BINF*6970 Statistical Bioinformatics

Winter 2020
Section(s): C01

College of Biological Science
Credit Weight: 0.50
Version 1.00 - December 18, 2019

---

1 Course Details

1.1 Calendar Description

This course presents a selection of advanced approaches for the statistical analysis of data that arise in bioinformatics, especially genomic data. A central theme to this course is the modelling of complex, often high-dimensional, data structures.

Pre-Requisites:	Introductory courses in statistics, mathematics and programming
Restrictions:	Restricted to students in Bioinformatics programs. Students in other programs may consult with course instructor.

1.2 Timetable

Timetable is subject to change. Please see WebAdvisor for the latest information.

1.3 Final Exam

Exam time and location is subject to change. Please see WebAdvisor for the latest information.

---

2 Instructional Support

2.1 Instructor

Khurram Nadeem, PhD

Department of Mathematics & Statistics

Email: nadeemk@uoguelph.ca

Office: MacNaughton 517

Times & Venue: Tuesday and Thursday 10:00-11:20 pm in SSC 1306

Office Hour: Wednesday 1:00-2:00 pm and by appointment.

2.2 Teaching Assistant

Jessmyn Niergarth

Email: jniergar@uoguelph.ca

Office Hour: TBA.

---

3 Learning Resources

3.1 Text

Lecture notes and assigned research articles. In addition, selected readings from the following books will be recommended to complement lecture notes. These resources are available directly online or as a PDF book for download through the University of Guelph library website: (https://www.lib.uoguelph.ca/).

1. Hastie, T., Tibshirani, R. & Friedman, J.H. (2009) The elements of statistical learning: data mining,

inference, and prediction. Springer Series in Statistics, 2nd edition. New York, NY: Springer.

2. James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) An introduction to statistical learning

(Vol.

112). New York: Springer.

3. Wickham, H. & Grolemund, G. (2017) R for Data Science. O'Reilly. (http://r4ds.had.co.nz/).

3.1 Software

All computational examples in the course (lectures, exams) will be done in the statistical language R, which is an open source software and is available for installation at:

https://cran.r-project.org/. We will mostly be using RStudio, an open-source integrated development environment for R and freely available at:
https://www.rstudio.com/.

Note: Please bring your own laptop to every class for hands-on practice and software implementation of the methods learned in class.

3.1 CourseLink

Course information and material (such as assignments, data sets, etc.) will be available on CourseLink. Students are responsible to check the website regularly for updated information and announcements. We mostly put stu on CourseLink for this course, but emergencies and big changes may get to you first via the university e-mail. It is equally important to check your e-mail regularly.

---

4 Learning Outcomes

The course covers advanced topics in statistics and data mining that arise during the analysis of bioinformatic data. The course will emphasize but not solely focus on genomic data. This course will use the language and packages for demonstration and analysis purposes.

4.1 Key Topics

Introduction to R computational environment. Principles and guidelines for statistical

computing and graphing. Aspects of statistical modeling. Linear and generalized linear models. Logistic regression and its use in classification. Family-wise error control in multiple hypothesis testing. Variable selection via LASSO – bias variance trade-off, bootstrapping and cross-validation.

Multivariate normal distribution. Principal components and their extensions. Classification techniques – Linear and quadratic discriminant analysis; KNN, SVM, classification trees; random forests. Clustering and dimension reduction techniques – hierarchical and non-hierarchical clustering, multidimensional scaling. Neural networks.

As the above-mentioned topics can be widely applicable to biological data analysis such as gene expression data, sequencing data, as well as genetics data; this course will also cover topics in statistical genetics, one of the main themes in bioinformatics.

Note: The final coverage may not include all these topics depending on time and other factors.

---

5 Teaching and Learning Activities

5.1 Lecture

Mon, Jan 6 - Fri, Apr 3

Topics:

Weekly Lectures	Topics Covered	Not
Week 1-2	Knowledge Discovery and Statistical Data Mining in Bioinformatics; Exploratory Data Analysis in R; Properties of Vectors and Matrices; Introduction to Multiple Linear Regression	n/
Week 3	Model validation and Selection in Linear Regression Models; Bootstrap and Cross-Validation	n/
Week 4	Bootstrap and Cross-Validation;	**Assignr**

	Regularization in Linear Models - Lasso	Due Friday
Week 5	Regularization and Cross-validation; Logistic Regression Model	n/
Week 6	Lasso-logistic regression; classification via supervised machine learning; KNN classification	**Assignm** Due Friday
Week 7	n/a	**Winter Br** Classes S
Week 8	Regression Trees; Random Forests	**Midt** Thursday,
Week 9	Dimension reduction - PCA	n/
Week 10-11	Statistical Genetics	n/
Week 11-12	Distance Measures and Clustering Techniques	**Assignm** Due Tuesda
Week 12-13	Cluster analysis; Support Vector Machines	n/
Week 13	Gradient Boosting; Artificial Neural Networks	**Assignm** Due Frida

---

6 Assessments

6.1 Marking Schemes & Distributions

Homework 40% (Four assignments, equally weighted)

Homework Due Dates: Assig. 1 (**Fr, Jan 31**); Assig. 2 (**Fr, Feb 14**);

Assig. 3 (**Tu, Mar 17**); Assig. 4 (**Fr, Apr 3**)

Midterm Exam 20% (Thursday **Feb 27**)

Final Exam 40% (Time, location TBA)

6.2 Exams

Midterm and final exams are take-home – specific instructions will be explained in the class and via CourseLink. They will aim at testing the understanding of the concepts and techniques covered in the course, mostly via interpreting the computer output or answering questions of a consulting nature. The most effective preparation for the exams is critical rethinking of topics: what is the objective of the method? How exactly the method is performed? Are there any underlying assumptions? Are there any modifications or extensions?

6.3 Homework

Assignments will consist of analyses of selected datasets. Students will turn in (typed) reports on these analyses, together with relevant graphics and conclusions, in plain language, avoiding explicit programming code and irrelevant material. Programming code for each assignment is required and will be relegated to the appendix. The evaluation will be based on the following criteria: amount and depth of the analysis, correct use of the statistical methods, correct and logical interpretations of the outcomes of the analyses, clarity and professional appearance of the text and graphics. The length of the text will not matter, and may be considered even as a negative factor, if overly excessive without a good reason.

6.4 Collaboration

While you are encouraged to discuss approaches to assignment questions with other students, the material turned in must be your own. Each individual assignment and the take-home exams are intended to be solely the work of a single student whose name appears on it.

6.5 Attendance

Although no explicit marks are given for class participation, attendance is crucial for successful completion of this course.

---

7 College of Biological Science Statements

7.1 Wellness

If you are struggling with personal or health issues:

- Counselling Services offers individualized appointments to help students work through personal struggles that may be impacting their academic performance.
- Student Health Services is located on campus and is available to provide medical attention.
- For support related to stress and anxiety, besides Health Services and Counselling Services, Kathy Somers runs training workshops and one-on-one sessions related to stress management and high performance situations.
  http://www.selfregulationskills.ca/

8 University Statements

8.1 Email Communication

As per university regulations, all students are required to check their e-mail account regularly: e-mail is the official route of communication between the University and its students.

8.2 When You Cannot Meet a Course Requirement

When you find yourself unable to meet an in-course requirement because of illness or compassionate reasons please advise the course instructor (or designated person, such as a teaching assistant) in writing, with your name, id#, and e-mail contact. The grounds for Academic Consideration are detailed in the Undergraduate and Graduate Calendars.

Undergraduate Calendar - Academic Consideration and Appeals
https://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-ac.shtml

Graduate Calendar - Grounds for Academic Consideration
https://www.uoguelph.ca/registrar/calendars/graduate/current/genreg/index.shtml

Associate Diploma Calendar - Academic Consideration, Appeals and Petitions
https://www.uoguelph.ca/registrar/calendars/diploma/current/index.shtml

8.3 Drop Date

Students will have until the last day of classes to drop courses without academic penalty. The deadline to drop two-semester courses will be the last day of classes in the second semester. This applies to all students (undergraduate, graduate and diploma) except for Doctor of Veterinary Medicine and Associate Diploma in Veterinary Technology (conventional and alternative delivery) students. The regulations and procedures for course registration are

available in their respective Academic Calendars.

Undergraduate Calendar - Dropping Courses
https://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-drop.shtml

Graduate Calendar - Registration Changes
https://www.uoguelph.ca/registrar/calendars/graduate/current/genreg/genreg-reg-regchg.shtml

Associate Diploma Calendar - Dropping Courses
https://www.uoguelph.ca/registrar/calendars/diploma/current/c08/c08-drop.shtml

8.4 Copies of Out-of-class Assignments

Keep paper and/or other reliable back-up copies of all out-of-class assignments: you may be asked to resubmit work at any time.

8.5 Accessibility

The University promotes the full participation of students who experience disabilities in their academic programs. To that end, the provision of academic accommodation is a shared responsibility between the University and the student.

When accommodations are needed, the student is required to first register with Student Accessibility Services (SAS). Documentation to substantiate the existence of a disability is required; however, interim accommodations may be possible while that process is underway.

Accommodations are available for both permanent and temporary disabilities. It should be noted that common illnesses such as a cold or the flu do not constitute a disability.

Use of the SAS Exam Centre requires students to book their exams at least 7 days in advance and not later than the 40th Class Day.

For Guelph students, information can be found on the SAS website
https://www.uoguelph.ca/sas

For Ridgetown students, information can be found on the Ridgetown SAS website
https://www.ridgetownc.com/services/accessibilityservices.cfm

8.6 Academic Integrity

The University of Guelph is committed to upholding the highest standards of academic integrity, and it is the responsibility of all members of the University community-faculty, staff, and students-to be aware of what constitutes academic misconduct and to do as much as possible to prevent academic offences from occurring. University of Guelph students have the responsibility of abiding by the University's policy on academic misconduct regardless of their location of study; faculty, staff, and students have the responsibility of supporting an environment that encourages academic integrity. Students need to remain aware that instructors have access to and the right to use electronic and other means of detection.

Please note: Whether or not a student intended to commit academic misconduct is not relevant for a finding of guilt. Hurried or careless submission of assignments does not excuse students from responsibility for verifying the academic integrity of their work before submitting it. Students who are in any doubt as to whether an action on their part could be construed as an academic offence should consult with a faculty member or faculty advisor.

Undergraduate Calendar - Academic Misconduct
https://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-amisconduct.shtml

Graduate Calendar - Academic Misconduct
https://www.uoguelph.ca/registrar/calendars/graduate/current/genreg/index.shtml

8.7 Recording of Materials

Presentations that are made in relation to course work - including lectures - cannot be recorded or copied without the permission of the presenter, whether the instructor, a student, or guest lecturer. Material recorded with permission is restricted to use for that course unless further permission is granted.

8.8 Resources

The Academic Calendars are the source of information about the University of Guelph's procedures, policies, and regulations that apply to undergraduate, graduate, and diploma programs.

Academic Calendars
https://www.uoguelph.ca/academics/calendars
