Statistical Methods for Machine Learning I

University of Toronto
Department of Statistical Sciences
STA314H1S Summer 2024

	LEC 0101:
Monday Lectures	10am-1pm ET at SF1101
Friday Lectures	2pm-5pm ET at SF1101
Tutorial 1	Friday 1pm-2pm ET at MS2173
Tutorial 2	Friday 1pm-2pm ET at GB119
Instructor OH:	TBA Starts Week 2
OH Locations:	TBA

Office Hours and tutorials will start on the second week. There will be additional office hours held by the instructors before the final exam.

COURSE OVERVIEW

Course Description: Briefly, the contents will focus on Statistical methods for supervised and unsupervised learning from data: training error, test error and cross-validation; classification, regression, and logistic regression; principal components analysis; stochastic gradient descent; decision trees and random forests; k-means clustering and nearest neighbour methods. Computational tutorials will support the efficient application of these methods.. Statistical analysis will be conducted using `R` and python. The students will be expected to simulate datasets using `R`, as well as interpret `R` and python code and output on tests and assignments.

Content, emphasis, etc. of the course is defined by means of the lecture material — not only the posted lecture notes. It is important to attend all lectures, as there is normally no simple way to make up for missed lectures (perhaps obtain another student's notes). There will also be many lecture examples using statistical software `R` and python.

Learning Outcomes: By the end of this course, all students should have a solid understanding of both the mathematical theory of machine learning models, as well as their application in the form of data analysis. Students should be prepared to show their understanding of the above through:

- application of methods through problem-solving questions;
- description and explanation of concepts relating to the mathematical theory;
- derivation and proof of topics based on machine learning concepts and theory;
- practical application of methods on real data using statistical software `R` or python, with appropriate justification of use of these methods;
- interpretation of data analysis results in clear and non-technical language

Pre-requisites: The prerequisites are:
Prerequisite: STA302H1/ STA302H5/ STAC67H3; CSC108H1/ CSC110Y1/ CSC120H1/ CSC148H1/ CSCA08H3/ CSCA48H3/ CSCA20H3/ CSC108H5/ CSC148H5; MAT223H1/ MAT224H1/ MAT240H1/ MATA22H3/ MATA23H3/ MAT223H5/ MAT240H5/ MATB24H3/ MAT224H5; MAT235Y1/ MAT237Y1/ MAT257Y1/ (MATB41H3, MATB42H3)/ (MAT232H5, MAT236H5)/ (MAT233H5, MAT236H5)
Exclusion: CSC411H1, CSC311H1, STA314H5, STA315H5, CSCC11H3, CSC411H5

Pre-requisites are **strictly enforced by the department, not the instructor.** If you do not have the equivalent pre-requisites, you will be un-enrolled from the course.

COURSE MATERIALS

Course Content: All lecture slides and materials will be posted on the Quercus course page for each lecture section. Furthermore, any important announcements will also be posted in Quercus. Please make sure to check it regularly so you don't miss anything.

Textbook: There are no required textbooks. All assessments will be conducted based on lecture materials. However, there are some useful books which will be used for references. The books are available online.

1. James, Gareth, et al. " *An introduction to statistical learning: With applications in R and python.*". Springer Nature, 2023. Link ISL Link
2. Hastie, Trevor, et al. "*The elements of statistical learning: data mining, inference, and prediction. Vol. 2.* New York: springer, 2009. ESL Link
3. Simon N. Wood. "*Generalized Additive Models: An Introduction with R*". CRC press, 2017.

Statistical Software: We will be using the R and python for performing statistical analyses in this course. Both is a free software that can either be downloaded onto your personal computer or used in a cloud environment. We encourage all students to use RStudio and Jupyter notebook through the JupyterHub for University of Toronto. This will allow you to login with your official UofT credentials and use RStudio without the need for a local installation and can be run on any device that has access to an internet connection. More information about using RStudio in JupyterHub will be provided early in the term. The codes shown in class will be available on the course page and, along with any additional resources, should be sufficient to complete any assessment involving data analysis.

COURSE COMPONENTS

Lectures: Lectures will be conducted in person in SF1101. Slides will be available after the class. Class time each week will comprise of a combination of lecturing, and code-along sessions. Where possible, you are encouraged to bring a laptop or tablet to follow along with the code.

Tutorials: There will be two tutorials sections in this course, which will be led by the TAs. Tutorials will be conducted in person in MS2173 and GB119. TAs will go over some codes and solve some mathematical problems. All the tutorials will start on Week 2.

Office Hours: Instructor and TAs will hold office hours in a combination of online and in-person formats. The office hour schedule and mode of delivery will be posted on Quercus once finalized. It is recommended that you visit office hours whenever you have a question about the material. It is always important to have material clarified as quickly as possible. Don't wait until the last minute to ask your questions!

Piazza: We will be using the Piazza as an online discussion forum, which can be accessed through the Quercus course page. **All questions about course material should be posted here** or asked during TA/instructor office hours. The instructor and TAs will monitor the board and will help answer questions but students are encouraged to answer posts and help their fellow classmates.

COMMUNICATION

How your instructor will communicate with you: All communication will be made through Quercus announcements or during lectures. Please ensure that you check Quercus regularly so you don't miss

anything important.

Where to send content questions: We will be using the Piazza to collect student questions regarding course content, assignments, etc. All questions should be posted here.

When to email the instructor: The instructor will only respond to emails of a private or sensitive nature. If you email the instructor with content related questions, you will be asked to repost your question on the content board so the answer may benefit all students. Should you need to email the instructor about a sensitive or personal nature, please use your official mail.utoronto.ca email, include your full name and student number in the text. Send all course related emails to sta314@utoronto.ca. Please allow up to 48-96 hours for a reply. Emails will not be monitored on evenings and weekends.

A note on email and discussion board etiquette: Please make sure that you communicate politely and respectfully with all members of the teaching team and your fellow classmates. Written communications can sometimes take a tone other than what was intended (e.g. can come off as dismissive, rude or insulting), so make sure you re-read or read out loud your email/post before sending it to make sure it has the tone you intended. For more tips on respectful communication, see professional communication tips. Piazza is a teaching and learning tool and therefore should only be used as such. Any posts that detract from the learning goal of the board will be removed to keep the board a safe space.

GRADING SCHEME

All the students will be evaluated in the following way:

Assessment	Date (tentative)	Weight
Assignment 1	July 19	15%
Term Test	July 26	25%
Assignment 2	August 13	15%
Final Exam	August 15-23	45%

Please note that the last day to drop the course without penalty is July 29, 2024.

EVALUATION BREAKDOWN

Term Test: The term test will be conducted in person during the scheduled Friday class time (see top of page 1). The test will be approximately 2 hours long. More details will be communicated closer to the test date. The test will cover material from Lectures 1-6.

Assignment: You will be given two assignments in the term. The purpose of this assignment is to develop your understanding of the properties of the machine learning models taught during the lectures. This will be useful for developing predictive modeling skills as well as to develop practical understanding of the methods taught in the class. The assignment will have a heavy focus on the use of statistical software (R and python), and will involve applying the methods learned during lecture to a data set. The format of the assignments will be as follows:

1. use the methods taught in lecture to perform a small data analysis by predictive modeling.

2. simulate unique datasets and writing your own functions instead of built in R/python functions.
3. solve some mathematical problems and explain the procedure with simulated datasets

Final Exam: The details about the final exam will be provided during the last week lectures. For the final exam we will be following standard University of Toronto Schedule. the final exam will be three hours in duration and will be scheduled by the Faculty of Arts and Science during the final assessment period.

LATE ASSESSMENT AND EXTENSION REQUEST POLICY

The assessment deadlines may change from the ones stated in the syllabus depending on how the lecture progresses. However, once the deadline(s) has been announced, the students need to submit the assignments by the deadline. Students will be able to still submit the assignments up to 5 days after the deadline, however, each additional day will be accounted for 20% penalty.
Extreme Situations/Prolonged Illness Extensions: Should a student be experiencing a prolonged illness or other situation that prevents them from turning in their work by the deadline, they should **immediately contact their instructor and College Registrar** to inform them of their situation. They should also submit an Absence Declaration form on ACORN that lists every day during which they were incapacitated and unable to work. Accommodations or further extensions will not be considered without a completed declaration, and will only be considered for extreme circumstances.

Accessibility-Related Extension Requests: Students registered with Accessibility Services should notify the instructor as soon as possible if additional time is needed on assessments that are eligible for extensions. Please **notify the instructor by email of your situation and cc your accessibility advisor** in the process. The instructor will work with the accessibility advisor to determine an appropriate extension for your situation.

MISSED ASSESSMENT POLICY

If you experience a prolonged absence due to illness or emergency that prevents you from completing any number of assessments, please contact your College Registrar as soon as possible so that any necessary arrangements can be made.

Missed Assignment: Missing assessments will receive a 0.

Missed Term Test: If a student is experiencing a serious personal illness or emergency on the date of the test, the student **must declare their absence on ACORN and notify the teaching team via email no later than one week after the date of the test**. The weight of the term test then will be transferred to the final exam. That is the student who missed the term test will have the final exam worth 70% of the total marks.

REGRADE REQUESTS

Regrade requests will be accepted for all assessments except the final exam. Regrade requests must provide a justification for where there exists a grading error and/or how the work meets the grading rubric. These justifications must further be backed up with concrete references to the course material. All regrade requests will be accepted through a form available on the Quercus course page and will be accepted no later than one week after the grade for that assessment is released. **No regrade requests will be accepted by email or after the 1 week deadline.** The instructor further reserves the right to re-evaluate the assessment in its entirety (i.e. grades can go up, down, or remain unchanged). Please allow a few weeks for regrade requests to be processed by the instructor.

INTELLECTUAL PROPERTY

Course materials provided on Quercus, such as lecture slides, assignments, tests and solutions are the intellectual property of your instructor and are for the use of students currently enrolled in this course only. **Providing course materials to any person or company outside of the course is unauthorized use**. This includes providing materials to predatory tutoring companies.

ACADEMIC INTEGRITY

The University treats cases of plagiarism and cheating very seriously. It is the students' responsibility for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters. All suspected cases of academic dishonesty will be investigated following procedures outlined in the above document. If you have questions or concerns about what constitutes appropriate academic behavior or appropriate research and citation methods, you are expected to seek out additional information on academic integrity from your instructor or from other institutional resources (see http://academicintegrity.utoronto.ca/). Here are a few guidelines regarding academic integrity:

- Sharing or discussing questions or answers with other students during tests is an academic offense.
- Students must complete all assessments individually. Working together is not allowed.
- Paying anyone else to complete your assessments for you is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- All work that you submit must be your own! You must not copy mathematical derivations, computer output and input, or written answers from anyone or anywhere else. Unacknowledged copying or unauthorized collaboration will lead to severe disciplinary action, beginning with an automatic grade

of zero for all involved and escalating from there. Please read the UofT Policy on Cheating and Plagiarism, and don't plagiarize.

RULES REGARDING THE USE OF GENERATIVE AI IN ASSESSMENTS

Generative Artificial Intelligence (AI), and specifically foundational models that can create writing, Computer code and/or images using minimal human prompting are proliferating and becoming ubiquitous. This includes not only GPT-4 (and its siblings ChatGPT, Gemini and Bing), but many writing assistants. that are built on this or similar AI technologies. There are now hundreds of these systems that are readily available. In this course, the use of such AI tools is limited. The students are most welcome to use the tools for learning purposes and to gather information about the topics that are taught during the course. However, these tools should not be used to complete any of the assessments in the course. For details please read the following statements:

- Students may use artificial intelligence tools for creating an outline for an assignment, but the final submitted assignment must be original work produced by the individual student alone.
- Students may not use artificial intelligence tools for taking tests, writing project report, creating R codes, or completing major course assignments. However, these tools may be useful when gathering information from across sources and assimilating it for understanding.
- Representing as one's own an idea, or expression of an idea, that was AI-generated may be considered an academic offense in this course.
- Students may not copy or paraphrase from any generative artificial intelligence applications, including ChatGPT and other AI writing and coding assistants, for the purpose of completing assignments in this course.
- This course policy is designed to promote your learning and intellectual development and to help you reach course learning outcomes.

ACCESSIBILITY NEEDS

Students with diverse learning styles and needs are welcome in this course. If you have an acute or ongoing disability issue or accommodation need, you should register with Accessibility Services (AS) at the beginning of the academic year by visiting https://studentlife.utoronto.ca/department/accessibility-services/. Without registration, you will not be able to verify your situation with your instructors, and instructors will not be advised about your accommodation needs. AS will assess your situation, develop an accommodation plan with you, and support you in requesting accommodation for your course work. Remember that the process of accommodation is private: AS will not share details of your needs or condition with any instructor, and your instructors will not reveal that you are registered with AS.

CLASS SCHEDULE - TENTATIVE

This is the tentative outline for Summer 2024. Topics may be reduced or additional topics may be added by course instructor's discretion.

Week	Content
1 (July 5)	Review of Linear and Logistic Regression. Gauss-Markov assumptions. Iteratively-reweighted-least-square method. ROC curve
2 (July 8)	Bias-variance trade-off. Regularized methods for regression. L2 and L1 regularization. Estimate parameters in LASSO using coordinate descent. K-nearest neighbors.
3 (July 12)	Linear and cubic splines. Scatter plot smoothing. Generalized Additive Model (GAM).
4 (July 15)	Classification. Linear discriminant analysis (LDA). Naive Bayes.
5 (July 19)	Resampling methods: Cross Validation and Bootstrap. Error rates, optimism.
6 (July 22)	Classification and Regression Trees (CART). Pruning. Bagging.
7 (July 26)	**Term Test**
8 (July 29)	Random Forest and Variable Importance. Gradient Boosting. Xtreme Gradient Boosting (XGBoost)
9 (August 2)	Artificial Neural Networks (ANN). Gradient descent and Stochastic Gradient Descent.
10 (August 5)	No Lecture.
11 (August 8)	Unsupervised learning. Principal Components Analysis (PCA). Singular Value Decomposition (SVD).
12 (August 12)	QR decomposition. K-Means clustering. Hierarchical clustering.
August 15-23	**Final assessment period**