ECON 3838 - 4848

Introduction to Applied Machine Learning with R

Winter 2020

CRN 26302

Instructors: Dr. Yigit Aydede **Office:** SB-343

Labs: Weiyi Li (wy785675@dal.ca)

Email: yigit.aydede@smu.ca **Office Hours: MW** 2-4pm.

Phone: 420-5673 **Web page:** Brightspace at SMUPort

Day, Time & Location of Classes: MW 4:00 – 5:15 pm, Sobey 155
Lab's: MW 5:30 – 6:45 pm, Sobey 155

Credit Hours: 3.0
Course Prerequisite(s): ECON 3303

Course Description

This class is an introductory undergraduate course for machine learning in social sciences with applications in R. As opposed to many economic applications that revolve around *parametric estimation* of $\boldsymbol{\beta}$ that underlies the relationship between **y** and **x**, machine learning revolves around the *problem of prediction* that uncovers generalizable patterns by discovering unobserved **y** from observed **x**. Using tools in statistical learning, the class will mainly focus on "supervised" machine learning and briefly cover "unsupervised" methods for understanding complex big datasets in economics and other social sciences. We will study "systems" that require "learning" algorithms built in penalized linear and logistic regressions (Lasso, Ridge, Elastic Net, Adaptive Lasso) and non-regression learning techniques (Decision Trees, Random Forest and Nearest Neighbors). We will use R, a free software package for statistical computing, which will be introduced and gradually developed to more advance algorithms as you learn by applications on big and complex datasets during the semester.

Machine learning is an exciting interdisciplinary field across statistics, computer science, neuroscience, and philosophy. The objective of this course is to give students enough understanding about the practical applications of statistical learning from big data in developing better strategies in business, public policy and social sciences.

Bachelor of Commerce Program Goals

Critical thinking – Graduates critically assess situations and use appropriate problem-solving skills. Students will:

- Formulate and justify positions on issues or situations using adequate and appropriate evidence
- Recognize and demonstrate competence in analytical reasoning

Information literacy – Graduates locate and use qualitative and quantitative information effectively using appropriate technology. Students will:

- Determine, retrieve, evaluate and manage relevant information
- Recognize and acknowledge copyright laws and intellectual property restrictions

Communication – Graduates use professional communication skills to facilitate business relationships. Students will:

- Prepare appropriate and effective written communications
- Prepare and deliver effective oral presentations

Ethics, corporate social responsibility, and leadership – Graduates exercise socially responsible leadership skills. Students will:

- Recognize the ethical dimensions of situations
- Consider a broad spectrum of stakeholders in the development of organizations' social responsibilities
- Recognize a variety of leadership styles and when each is appropriate
- Work effectively as part of a team

Global perspective – Graduates recognize the diversity and global opportunities their local, national, and world environments present. Students will:

- Recognize the effects of different economic, political, cultural, social and technological environments and integrate them into their decision-making

Business knowledge and competency – Graduates use their business knowledge and professional skills successfully. Students will:

- Demonstrate a fundamental understanding of each of the functional areas of business, and how to effectively integrate and apply this knowledge
- Assess the opportunities and risks faced by organizations of different size, ownership and governance structures

Course Objectives

After completing this course, if you have attended class regularly, read the assigned material, and applied the teaching presented in the course, you will be able to understand:

- The difference between estimation and prediction processes,
- Overfitting problems and variance-bias trade-off in predictions,
- Basics of nonparametric estimations,
- Regression and classification models,
- Developing a learning system by grid search and cross-validations,
- Performance evaluations for predictive models in classification,
- Regularization with LASSO families,
- K-Nearest Neighbors (kNN),
- Classification and Regression trees,
- Random Forest, Bagging and Boosting,
- Model selection with CARET,

- Naïve-Bayes in Text Mining.

Instructional Approach

We will have a combination of 2 lectures and 2 labs each week. We will use **R,** which is a free statistical program and also available at the campus.

How to succeed in this course?

- **Lecture notes are APPLIED. Any code/script that we apply in the class, you will have them in lecture notes. Hence you will be able to STUDY them after the class. Please before and after you spend some time on them to increase your understanding!**
- **Learning is a struggle**. Expect more confusion and less progress in each lecture. More practice and genuine focus will get you to the end. Study regularly. Your attendance will improve your understanding on the course. I will not follow the book word by word. If you miss a class you will definitely miss some valuable information which may or may not be found in textbooks.
- I'll give you some practice questions but you find many sources online for more practice. The assignments are good ways to understand the subject. Spend enough time to solve them and then compare your results with the posted answers.
- During the lectures I mostly use the board and my lecture nots in rmd format **not Power Point slides**. You are expected to listen and apply some of the codes/scripts.
- Try to understand the intuition not the mechanics first! If you believe that you have difficulties in understanding some concepts use my office hours. These are your hours. Besides, I have an open-door policy. As my time permits, I would love to see you in my office asking questions.
- Instructions about submitting the project will be given in advance. The mark will reflect how close the submission follows them.
- Each assignment will be given in advance and a set of instructions. The format of submission will be marked as well.

Required and Recommended Texts and/or Materials

- **An Introduction to Statistical Learning with Applications in R** by James, Witten, Hastie, and Tibshirani. It is a free book available on the Internet. http://www-bcf.usc.edu/~gareth/ISL/
- *The first draft of* **my own book**. You will be the first reader of its first draft. I'll give you all and every chapter (lecture files) in advance. Lab files will be given after the labs.

Evaluation

Midterm (each 25)	25% (Monday, 3/9)
Final Project	25% (The date and the place will be announced later)
7 - 10 Assignments.	50% (The weights will not be same. Dates will be determined)

The numeric grades you earn in this course convert to a letter grade as follows:

A+	90-100	B+	77-79	C+	67-69	D	50-59
A	85-89	B	73-76	C	63-66		
A-	80-84	B-	70-72	C-	60-62	F	0-49

Please refer to the Undergraduate Academic Calendar for related Grading System policies and procedures.

Academic Policies

Academic Integrity Policy and Student Responsibility

As a Saint Mary's University student, you are responsible for understanding and avoiding academic offences, including plagiarism, cheating, and falsification. Working with another person (or in a group) when individual work is required for a grade is considered a form of academic dishonesty.

Plagiarism is when you present someone else's words, ideas or techniques as your own. When you want to refer to someone else's work, you must reference it either by direct quotation or paraphrase (expressing the idea in your own words), which must be acknowledged using correct citation. When you are in doubt about what, when, and how to cite your information sources, consult with me, or the Writing Centre, before submitting your work. Academic dishonesty is a serious offense, so be sure you understand how to correctly acknowledge and use sources when preparing your work.

Plagiarism applies to all forms of information or ideas that belong to someone else (e.g., literary works, computer programs, mathematical solutions, scientific experiments, pictures, website or data).

Furthermore, submitting the same piece of work (even if it's your own) for a grade or credit in more than one course is usually not permitted. The approval of the course instructors involved must be obtained before submitting the assignment.

If you are found in violation of this academic honesty policy, your work is subject to a grade of zero, and you will be reported to the Registrar. If the Registrar finds this is a second offence, you will be referred to the university's Academic Honesty Committee and subject to disciplinary action.

Please read the entire "Academic Integrity" section (Academic Regulation 19) of Saint Mary's Academic Calendar for a complete description of each offence, noting especially the examples of plagiarism and penalties.

Late Assignments, Missed Tests, Mid-term Exams, or other Due Dates

All assigned tasks, including class presentations, must be completed as scheduled. Failure to do so will result in penalties, or a zero mark. There will be no make-up tests/or project. If you miss a project for a valid reason (with a University recognized excuse and proper documentation), the weight will be added to the final project. You will be given problem sets.

Senate Policy on Final Examinations

The Senate policy on final examinations can be read here:
http://www.smu.ca/webfiles/8-1016_SenateRescheduleExams.pdf

Special Examinations

We are guided by Academic Regulation 10 found in the Academic Calendar.

> ***10. Special Examinations***
> ***A student who, due to a serious illness or emergency, was detained or rendered unfit to write a required final examination, may appeal for a special examination. Elective***

arrangements (such as travel plans) are not considered acceptable grounds for granting an alternative examination time. A student who wishes to have such an appeal considered must:

a. Within 48 hours after the end of the final examination, report, or have a representative report (in writing if possible), to the Instructor and the Dean of the Faculty in which the course is offered, intention to appeal for a special examination and

b. Within one week after the end of the examination submit to the Dean a written request for a special examination. This request must be accompanied by an explanation of the circumstances which made it impossible for the student to write the regular examination and a medical doctor's report, or other document, which supports the appeal.

c. The Dean's decision will be communicated to the student and the instructor within one week of receipt of the request.

d. If the Dean approves the special examination, the responsibility for setting and conducting special examinations will lie with the instructor and the department. Special examinations should be completed as soon as possible and normally (i.e. wherever possible) as follows: for first term by Jan 31, for the second term by May 31, for summer session one by July 31 and for summer session two by Sept 30.

The Sobey School of Business has an "Appeal for Special Examination" form which is available at the front desk of the BComm Advising Centre (Sobey 252) or via email request at sobey.bcomm@smu.ca. As per paragraphs a. and b. above, this form is to be used to request an appeal for a Special Examination. Supporting documentation should also be included. The completed form and supporting documentation can be delivered to the BComm Advising Centre or emailed to sobey.bcomm@smu.ca.

Once the form has been received, a committee designated by the Dean will review the appeal request and the supporting documents as well as contacting the respective instructor(s) before informing the student of the final decision.

Communication devices during examinations

Students will need to turn off and store ***all communication devices*** such that they cannot be seen or accessed during the examination. Holding or using a cell phone during an examination is strictly prohibited and is an academic dishonesty offence under the Academic Regulation 19 – Academic Integrity.

Tentative Schedule (Chapters follow ISLR)

Note that we can change the order during the semester. The plan is to cover each chapter about in a week!

1. *Introduction to Machine Learning and the course* (Ch 1)

   Statistical learning: Supervised vs. Unsupervised learning

2. *Review on Statistics – 2* (Ch.2-3 & Lecture 2 Notes)
   a. Data types, dataset types, and descriptive statistics,
   b. Plots,
   c. Probability distributions

    d. Regression with OLS and MLE

3. *A Formal Look at Learning: Variance-bias tradeoff* (Ch 2 & Lecture 3a Notes)
    a. Estimator and MSE,
    b. Predictor and MSPE,
    c. Uncertainty in prediction,
    d. Dropping a variable in regression,
    e. Prediction interval,
    f. Overfitting and model selection
4. *Prediction in Classification* (Ch.4 & Lecture 4 Notes)
    a. Linear Probability Models
    b. Logistic Regression
5. *Nonparametric Estimations* (Ch 7 & Lecture 5 Notes)
    a. Density estimations,
    b. Kernel regressions
    c. Regression splines,
    d. MARS,
    e. GAM
6. *Grid Search and Cross-Validation* (Ch 5 & Lecture 6 Notes)
    a. Splitting the data: training and test sets,
    b. k-fold cross-validation,
    c. Grid search,
    d. Cross-validated grid search,
    e. Time-series data,
    f. Caret.
7. *Classifications with Caret and kNN* (Ch 7 & Lecture 7 Notes)
    a. Linear classifiers: LPM and logistic regression,
    b. kNN smoothing,
    c. Application with Caret,
8. *Classification and Goodness of fit* (Lecture 8 Notes)
    a. Confusion table,
    b. Evaluation metrics,
    c. Receiver Operating Characteristics Curve (ROC),
    d. Area Under the Curve (AUC).
9. *Tree-based models* (Ch 8 & Lecture 9 Notes)
    a. CART – classification Tree,
    b. CART – Regression Tree,
    c. Bagging and boosting – Random Forest,
    d. Ensemble models.
10. *Regularization with LASSO* (Ch 6 & Lecture 10 Notes)
    a. Ridge,
    b. Lasso,
    c. Adaptive Lasso,
    d. Elastic Net.
11. *Text Mining* (Lecture 11 Notes)

a. Naïve-Bayes models,
b. Spam removal example.

Tentative Lab Schedule: (Lab sessions will be adjusted in line with the progress in lectures)

LAB1: *Intro to R and R Markdown*

- R, RStudio, and R packages
- Starting with RStudio,
- Working directory,
- Data types and Structures,
- R-Style Guide

LAB2: *Data frames and Programming basics*

- Lists, Data frames, Reading and writing data files, Subsetting data frames, Plotting, Categorical variables,
- Programming basics: if/else, loops, apply() family, functions.

LAB3: *Modeling with R*

- Regression with lm(),
  - Data preparation,
  - Factor variables,
  - Dummy coding,
  - Column names,
  - Data subsetting and missing values,
  - Dummy Variable Models,
  - mtcar example,
  - model.matrix(),

LAB4: *Simulations with R*

- Random sampling,
- Random number generating,
- Simulation for inference,
- DGM,
- Bootstrapping,
- Example

LAB5: *Examples for Prediction in classification*

- Boston Housing example,
- Coronary Heart Disease,
- SPAM.

LAB6: *Use of Nonparametric Methods*

- Smoothing,
- kNN,
- Kernel smoothing,
- Loess(),
- Multivariate loess().

LAB7: *Prediction in Classification with Training*

- LPM,

- Logistic,
- kNN,
- Loess(),
- Multivariate loess().

LAB8: *Algorithmic Optimization*

- Brut-force optimizations,
- Derivative-based methods,
- Gradient descent,
- Numerical optimizations with R.

LAB9: *Tree-based models*

- CART,
- Bagging, boosting, Random Forest,

LAB10: *Regularized regressions*

- Lasso family examples with big data high-dimensional,

We might add/drop some topics as we proceed