STAT 452/652: Statistical Learning and Prediction

Syllabus

Fall, 2020

Instructor: Tom Loughin (Call me "Tom")

Contact: Office: HOME! (Thanks, Covid-19...) Phone: NA E-mail: tloughin@sfu.ca

Class Time: All course elements except final project presentations are remote, asynchronous recordings (you view lecture materials when you want to).

Questions: Instead of office hours, I will hold online live question-and-answer (QA) sessions Mon 17:00-18:00, Thu 11:00-12:00.
The QA sessions will be held virtually, likely on Zoom, and are intended to address questions about course and assignment (statistical) content. All students are welcome to attend, and questions will be answered publicly. I will try to remember to record the sessions and post them for later viewing.
If you have a more personal matter to discuss, please contact me privately by e-mail. However, please note that I *cannot* answer statistical course content questions or homework questions over e-mail. There are too may students to make that possible.

Discussion: I have set up two discussion boards in Canvas. One will mainly be for questions regarding clarifications of notes or assignments, which might also be useful for other students. I will check the board periodically, but I also welcome replies from other students. Participation is voluntary and not graded. Answers are generally public. The other is for questions about course procedures and conduct, not relating to statistics. Again, these will be public.

Book: *An Introduction to Statistical Learning with Applications in R* (ISLR) by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013). New York: Springer. It is available online through the library and also for free (LEGAL!) download here:
http://faculty.marshall.usc.edu/gareth-james/ISL/

Prerequisite: This course is a required course for the Data Science major and an elective course for the Statistics and Actuarial Science majors. It is also an elective for the Statistics minor. The prerequisites are STAT 302, 305, 350, or BUEC 333.

All of these are "Stat-2" courses focusing on regression analysis.
Basically, I expect you to have had a course beyond an intro stat course that taught you regression and a little bit of ANOVA.

Computing: We will use **R** for all statistical computations in this class. I assume that all students have had some previous exposure to R. Students who plan to use R but have never used it before should begin getting familiar with it *immediately*. (I suggest RStudio as a front-end for R. Students not wanting to deal with a free software download could use Jupyter notebooks through https://sfu.syzygy.ca/.) I will provide example programs for each different kind of analysis we learn; these will provide a template for students to work from on their homework and in their futures!

Logistics: Each week I will post lecture notes in pdf to Canvas and recorded lectures that explain these notes. The lectures will describe and explain the statistical analysis methods, and there will be a few short examples with bits of R code. The lecture notes will have exercises at the end. The exercises relating to the lectures covered each week will be due on Friday of the following week, so there will be a predictable assignment schedule.

As well, tutorials will be posted each week to support the lectures from that week. These tutorials will dig deeper into examples and demonstrate R code step-by-step. The tutorials will not teach the methods, so they will assume that you have already seen the lectures. It is my expectation that students will view both lectures and tutorials before attempting homeworks or answering questions.

GRADING and POLICIES

- Exams: There are no midterms. Students in 452 will have a final, the date for which is yet to be announced. Grad students in 652 will have a final project instead.
- Homework: There will be about 10 graded homework assignments. Essentially, there will be something due *each week*. Assignments will consist mostly of applications and simulations. We will use Crowdmark and Canvas for submitting homework assignments. Instructions will be given later when the first assignment is due.
- Projects:
  - All students will do two main prediction projects, one on regression and one on classification. Both will be in the form of a competition. Students will provide predicted values for sets of data that I provide and write a short report on the process that they took to create the predictions. *Grades will be assigned partly competitively, based on the performance of students' prediction models!*
    - For these projects, 452-D100 (non-ACMA) students will be assigned a partner at random at least a week prior to the due date. If a partnership is not working out, the students can contact me at least three days prior to the due date and

both members will a project alone instead. Students will not have the option of finding another partner. Details will be given later when the projects are formally assigned.

  - Students in 652 will do another project in lieu of a final exam. They will be responsible for locating an appropriate paper or other source for data and performing appropriate analyses on data. They will provide a written report and a presentation on their results. Topics will be need to be approved by me in advance.
  - We will talk about both projects more as the semester progresses.

- Teams: Homeworks may be completed either individually or in pairs. There is no bonus or penalty associated with either choice. You may choose your own partner. Each team submits one assignment in Crowdmark as a group submission. Instructions will be given later. Both students receive the same grade for the assignment. (Complaints about non-contributing team members, or errors committed by team members, will be ignored. You have the choice do the assignment alone!) Each member of a team is expected to know and understand everything that the team turns in. **No pair may produce more than ONE assignment together. Students may complete at most 3 assignments in a team (there will be approximately 10 assignments). Students who do more than one if the same pair or more than three in a team will get zeroes for all assignments beyond the allowance.**

- Collaboration: Discussion of ideas learned in class is highly encouraged. This often helps in the learning process. Discussion of assignments, however, must be done carefully. Because assignments count toward a student's mark, the work turned in by each student should represent that individual's understanding, ideas, and creativity. By submitting an assignment for marking, you assert that the work contained therein is fundamentally your own. In particular, while students may discuss with one another techniques and approaches to solving problems (only after trying to solve it alone, of course), students should not get answers to problems—or the R code for solving them—from another source, within or outside the class. This constitutes plagiarism and is covered by university policy:
  https://www.sfu.ca/students/academicintegrity/students.html.
  Note that anyone who provides answers to another student is also guilty of academic dishonesty, and will be held accountable for these actions. If you have questions about collaboration, please contact me before it becomes a problem!

- Late Homework: Homework is considered late if it is not received before the specified due time. Homework handed in later the same day may be subject to a penalty of 20% of the homework's value and on later days at an additional 20% per day. Homeworks will not be accepted after they are marked and returned to the class.

  - I always make allowances for emergencies. Talk to me as soon as possible if a problem arises. Expect to offer some evidence

  - I will also usually give someone a day or two break if they are just swamped at the moment or an assignment is taking longer than expected. My patience will wear thin, however, if a student asks for multiple breaks. That signals poor time management or refusal to prioritize the course and get assignments done on time.

- Marking Errors: TAs try to be consistent during marking but sometimes mistakes still happen. Also, students sometimes disagree with the proportion of marks granted for partially correct solutions. Dealing with weekly appeals for marking changes is time consuming and often pays little in return. I therefore want to substantially reduce the effort that you, the TA's, and I must put into marking appeals. As an allowance for mistakes that happen over the course of the term, all students in the course will receive 2% bonus marks added in at the end of the term. In exchange, students will not request extra marks on assignments and midterms. (Exceptions will be made for errors in computing the final score on the assignment). o If a student truly feels that a huge error has been made, they may forfeit these 2 free percentage points and request that their appeal be considered. The 2 points will NOT be returned to the student, regardless of the outcome of the appeal. o The appeal process will be posted later.

- Missed Projects and Homeworks: Students are responsible for meeting class deadlines. Assessments that are received in tome to be marked will count as zeroes except in the case of illness or other emergencies, in which case students are expected to provide documentation (e.g. Doctor's note, funeral notice) to support the necessity of missing the deadline. Even then, there is lots of question about the legitimacy of these "excuses".

  - In 2016, three students missed my first midterm. All three gave me a vague doctor's note, two from the same doctor, and all were received 2-5 days after the midterm. Those same three students were the only three students who had not yet turned in a homework. Since I am a statistician, I can tell you that the chances that exactly these three students out of 76 would be the three who missed the exam is 1/70,300 (i.e., p-value is 0.00001 for a test of the null hypothesis that students' previous homework status was not related to missing the midterm). Not likely in a real world.
  - In order for your request for excused absence to be considered, you must notify me as soon as you know that you will miss the assessment. This usually means before the exam or due time. If you wait until later tell me that you missed something, I will expect clear evidence that you were so incapacitated that you could not possibly have used your cell phone to send an e-mail. If you don't have that, you get a zero. Period.
  - Unlike other instructors, I do not move the weight of missed assessments onto the final. Every instructor has seen students who come up with excuses to miss most of the class, so that they can just pass the final to pass the course. Rarely does it work. So I don't allow it. Your final exam or project will be worth no more than what is stated.

- If you miss a midterm or an assignment, I will assign an "imputed value" to you based on how well you do on the rest of the marked elements in the class. If you are a median

student on average, you will be assigned a median score for the missing element. If you are a top student in other elements, you will get a top score. If you are a bottom student, you will get a bottom score.

- Scoring (This is subject to change):

  Assignments 10% (20% for 652)

  Project1 25%

  Project2 25%

  Final 40% (30% for 652)

- Grades will be assigned as follows:

  - STAT 452:

    A+: 95.00+ %
    A: 90.00–94.99%
    A-: 85.00–89.99%
    B+: 80.00–84.99%
    B: 75.00–79.99%
    B-: 70.00–74.99%
    C+: 66.00--69.99%
    C: 62.00-65.99%
    C-: 58.00--61.99%
    D: 50.00-57.99%
    F: 0–49.99%

  - STAT 652:

    A+: 95.00+ %
    A: 90.00–94.99%
    A-: 85.00–89.99%
    B+: 80.00–84.99%
    B: 75.00–79.99%
    B-: 70.00–74.99%
    C+: 65.00--69.99%
    C-: 60.00--65.99%
    F: 0–59.99%

  Students who wish to receive a grade for the class must be enrolled in time to be placed on the class roster. Under no circumstances will credit for the class be given to a student who is not enrolled.

ANTICIPATED COURSE OUTLINE (SUBJECT TO CHANGE!)

1. INTRODUCTION: STATISTICAL LEARNING AND PREDICTION
2. MEASURING PREDICTION ERROR
   (a) Basic concepts of bias and variance
   (b) Training vs generalization error
3. LINEAR REGRESSSSION ESSENTIALS AND EXTENSIONS
   (a) Review of Least Squares
   (b) Residuals and model fit
   (c) Different forms of multiple linear regression
4. VARIABLE SELECTION IN LINEAR REGRESSION
   (a) Traditional methods and their problems
   (b) New approaches to variable selection: Shrinkage, regularization, penalization
5. NONLINEAR REGRESSION METHODS
   (a) Splines and smoothing
6. TREES AND ENSEMBLES
   (a) Bagging
   (b) Random Forests
   (c) Boosting
7. CLASSIFICATION
   (a) Basic problem and historical approaches
      i. Linear Discriminant Analysis
      ii. Logistic Regression
   (b) Nonparametric methods
      i. Trees
      ii. k nearest neighbours
   (c) Support Vector Machines
   (d) ROC Curves
8. DIMENSION REDUCTION TECHNIQUES
   (a) Principal Components
   (b) Clustering TOPICS COVERED:

Learning Outcomes

Upon successful completion of this course, you will be able to:

1. Use several modern techniques of regression, classification, and clustering, including splines, tree ensembles, and support vector machines;
2. Apply methods for variable selection;
3. Distinguish the importance of bias and variance in prediction;
4. Use programs in R to complete many forms of modern statistical analysis