STAT 4250: Statistical Learning, Winter 2021 (A01), CRN: 60145

Tentative Course Outline

Course Details

Course Title & Number:	Statistical Learning
Credit Hours:	3
Class Schedule:	11:30 AM–12:20 PM, MWF
Location for Lectures:	This is a Remote Learning course
Location for Tutorials:	Statistics Lab will be performed online
Course Material:	All course materials are posted on UMLearn (D2L) web
Pre-Requisites:	[one of STAT 3100, the former STAT 3600, or the former STAT 3800] and STAT 3150 and [STAT 3690 or the former STAT 4690].
Course Description:	Topics related to the use of Statistics and inferential methods in machine learning, including the lasso and ridge regression, classification and clustering, support vector machines, bagging, boosting, neural networks, and ensemble methods.
Lectures:	Live lectures and office hours will be conducted over Zoom. You do not need a Zoom account; you can simply access my lectures through the link that will be posted on UM Learn webpage. Lectures will not be recorded and students are strongly encouraged to attend the live lectures.

Instructor Contact Information

Instructor:	Mohammad Jafari Jozani
Office:	365 Machray Hall
Office Hours & Availability:	By appointment. Send me an email at least a day before to arrange a meeting using the Cisco Webex system or Zoom.
E-mail:	m_jafari_jozani@umanitoba.ca
	I will only respond to e-mail from UMNet ID's. When feasible, I normally return a call or an email within 24 hours.

Tutorial

Instructor:	Keith Uzelmann
E-mail:	uzelmank@myumanitoba.ca
Lab Time:	To be scheduled by Keith.

Attendance of lab tutorials is mandatory. During the tutorials, the TA will solve some selected theoretical and applied problems and will provide solutions to real data analysis using R. He will also be answering your questions. The first few Labs is dedicated to teaching the basics of R and R-markdown.

General Course Information and Course Registration

This course is a survey of statistical learning methods and covers major techniques and concepts for both supervised and unsupervised learning. Students will learn how and when to apply statistical learning techniques, their comparative strengths and weaknesses, and how to critically evaluate the performance of learning algorithms. The goal is to

(i) apply basic statistical learning methods and perform exploratory analysis,

(ii) properly select statistical learning models,

(iii) implement these methods using the R programming language,

(iv) correctly assess model fit and error,

(v) build an ensemble of learning algorithms.

Course Registration

It is **your responsibility** to ensure that you are entitled to be registered in this course. This means that you:

1. have the appropriate prerequisites, as noted in the calendar description, or have an appropriate permission from the instructor to waive these prerequisites;

2. have not previously taken, or are concurrently registered in, this course and another that has been identified as "not to be held with" in the course description.

The registration system may have allowed you to register in this course, but it is **your responsibility to check**. If you are not entitled to be in this course, you will be withdrawn, or the course may not be used in your degree program. There will be no fee adjustment. This is not appealable. Please be sure to read the course description for this and every course for which you are registered.

Textbook, Readings, Materials

I will have my own course notes. However, the main textbooks for this course are listed below. Other references will be suggested during the course if required. You can download these references as I describe below. Lecture notes will be available through the UM Learn system (see below). In this course I will be heavily using R for data analysis. I will provide a short note regarding R programming which is left to the students to make themselves familiar with it.

1. *An Introduction to Statistical Learning with Applications in R*. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Springer Texts in Statistics. ISBN 978-1-416-7138-7. NewYork (2013). E-book is available for download here https://www-bcf.usc.edu/ gareth/ISL/ISLR%20First%20Printing.pdf.

2. *The Elements of Statistical Learning (2nd Edition)*. Trevor Hastie, Robert Tibshiani and Jerome Friedman. Springer Series in Statistics. (2009). E-book is available for download using https://web.stanford.edu/ hastie/Papers/ESLII.pdf.

In order to prepare for class, I will normally ask you to read about the topics to be covered before the lecture. I am not expecting you to learn the material on your own, only to familiarize yourself with the main ideas and vocabulary so that the lectures are easier to follow. Do not get bogged down in formulae or minute details. If you come across something that is confusing or troubling, don't despair. If your questions are not resolved during the lecture, please ask. As you work on the problem sets, it will be helpful to re-read the material on a more detailed level.

Topics To Be Covered

Here is the outline of the course material (not necessarily in the same order that I will teaching in the class), which is subject to change, depending on time and class interests.

1. Introduction

- Brief overview of statistical learning concepts such as supervised and unsupervised learning with several examples (Regression versus Classification problems, supervised versus unsupervised learning, trade-off between prediction accuracy and model interpretability)
- Introducing methods for assessing the model accuracy
- A brief introduction to R programming (in the Lab)

2. Supervised Learning

- An overview of linear regression: Univariate, multivariate and multiple linear regression.
- Regression versus K-nearest neighbours.
- Moving beyond linearity by working with Polynomial regression and regression Splines.
- Ridge regression, regularized regression and some elated topics.
- Perceptron and Support Vector Machine
- Resampling Techniques such as the Cross-Validation and Bootstrap.
- Classification using the logistic regression, LDA, QDA, finite mixture models and Naive Bayes approaches.
- (if time permits) Trees, Boosting and Neural Networks

3. Unsupervised Learning

- Clustering methods such as K-means clustering, Hierarchical clustering
- Understanding dendrograms and different similarity and dissimilarity measures.
- Dimension Reduction techniques such as principal component analysis (if time permits)

Course Technology

Course web-page: Course materials will be made available through the University of Manitoba's UM Learn system (`umanitoba.ca/d2l`).

Software: We will extensively be making use of the R statistical software. R is freely available for Linux, Macintosh and Windows from *The Comprehensive R Archive Network* at `http://cran.r-project.org/`. Please download and install.

Other Technology: It is the general University of Manitoba policy that all technology resources are to be used in a responsible, efficient, ethical and legal manner. Students should restrict their use of technology to those approved by the instructor and/or University of Manitoba Accessibility Services for *educational purposes only*. Electronic messaging, e-mail, social networking, gaming, etc. should be avoided during class time. Cell phones should be turned off. If a student is on call for emergencies, his/her cell phone should be on vibrate mode and the student should leave the classroom before using it.

Important Dates

These dates are tentative and subject to change at the discretion of the instructor and/or based on the learning needs of the students but such changes are subject to Section 2.8 of the ROASS Procedure.

Date	Information	Date	Information
January 8	Classes Begin	March 31	Last Day for VW
February 15	Louis Riel Day	April 2	Good Friday
February 16–19	Winter Term Break	April 16	**Project Report Deadline**
February 20	**Project Proposal Deadline**	April 16	End of Classes
February 26	**Term Test**	April 19– May 1	Final Exams Period

Course Work, Examinations & Grading

Midterm and Final Exams: This course does not have a final exam. You will be having ONE term test worth 20% of your final grade. The tentative date for your term test is **February 26, 2021** from 11:30 to 12:30. Test content is defined by the lecture notes. **There will be no make-up test.** If you miss the term test with a valid reason and inform me within 24 hours, the weight of the test will be shifted to the final project. As the mid-term exam will be online, each student must submit his or her own copy of solutions, including comments, discussions and interpretations. **For any document submitted online you need to confirm in writing that submitted solutions are your own work and you have not cheated and/or consulted with anyone, or used any sources other than your course notes.** Note that actions will be taken against students who are found guilty of acts of academic dishonesty.

Assignments: Assignments worth 35% of your final grade (5 assignments each worth 7%). Assignments are due at the start of the class (time will be announced). Each assignment involves three parts (theory, application and simulation). Assignments submitted late will be severely penalized. Assignments submitted after the solutions are posted or after the graded assignments are return to students will not be marked and receive a grade of 0. Obviously, exceptions can be made to the above policy if special/exceptional circumstances warrant them (e.g., serious illness). Students are encouraged to discuss and work together on the solutions to the assignments. However, each student must hand in his or her own copy of each assignment with personalized solutions, including comments, discussions and interpretations. Note that actions will be taken against students who are found guilty of acts of academic dishonesty.
Your assignments should conform to the following standards:

- Theoretical part of the assignments are to be done on 8.5×11 paper, scanned and submitted as a high quality PDF file
- Name the file you submit with your name and student ID.
- Applied and simulation parts of each assignment that involve R programming should be accompanied with the R codes and results should be reproducible. I do encourage you to use R-markdown to hand in your R assignments.
- Revise your assignments so they are reasonably free of grammatical and typographical errors.
- Make sure each step in your solutions is well justified: I mark what is written on paper and should not have to guess what you mean.
- Messy or unreadable assignments will be returned with a mark of zero.

Real Data Analysis Projects There will be a real data project worth 45% of your final grade. Students should analyze a real data of their own choice from the UCI Machine Learning Repository at `https://archive.ics.uci.edu/ml/datasets.php` using the techniques covered in the course. Students should choose a more recent dataset that is published on the UCI website (i.e., not before 2016). Then **a proposal should be submitted to me by February 20, 2021** outlining the details of the data set, motivation of the study and a tentative plan for the analysis. After your plan is approved by myself (with or without revision) then you can start working on your project and completing the analysis while we are going through the course materials. Final reports should be prepared in Rmarkdown and in the PDF format. **The due date for submitting your final report is April 16, 2021**. More details regarding the data project will be submitted on D2L.

Your report should conform to the following standards:

- Be sure that you explain as clearly as possible the connection of your project and the concepts you learned from class.
- Your report should have a motivation and a quick summary of the problem.
- Real data analysis should be accompanied with the R codes and I should be able to get your answers by running your codes. If your R code does not work you will not get any mark. You are highly encouraged to your Rmarkdown to prepare your homework solutions.
- Revise your report so they are reasonably free of grammatical and typographical errors. Messy or unreadable report will be returned with a mark of zero.
- Make sure each step in your analysis is well justified. I mark what is written on paper and should not have to guess what you mean.
- Each report should have a conclusion section that includes comments on the meaning of the results and open questions.

Grading Scheme: The following are the minimum percentage grades required to receive each of the various letter grades: A+ (90%), A (80%), B+ (75%), B (70%), C+ (65%), C (60%), D (50%).

Item	Percent
Mid-Term	20%
Assignments	35%
Read Data Project	45%
Total	100%

Class Communications

The University requires all students to activate an official U of M email account, which should be used for all communications between yourself and the university (including all your instructors). All these email communications should comply with the University's policy on electronic communication with students, which can be found at: `http://umanitoba.ca/admin/governance/governing_documents/community/electronic_communication_with_students_policy.html`

Using Copyrighted Material

Please respect copyright and we will use copyrighted content in this course. All course notes, assignments, tests, exams, practice exams and solutions are the intellectual property of your instructor or the Department of Statistics. Reproduction or distribution of these materials is strictly forbidden without their consent. For more information, see the University's Copyright Office website at http://umanitoba.ca/copyright/orcontactum_copyright@ umanitoba.ca.

Recording Class Lectures

Mohammad Jafari Jozani and the University of Manitoba hold copyright over the course materials, presentations and lectures which form part of this course. **No audio or video recording of lectures or presentations is allowed in any format, openly or surreptitiously, in whole or in part without permission of Mohammad Jafari Jozani**. Course materials (both paper and digital) are for the participant private study and research. If class recordings are provided by the instructor those are meant to be for your own personal use only. **It is not permitted to copy or distribute the recordings.**

Student Accessibility Services

If you are a student with a disability, please contact Student Accessibility Services (SAS) for academic accommodation supports and services such as note-taking, interpreting, assistive technology and exam accommodations. Students who have, or think they may have, a disability (e.g. mental illness, learning, medical, hearing, injury-related, visual) are invited to contact SAS to arrange a confidential consultation.

Student Accessibility Services, http://umanitoba.ca/student/saa/accessibility/
520 University Centre, (204) 474-7423, Student_accessibility@umanitoba.ca

Academic Integrity

It is important that you understand what constitutes academic dishonesty and that you are familiar with the very serious consequences. Links to resources that describe academic dishonesty (including plagiarism, cheating, inappropriate collaboration and examination impersonation, as well as typical penalties) can be found at:

`http://umanitoba.ca/faculties/science/undergrad/resources/webdisciplinedocuments.html`
or
`http://umanitoba.ca/faculties/science/undergrad/resources/webdisciplinedocuments.html`

ROASS Schedule A

Schedule "A" of the *Responsibilities of Academic Staff with regards to Students (ROASS)* policies of the University of Manitoba lists resources and policies for students. It is important that you familiarize yourself with these resources and policies. This document will be posted to the Department of Statistics web page and to the UM Learn system.

`http://umanitoba.ca/science/statistics/files/pages/2016/09/Schedule-A-ROASS-Statistics.pdf`