BINF 6970: Statistical Bioinformatics

Winter 2019 Course Outline

Instructor: **Khurram Nadeem**, Assistant Professor
Department of Mathematics & Statistics
Email: nadeemk@uoguelph.ca
Office: MacNaughton 517
Times & Venue: Tuesday and Thursday 1:00-2:20 pm in SSC 1306
Office Hour: Monday 1:00-2:00 pm and by appointment

Guest Lecturer: **Zeny Feng**, Professor
Department of Mathematics and Statistics
Email: zfeng@uoguelph.ca
Office: MacNaughton 540
Times: March 14, 19, 21, 26, and 28 (Tuesday and Thursday), 1:00-2:20 pm, SSC 1306
Office Hour: Monday and Wednesday, 9:30-10:30 am, Tuesday 10-11 am
(the weeks of March 18-22, 25-29).

Teaching Assistant: TBA

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 modeling of complex, often high-dimensional, data structures.

Prerequisites: Introductory courses in statistics, mathematics and programming.

Restrictions: Membership in the bioinformatics program or instructor consent.

Course Objectives

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 R language and packages for demonstration and analysis purposes.

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/).

BINF 6970: Statistical Bioinformatics

Winter 2019 Course Outline

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/).

Assignments and Important Dates

Homework 40% (Four assignments, equally weighted)

Homework Due Dates: Assig. 1 (**Tu, Jan 29**); Assig. 2 (**Th, Feb 14**);

Assig. 3 (**Tu, Mar 19**); Assig. 4 (**Th, Apr 4**)

Midterm Exam 20% (Thursday **Feb 28**; In-class)

Final Exam 40% (Monday **Apr 8**, 8:30 am-10:30 am; location TBA)

Note: Final exam will consist of two components: i) An in-class written exam on the exam day (08/04/2019), and ii) a take-home exam involving analysis of a given dataset using statistical data mining methods covered in class.

Drop Date: 40th class day (Friday, March 8) is the last day to drop one semester courses.

Course 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. Familywise 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. Topics in statistical genetics for this course might include: basic concepts and terminology in genetics, population structure, Hardy-Weinberg Equilibrium,

linkage equilibrium (LE) or linkage disequilibrium (LD), genetic association test, genome-wide genetics association studies, haplotype inference, haplotype association analysis, and imputation methods. (Note: The final coverage may not include all these topics depending on time and other factors.)

Comments on the Course Work

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

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.

Exams: Closed book, closed notes (except for the take-home exam – 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?

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 final take-home exam is intended to be solely the work of a single student whose name appears on it.

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.

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.

Handing in the assignments: In addition to handing in the hard copies on the due dates, students are also required to submit electronic copies of their homework assignments to CourseLink Dropbox.

Note: We mostly put stuff 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.

University Statements

Rights and Responsibilities: A complete listing of rights and responsibilities appears in the graduate calendar:

https://www.uoguelph.ca/registrar/calendars/graduate/2017-2018/genreg/index.shtml

Keep Copies of Everything: Sometimes homework's get lost and quiz grades are not recorded correctly. Please keep copies of any assignments you hand in and keep a folder with all your work in case there is a problem.

Attendance: Illness, etc.: Attendance is, of course, very important. If you miss class because of illness or for compassionate reasons, please see the instructor for possible academic consideration.

Course Feedback: The sponsoring department require student assessments of all courses taught by the departments. These assessments provide essential feedback to faculty on their teaching by identifying both strengths and possible areas of improvement. In addition, annual student assessment of teaching provides part of the information used by the department's Tenure and Promotion Committee in evaluating the faculty member's contribution in the area of teaching. The department's teaching evaluation questionnaire invites student response both through numerically quantifiable data, and written student comments. In conformity with University of Guelph Faculty Policy, the department's Tenure and Promotions Committee **only considers comments signed by students (choosing "I agree" in question 14)**. Your instructor will see all signed and unsigned comments after final grades are submitted. Written student comments may also be used in support of a nomination for internal and external teaching awards. NOTE: No information will be passed on to the instructor until after the final grades have been submitted.

Electronic Recording of Classes: The electronic recording of classes is expressly forbidden without the prior consent of the instructor. This prohibition extends to all components of the course, including, but not limited to, lectures, tutorials, and lab instruction, whether conducted by the instructor or teaching assistant, or other designated person. When recordings are permitted they are solely for the use of the

authorized student and may not be reproduced, or transmitted to others, without the express written consent of the instructor.

Academic Misconduct: 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 offenses 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 discourages misconduct. 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 offense should consult with a faculty member or faculty advisor. The **Academic Misconduct Policy** is detailed in the Undergraduate Calendar:

https://www.uoguelph.ca/registrar/calendars/graduate/2017-2018/genreg/sec d0e3039.shtml

Accessibility: The University of Guelph is committed to creating a barrier-free environment. Providing services for students is a shared responsibility among students, faculty and administrators. This relationship is based on respect of individual rights, the dignity of the individual and the University community's shared commitment to an open and supportive learning environment. Students requiring service or accommodation, whether due to an identified, ongoing disability or a short-term disability should contact Student Accessibility Services as soon as possible. For more information, contact 519-824-4120 ext. 56208 or email sas@uoguelph.ca or see the website:

https://wellness.uoguelph.ca/accessibility/