NBA 4920/6921: Macihne Learning Applications in Business

Syllabus v3

Fall 2021

Instructor: Murat Unal	E-mail: mu96@cornell.edu
TA: Ishneet Kaur Sachar	E-mail: iks23@cornell.edu
Class Hours: T/Th 1:00-2:15PM	Class Room: Sage Hall 141
Office Hours: T/Th 2:30-3:30PM	Office: Sage Hall 127
TA Office Hours: M/W 3:00-4:00PM	Office: Sage Hall 127

Course Description

This course introduces students to machine learning techniques used in business applications to draw managerial insights from data.

Course Objectives

The overarching goal of the course is to assist students in developing the set of skills required for applying machine learning methods. Like any skill, this can be accomplished through first absorbing the related knowledge and then developing the ability to use them. As such we aim to accomplish the following:

1. Develop the ability to distinguish between problems that require inference and those that call for prediction.
2. Develop familiarity with common machine learning algorithms and their strengths/weaknesses.
3. Build intuition for prediction—especially the bias-variance trade-off.
4. Develop the ability to apply the learned methods in R.

Prerequisites

I expect you to be comfortable with the basic concepts of probability and statistics typically taught in any basic probability course: BTRY 3080, ECON 3110, ECON 3130, ILRST 3110, ILRST 3080, MATH 4710, STSCI 3110, or STSCI 3080.

Should you take this course?

- **Yes** if:
- You are not only interested in learning machine learning methods but also eager to apply them.
- You have the time and willingness needed to put in the effort.
- **No** if:
- You have never taken any basic probability and statistics class.
- You already took a related class and/or have experience in applying machine learning methods.
- You are looking for a more theoretical and mathematical treatment of machine learning algorithms.
- You think this would be an opportunity for easy credits.

Required Material

Software and Tools

We will be using the R programming language. R is the industrial strength software for data analysis. It's free, cross platform, and hugely capable. The learning curve can be steep, but is very worthwhile.

Books

Lectures will be based on the chapters from

1. **Introduction to Statistical Learning** (ISL)
2. **Data Visualization** (DV)
3. **The Hundred-Page Machine Learning Book** (100ML)

Readings

1. Miller & Kosanagar (MK), Harvard Business Review (2019), *How Targeted Ad and Dynamic Pricing Can Perpetuate Bias*
2. Barocas & Selbst (BS), California Law Review (2016), *Big Data's Disparate Impact*

Slides and Code

I will post the lecture slides and R code to Canvas.

Additional Resources

1. **R Studio's repository** for learning R. Books, tutorials, cheatsheets.
2. **R for Data Science**
3. **Machine Learning Crash Course**

Course Policies

Covid-19

In order to maintain a safe environment, please note that all students in this course are expected to abide by the Cornell University Campus Behavioral Compact. This includes, but is not limited to: adhering to all in-person course attendance rules (e.g. mask wearing, social distancing, assigned seating, etc.), submitting to regular testing, completing the Daily Check, avoiding out-of-area travel, and staying abreast of all Cornell COVID-19 news and requirements (including any rule changes) as outlined on the University's COVID-19 **website**.

Mode of Instruction

Per university guidelines, the primary mode of instruction will be live on campus. If you believe you qualify for medical accommodation, please reach out to Student Disability Services. We are happy to offer attendance by Zoom as an option for those who qualify.

Laptop/Tablet

You may use your laptop/tablet for classroom activities. If you are watching videos or engaging in other distracting activities on your laptop/tablet, I will ask you to leave.

Cellphone

No cellphones. You are not allowed to use your cellphone in the classroom for any purpose. If you have to use your cellphone you may leave the classroom and do so outside. Offenders will lose 1 percentage point off of their final grade.

Zoom recordings

Lectures will be recorded and the recordings will be made available after the class. You can access them in Canvas under the Zoom tab.

Communication

If you want to send me an email about anyting related with this course please write NBA 4920/6921 in the subject line. I give priority to your emails and this way I will respond to them in a timely manner. Otherwise, your emails might sit in my inbox until I find the time to go through them.

Attendance

Attendance is expected in all lectures. Valid excuses for absence will be accepted before class. I will take attendance through quizes on Canvas. Every class I will publish a quiz, which will be worth 4 points. The correct answer will be worth 1 point and you will get 3 points for attempting the quiz in class.

Grading

The typical Cornell University grading scale will be used. I reserve the right to curve the scale dependent on overall class scores at the end of the semester. The grade will count the assessments using the following proportions:

- **30%** final exam.
- **25%** class project
- **25%** 4 assignments
- **20%** participation (10% attendance, 10% discussion participation)

Assignments

Assignments must be typed in and submitted through Canvas before class on the day they are due. You may work in groups but everyone must submit their own unique work. You may complete your analyses in any programming language your prefer. However, keep in mind that questions from R related material that we will be covering in the lectures can appear in the quizes, assignments and exams.

Late Assignments

After the deadline, assignments will be accepted for a 10% deduction to the score for every late day up to 3 days after the deadline. After this any assignments handed in will be given 0.

Re-grading

If you believe that we made a mistake in grading your deliverable or exam and would like to request a re-grade, send me an email explaining the mistake by 6:00 PM of the day after your work has been returned to you. We will re-evaluate your work, which may result in your grade going up, going down, or remaining unchanged.

Class project

Description

1. Create a 2-4 person group.
2. Find a prediction problem and confirm it with me.

- Important considerations:
  - Is this an important problem? Why should we care?
  - Can you collect the data?
  - Data sets from the class and ISL are off limits.
  - Do you have a clearly defined outcome and a good amount of predictors?

3. Collect the data.
4. Tell a story with the data: explore and visualize.
5. Train four different algorithms you have seen in class.
6. Estimate your cross validation error with a metric suitable for your problem.
7. Test your model on a held out data.
8. Create your report and slides.

Deliverables

Project groups and topics are due before class on 10.07.
Submit the following on Canvas before class on 11.30.
Late submissions will not be accepted.

1. An R markdown/Jupyter notebook/Kaggle notebook file showing all your analyses, code, and figures
2. 1-2 page executive summary
3. 4-5 slide presentation
4. Evaluation of your group member's contribution (submitted individually)

Markdown/notebook

Organize your markdown/notebook in accordance with the steps you followed in your analysis.

1. Data
2. Cleaning
3. Tuning
4. Training
5. Prediction

Treat this as something you would include in your potential job application. Anyone who receives it should be able to follow your steps and replicate your analysis.

Executive summary

The executive summary should be typed, clearly organized and written well. It should have the following structure:

1. The big picture:
    - What is your problem?
    - Why should we care?
2. Data:
    - Sources
    - Cleaning
    - Challenges
3. Methods:
    - ML methods applied
    - Tuning methods applied and tuned parameters
4. Results and conclusion:
    - Performance metrics you used
    - Model performance
    - Managerial insights you obtained

Presentation

You will present your work to the class in a 10-minute presentation during one of the two sessions: 11.30 and 12.02

Evaluation

Describe a short evaluation whether each member in your group contributed equally or differently. If the work was unequal, describe the differences and whether you believe you deserve a different grade than the group.

Academic Integrity and Honesty

You are expected to abide by the Johnson School Honor Code and the Cornell University Code of Academic Integrity. Any work you submit must be your own or your fair share of a team project. It is a violation of the Honor Code to seek or use case or problem-specific help from others (online or in person) who have previously studied the same case or problem. Cheating or plagiarizing of any sort on any component of this class will result in a failing grade for the term and a report of the offense to the university.

Accommodations for Disabilities

Students with Disabilities: Your access in this course is important to me. Please request your accommodation letter early in the semester, or as soon as you become registered with Student Disability Services (SDS), so that we have adequate time to arrange your approved academic accommodations.

- Once SDS approves your accommodation letter, it will be emailed to both you and me. Please follow up with me to discuss the necessary logistics of your accommodations.
- If you are approved for exam accommodations, please consult with me at least two weeks before the scheduled exam date to confirm the testing arrangements

Tentative outline

- This class schedule is tentative and may change as the term proceeds. All changes to the class schedule will be announced in class and posted on Canvas.
- Two 90 min introduction to R sessions will be held by the TA on 08/30 Monday in Sage 141 and on 09/01 Wednesday in Sage B08 both between 5:00 and 6:30pm. I urge you to take this opportunity and start expanding your R knowledge early on.

Date	Topic	Reading
08.26 Th	Course introduction	
08.31 T	Introduction to machine-learning	ISL Ch 2.1
09.02 Th	Data exploration & visualization	DV Ch 1.1 & Ch 3
09.07 T	Linear regression part 1	ISL Ch 3
09.09 Th	Linear regression part 2	ISL Ch 3
09.14 T	Logistic regression	ISL Ch 4.3
09.16 Th	Classification performance metrics	ISL Ch 4.3
09.21 T	Prediction errors & the variance-bias trade-off	ISL Ch 2.2
09.23 Th	Hold-out methods	ISL Ch 5.1
09.28 T	Hold-out methods in R	ISL Ch 5.1
09.30 Th	Linear model best subset selection	ISL Ch 6.1
10.05 T	Linear model stepwise selection	ISL Ch 6.1
10.07 Th	Linear model selection in R	ISL Ch 6.2
10.12 T	Fall break - no class	
10.14 Th	Shrinkage methods: ridge regression	ISL Ch 6.2
10.19 T	Shrinkage methods: lasso regression	ISL Ch 6.2
10.21 Th	Shrinkage methods: elastic net	ISL Ch 6.2
10.26 T	Tree-based methods: regression	ISL Ch 8.1
10.28 Th	Tree-based methods: classification	ISL Ch 8.1
11.02 T	Ensemble methods: bagging	ISL Ch 8.2
11.04 Th	Ensemble methods: random forests	ISL Ch 8.2
11.09 T	Ensemble methods: boosting	ISL Ch 8.2
11.11 Th	Support vector machines	ISL Ch 9
11.16 T	Support vector machines	ISL Ch 9
11.18 Th	Unsupervised learning	100ML Ch 6
11.23 T	Bias & fairness in learning systems	BS Ch 1 & MK
11.25 Th	Thanksgiving - no class	
11.30 T	Project presentations	
12.02 Th	Project presentations	
12.07 T	Final review	