STA 314: Statistical Methods for Machine Learning I

July 8th, 2023

Objectives: This course aims to provide a broad introduction to the fundamental concepts used in common Machine Learning (ML) methods. After taking this course, students are expected to have a solid understanding of the statistical concepts underlying the ML methods, and fluency in applying these ML methods through the programming environment R. It serves as a foundation for more advanced courses, such as STA414 (Statistical Methods for Machine Learning II).

We will cover statistical methods for supervised and unsupervised learning from data, including topics such as regression and classification; regularization; cross-validation; principal component analysis (PCA); Bootstrapping methods; clustering and Bayesian inference. Tutorials will focus on the computation and application of these methods.

Instructor: Ziang Zhang (aguero.zhang[at]mail.utoronto.ca)

Lecture: In-person (no video recording), Friday 14:00 – 17:00 PM (Location: MS 2170)

Instructor Office Hours: After class, or by appointment through email.

Teaching Assistants:

- Arturo Esquivel (a.esquivel[at]mail.utoronto.ca) for TUT0101 (Location: MS 2173)
- Vedant Choudhary (vedant.choudhary[at]mail.utoronto.ca) for TUT0102 (Location: MS 3278)

Tutorial: In-person, Friday 13:00 – 14:00 PM.
Note that you **have to attend** the exact tutorial session that you enrolled on acorn, since the quiz will be taken during the tutorial session.

Course Pages:

1. Quercus `https://q.utoronto.ca/courses/308730` will be used for announcements regarding lectures, tutorials, quizzes, tests and exams etc. The lecture materials and assigned problem sets will be posted here.
2. Piazza `https://piazza.com/class/lhtkgqcuj8e424/` will be used for the course forum. If your question is about the course material and doesn't give away any hints for the homework, please post to Piazza so that the entire class can benefit from the answer.
   As a motivation for students to participate on Piazza, the instructor team will regularly endorse good questions and response. By the end of the term, students who have received ≥ 5 good questions or endorsed responses will be awarded a 1 percent bonus at their final grades.

Textbook: The majority of the course content will be based on the textbook:

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, which can be accessed freely online.

Students are only responsible for the material covered in lectures, tutorials, and homeworks. There are some additional references that may be helpful:

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*.
- Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

- Kevin Murphy. *Machine Learning: a Probabilistic Perspective*.

Prerequisites: An undergraduate-level understanding of probability, statistics, calculus and linear algebra is assumed. Some basic familiarity with programming in R is also assumed. Check out the department website for more detailed requirements.

Grading Policy: Quizzes (30%), Midterm (30%), Final (40%).

Quizzes and Problem sets: The quizzes will be held *during the tutorial time*, with *questions based on the assigned problem set* each week. The assigned problem set will be posted approximately weekly. Most problem sets will include a computational component. *You should bring printouts of your work to the quiz, as you will be asked to answer questions based on your printouts.* Possibly, one of the quiz questions will require you to hand in your printout. The non-computational parts of the assignments are intended to help you better understand the lecture content, and *will not be collected or graded.* The *lowest quiz will be dropped.*

Midterm: The midterm will be held during the lecture time on July 28th at EX320. A representative practice midterm will be posted before the midterm to help you understand the midterm format.

Collaboration policy:
For the non-computer part of each assigned problem set, students are allowed to collaborate with each other. For the computer part of each assigned problem set, students are **required to work on the question alone**. Never, ever, bring a copy of somebody else's printout, or allow anyone to have a copy of yours. If you allow anyone to have an electronic copy of your computer work, for any reason, you are committing a serious academic offence.
For quizzes, midterm and final exam, collaboration with other students is strictly forbidden.

Regrading policy: For the midterm and quizzes, regrading requests should be submitted via email to the instructor. Regrading requests must include the student's name, student number, and a justification for the request, which refers specifically to the student's answers and the course materials. Requests without this justification will not be considered. The deadline for requesting a regrading is one week after the marks are returned. Remarks may result in a decrease in the grade.

Missed test and quizzes: If a student self-declares the absence through the online absence declaration, and therefore misses the quizzes or midterm test:
Quizzes: The grade for the first missed quiz will be dropped and your quiz grade will be based on the remaining three quizzes. If you miss another quiz, the grade for your second missed quiz will be equally divided into the midterm and final (so now the midterm weighs 35% and final weighs 45%). If you miss more than two quizzes, the grade of each additionally missed quiz will be counted as 0, which means you will lose 10% of your grade for each missing quiz.
Midterm Test: If you miss the midterm test, all of the midterm grades will be reweighted into the final exam.

Important Dates:

Quiz 1	July 14th
Quiz 2	July 21st
Midterm	July 28th
Last day to drop	July 31st
Quiz 3	Aug 4th
Quiz 4	Aug 11th
Last day for CR/NCR, or LWD	August 15th
Final Exam	TBA

Tentative Course Outline:

- Week 1 (July 7th): Review of Linear Algebra and Basic Statistics; Bias-variance Tradeoff; Training/Testing Separation; Nearest Neighborhood Methods.
- Week 2 (July 14th): Cross-Validation; Model Selection& Regularization; Principal Component Analysis (Quiz 1).
- Week 3 (July 21st): Classification; Generative and Discriminative Methods; Support Vector Machine (Quiz 2).
- Week 4 (July 28th): Midterm Exam (held in lecture time).
- Week 5 (Aug 4th): Decision Trees; Bootstrapping and Ensamble Methods (Quiz 3).
- Week 6 (Aug 11th): K-means Clustering; Introduction to Bayesian Methods (Quiz 4).

Academic Honesty: The University supports acting in honesty, trust, fairness, respect, responsibility, and courage in all aca- demic matters. Students are responsible for knowing the content of the Universitys Code of Behaviour on Academic Matters. All suspected cases of academic dishonesty will be investigated following procedures outlined in the Code of Behaviour above. If you have questions or concerns about what constitutes appropriate academic behaviour 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 (http://academicintegrity.utoronto.ca).

Copyright: The unauthorized use of any form of device to audiotape, photograph, video-record or otherwise reproduce lectures, course notes or teaching materials provided by instructors is covered by the Canadian Copyright Act and is prohibited. Students must obtain prior written consent to such recording. In the case of private use by students with disabilities, the instructor's consent must not be unreasonably withheld.

Accommodations for disability policy: If you have a disability or health condition that may require accommodations, please approach the Accessibility Services. To arrange accommodations, you will need to submit a letter of accommodation from Accessibility Services to the instructor at least 7 days before the exam date.