STA 314: Statistical Methods for Machine Learning I

Logistics and why this course

Xin Bing

Department of Statistical Sciences
University of Toronto

About this course

- This course is a broad introduction to machine learning from a statistical perspective (aka statistical learning). We put emphasis on intuition and basic mathematical derivations of how and why popular machine learning methods work.

- We will focus on understanding methodology rather than implementing complicated machine learning algorithms or delving into deep theory.

- You will learn examples of applying popular machine learning methods to real data sets in R & python.

About this course

We cover two types of learning problems:

- Supervised learning (80%)
  - Regression
  - Classification
- Unsupervised learning (20%)
  - Dimension reduction
  - Clustering
  - Matrix factorization
- This includes a variety of important methods:
  - linear regression, logistic regression, non-parametric regression, nearest neighbours, decision trees, bagging, boosting, random forests, SVMs, (deep) neural networks
  - PCA, K-means, matrix completion, topic modeling

Do I have the appropriate background?

Coursework is aimed at advanced undergrads and graduate students. We will use multivariate calculus, probability, and linear algebra.

- Linear algebra: vector/matrix operations such as eigenvalues and eigenvectors, eigen and singular value decompositions, inverse, trace, norms.
- Calculus: partial derivatives/gradient.
- Probability & Statistics: expectation, variance, covariance; Bayes' theorem; common distributions; maximum likelihood estimation, simple linear regression, point and interval estimation, hypothesis testing, p-values.

Do I have the appropriate background?

- Programming language: we are using R in this course.
  - Useful resources: `https://cran.r-project.org/`. A good review of some basic R commands is in Chapter 2.3 of the textbook.
  - How much do you need to know?
    - Basic knowledge on R is required (e.g., load data, create a vector or matrix, etc.)
    - The tutorials will provide you demonstrations of using R to perform statistical analysis.
    - The emphasis of coding will be on the use of the various R packages and on the implementation of the key subroutines of ML methods.
    - You will not be required to **implement complicated machine learning algorithms** nor to **write an entire R package**.

Textbook and other suggested readings

We mainly use the textbook

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*.
- You can find it via `https://www.statlearning.com`.

You are only responsible for the material covered in lectures, tutorials, and practical problem sets. You may also find the following references useful throughout the course:

- Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning*.
- Christopher Bishop. *Pattern Recognition and Machine Learning*.
- Kevin Murphy. *Machine Learning: a Probabilistic Perspective*.

There are lots of other freely available, high-quality ML resources.

Course Information

Course website: the main source of information; check regularly!
`http://courses.utstat.utoronto.ca/sta314_f24/`

Course email: sta314@course.utoronto.ca.

Crowdmark: course project & exams (submission and grading).

Quercus: annoucements only.

Piazza: the main place for discussions. Sign-up:
`https://piazza.com/utoronto.ca/fall2024/sta314h1`

Email policy

- Please, *do not* send emails to either the instructor or the TAs' personal / professional emails, except for absolutely urgent requests.
- For questions / requests,
  - if it is about course material such as lectures / tutorials, or clarification question, post it to Piazza so that other students will benefit.
  - if it is about requests such as regrading request, use the course email.
  - if it is related with solving practical problem sets, use the office hours.

Course delivery instructions

- All sections (LEC0101 and LEC0201) have the same delivery layout.

Weekly	delivery mode
1-hr lec on Mon & 2-hr lec on Wed	in-person
1-hr tutorial	in-person
5-hr office hour $(1 + 4)$	zoom / in-person

- Tutorials are highly recommended as they contain supplementary materials to the lectures. Some weeks might not have tutorials. Quizzes will be given during tutorials.

Course Information

All information is in the syllabus on the course website.
If you remember just one thing:

Check the course website regularly.

Grading

- (5%) 5 quizzes
  - Basic and short questions on course materials
  - Weighted equally
  - Hand-in during tutorials
- (50%) Two midterm tests
  - Each has 25% weight
  - 2-hour held during normal class time
  - See the syllabus or course website for exact date and location.
- (25%) Final test
  - 2-hour held during the final assessment period
  - Date, time and location are TBA.
- (20%) Course project
  - Initiated later and is due in the final assessment period
  - Group of size 1-4, one report
- (2-3%) Bonus

More on tests

- Test will be closed-book without aid of any form such as calculator, cheat sheet, etc.
- For any midterm test you missed, its grading weight will be equally added up to the other exams that have not been taken.
  - If you missed the first midterm, both the second midterm and the final exam will be worth 37.5% per each. If you further missed the second midterm, the final will be worth 75%.
  - If you took the first midterm but missed the second midterm, your first midterm will still be worth 25% but your final will be worth 50%.

Practical problem sets

- About 4 sets of practical problems
- Deepen your understanding of the course material
- Practical application of the methods taught in class
- Not graded but you are expected to be able to solve them independently

Similar course: CSC311–Introduction to Machine Learning

- **STA314 takes a more statistical perspective than CSC311** while their core contents share the same machine learning methods.
  - The course will focus on the methodology and statistical insight rather than algorithm (or coding).
  - We do not cover reinforcement learning.
  - We will cover model selection, high-dimensional statistics, bootstrap, etc.

Statistical Learning versus Machine Learning

- Machine Learning (ML) is a subfield of Artificial Intelligence while Statistical Learning (SL) is a subfield of Statistics.
- They both try to uncover patterns in data.
- Both fields draw heavily on calculus, probability, and linear algebra, and share many of the same core algorithms
- ML puts more emphasis on algorithms, computation and prediction accuracy while SL emphasizes more on models and their interpretability, and how to evaluate uncertainty of the learning procedure.
- This course focuses on **Statistical Learning**.

Advanced Courses

This course will help prepare you for the following courses.

- **STA414** (Statistical Methods for Machine Learning II)
  - This course is the follow-up course, which delves deeper into the probabilistic interpretation of machine learning.
- **CSC413** (Neural Networks and Deep Learning)
  - This course covers deep learning and automatic differentiation.
- **CSC412** (Probabilistic Learning and Reasoning)
  - The CSC analogue of STA414.

Why this class?

"I've heard that neural networks solve everything, can we just learn those?"

- There's a whole world of problems where neural nets do not work.
- The techniques in this course are still the first things to try for a new ML problem.
  - E.g., try logistic regression before building a deep neural net!
  - It is important to accurately assess the performance of a method, to know how well or how badly it works or will work.
    (Easier for simple methods)

Why this class?

- The principles you learn in this course will be essential to understand and apply neural nets.
  - It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
  - Advanced algorithms are built on the simpler ones.
- Statistical learning is a fundamental ingredient in the training of a modern data scientist / quantitative analyst
  - science (biology, neuroscience, medicine)
  - industry (tech company, transportation)
  - finance (quant, trading, bank)
  - $\vdots$

Questions on logistics?