchronously (real-time) or emails, please use correct spelling, punctuation, and grammar that are consistent with the academic environment and scholarship.
- Learners who do not adhere to this guideline may be removed from the course.

Tentative Course Outline (Most lectures will be released at 8 am on Fridays ET, and most assignments are due at midnight on Sundays ET unless holidays/Institute breaks)

Tentative Course Outline

Weeks	Course Topics	Release Dates
Week 1	Module 1: Introduction ● Topic 1: Intro. To Data Mining o Lesson 1: Course Overview o Lesson 2: Overview of Data Sciences o Lesson 3: Data Mining Process o Lesson 4: Introduction to R ● Topic 2: Overview of Supervised Learning o Lesson 1: Overview of Supervised Learning o Lesson 2: Cross-Validation o Lesson 3: KNN o Lesson 4: Tuning Parameters in KNN	Monday, January 6, 2025
Homework	Homework 1	Mon Jan 6 at 8:00 am Eastern - Monday, Jan 20 at 11:59 p.m. Peer Assmt: Tuesday, Jan 21 at 9:30 a.m. Eastern – Saturday, Jan 25 at 11:59 p.m. (all deadlines shift one day due to MLK Day)
Week 2	Module 2: Linear Regression ● Topic 1: Linear Regression Model o Lesson 1: Estimation o Lesson 2: Inference o Lesson 3: Example: Ball weight o Lesson 4: Evaluation & Diagnostics o Lesson 5: Model & Variable Selection o Lesson 6: Air Pollution Example	Friday, January 10, 2025
Quiz	Quiz #1	Monday, January 13 at 8: 00 a.m. Eastern – Sunday, Jan 26 at 11:59 p.m.
Week 3	Module 2: Linear Regression (Cont.)	Friday, January 17, 2025

Weeks	Course Topics	Release Dates
	• Topic 2: Advanced Topic for Linear Regression o Lesson 1: James-Stein estimator o Lesson 2: Shrinkage Method o Lesson 3: Ridge Regression o Lesson 4: LASSO o Lesson 5: Principal Components o Lesson 6: Partial Least Square	
Homework	Homework 2	Monday, Jan 20 at 8:00 a.m. Eastern – Feb 2 at 11:59 p.m. Peer Assmt: Feb 3 at 9:30 a.m. Eastern – Feb 7 at 11:59 p.m.
Week 4	Module 3: Linear Classification • Topic 1: Linear Discriminant Analysis o Lesson 1: Overview of Discriminant Analysis o Lesson 2: Linear Discriminant Analysis (LDA) o Lesson 3: QDA and Naïve Bayes Classifiers • Topic 2: Logistic Regression I o Lesson 1: Logistic Regression: Estimation o Lesson 2: Optimization in Logistic Regression o Lesson 3: Simple Logistic Regression	Friday, January 24, 2025
Quiz	Quiz #2	Monday, Jan 27 at 8:00 a.m. Eastern – Feb 9 at 11:59 p.m.
Week 5	Module 3: Linear Classification (Cont.) • Topic 3: Logistic Regression II o Lesson 4: Example for Logistic Regression (CHD) o Lesson 5: Prediction in Logistic Regression o Lesson 6: Model Selection in Logistic Regression • Topic 4: Case Study: Golf Putting o Lesson 1: Golf Putting by linear regression o Lesson 2: Golf Putting by Linear Classification o Lesson 3: Golf Putting by new domain-knowledge-based model	Friday, January 31, 2025
Homework	Homework 3	Mon, Feb 3 at 8:00 a.m. Eastern – Feb 16 at 11:59 p.m. Peer Assmt: Feb 17 at 9:30 a.m. Eastern – Feb 21 at 11:59 p.m.
Week 6	Module 4: Local Smoothers and Additive Models • Topic 1: Local Smoothers and Kernel o Lesson 1: Overview of Local Smoothers o Lesson 2: LOESS o Lesson 3: Kernel Smoothing	Friday, February 7, 2025

Weeks	Course Topics	Release Dates
	○ Lesson 4: Smoothing for Deterministic Design ○ Lesson 5: Smoothing in Stochastic Design ○ Lesson 6: Advanced topics in Kernel Smoothing [Information on the course project is released this week]	
Quiz	Quiz #3	Friday, February 7 at 8:00 a.m. Eastern – Tuesday, Feb 25 at 11:59 p.m.
Week 7	Module 4: Local Smoothers and Additive Models (Cont.) ● Topic 2: Spline Smoothing ○ Lesson 1: Introduction to interpolation Splines ○ Lesson 2: Cubic Spline Smoothing ○ Lesson 3: Optimality and Algorithm in Splines ● Topic 3: Additive Models ○ Lesson 1: Additive Model ○ Lesson 2: Generalized Additive Model ○ Lesson 3: Case study: Spam Data Example	Friday, February 14, 2025
Homework	Homework 4	Fri, Feb 14 at 8:00 a.m. Eastern – Mar 2 at 11:59 p.m. Peer Assmt: Mar 3 at 9:30 a.m. Eastern – Mar 7 at 11:59 p.m.
Week 8	Module 5: Tree-Based and Ensemble Methods ● Topic 1: Tree-Based method ○ Lesson 1: Introduction to Tree ○ Lesson 2: Growing and Pruning for Regression Tree ○ Lesson 3: Classification Tree ○ Lesson 4: Practical Issues in Tree ○ Lesson 5: Tree method in R ○ Lesson 6: Case Study: Tree in R	Friday, February 21, 2025
Course Project Proposal	Course Project Proposal (Group Assignment)	Fri, Feb 21 at 8:00 a.m. Eastern – Mar 9 at 11:59 p.m. Peer Assmt: Mar 10 at 9:30 a.m. Eastern – Mar 14 at 11:59 p.m.
Week 9	Module 5: Tree-Based and Ensemble Methods (Cont.) ● Topic 2: Ensemble Methods ○ Lesson 1: Introduction to Ensemble Methods ○ Lesson 2: Bayes Model Averaging & Stacking ○ Lesson 3: Bootstrapping ○ Lesson 4: Bagging Techniques ○ Lesson 5: Random Forest ○ Lesson 6: R Lab: Random Forest	Friday, February 28, 2025

Weeks	Course Topics	Release Dates
	(No new assignment. Continue to work on the project)	
Week 10	Module 5: Tree-Based and Ensemble Methods (Cont.) ● Topic 2: Ensemble Methods (cont.) o Lesson 7: Introduction to Boosting o Lesson 8: AdaBoosting o Lesson 9: A Statistical View of AdaBoosting Algorithm o Lesson 10: General Boosting Algorithm o Lesson 11: Boosting Algorithm in R o Lesson 12: R Lab for Boosting	Friday, March 7, 2025
Homework	Homework 5	Fri, Mar 7 at 8:00 a.m. Eastern – Mar 16 at 11:59 p.m. Peer Assmt: Mar 17 at 9:30 a.m. Eastern – Mar 21 at 11:59 p.m.
Week 11	Spring Break	March 17, 2025 (Mon) to March 21, 2025 (Fri)
Week 12	Module 6: Advanced Supervised Learning ● Topic 1: Support Vector Machine in Linear scenario o Lesson 1: Introduction to Support Vector Machine o Lesson 2: Maximum Margin Optimization for SVM o Lesson 3: SVM for Linearly Separable o Lesson 4: Slacking variables for SVM o Lesson 5: Optimization for SVM o Lesson 6: SVM for Linearly Non-separable	Friday, March 21, 2025 (* we release the submission of the project presentation/final report so that you can submit it earlier if you want)
Course Project Presentation	Course Project Presentation File (Group Assignment)	Mar 21 at 8: 00 a.m. Eastern – Apr 6 at 11:59 p.m. Peer Assmt: Apr 7 at 9:30 a.m. Eastern – Apr 11 at 11:59 p.m.
Course Project Written Report	Course Project Written Report (Group Assignment)	Mar 21 at 8: 00 a.m. Eastern – Apr 21 at 11:59 p.m. Peer Assmt: Apr 22 at 9:30 a.m. Eastern – Apr 26 at 11:59 p.m.
Week 13	Module 6: Advanced Supervised Learning (Cont.) ● Topic 2: Support Vector Machine in nonlinear scenario o Lesson 1: SVM for General Non-linear scenario o Lesson 2: Linearization by Kernel Function o Lesson 3: SVM in the General Nonlinear Case o Lesson 4: Regularization framework for SVM o Lesson 5: Support Vector Regression and Multi-class SVM o Lesson 6: R lab	Friday, March 28, 2025

Weeks	Course Topics	Release Dates
Quiz	Quiz #4	Fri., Mar 28 at 8:00 a.m. Eastern – Apr 13 at 11:59 p.m.
Week 14	Module 6: Advanced Supervised Learning (Cont.) ● Topic 3: Neural Networks o Lesson 1: Introduction to Neural Networks o Lesson 2: Backpropagation Algorithm o Lesson 3: Fitting Neural networks o Lesson 4: Deep Neural Networks o Lesson 5: R Lab 1 o Lesson 6: R Lab 2 (Work on course project)	Friday, April 4, 2025
Week 15	Module 7: Unsupervised Learning ● Topic 1: Cluster Analysis o Lesson 1: Introduction to Unsupervised Learning o Lesson 2: K-means Algorithm o Lesson 3: Example for K-means o Lesson 4: Kernel K-means o Lesson 5: R lab: K-means o Lesson 6: EM Algorithm	Friday, April 11, 2025
Course Project Written Report	Course Project Written Report (Group Assignment)	Mar 21 at 8: 00 a.m. Eastern – Apr 21 at 11:59 p.m. Peer Assmt: Apr 22 at 9:30 a.m. Eastern – Apr 26 at 11:59 p.m. Peer Evaluation of Teammates within the project group, if applicable: Apr 22 at 9:30 a.m. Eastern – Apr 26 at 11:59 p.m.
Week 16	Final Exam (Take Home)	Friday, April 18 at 8:00 a.m. Eastern – Sun, April 27 at 11:59 p.m. Eastern

Course Materials/Textbook

- All content and course materials can be accessed online
- There is no required textbook for this course however "An Introduction to Statistical Learning" is highly recommended and is available for free at https://www.statlearning.com/.

Technology/Software Requirements

- Internet connection (DSL, LAN, or cable connection desirable)
- Adobe Acrobat PDF reader (free download; see https://get.adobe.com/reader/