Methods for Statistical Learning for Big Data

Canvas title: "Algorithms in Biomedical Data Science"

QBS 177
Winer Quarter
Instructors: Jiang Gui and Nicholas Jacobson
Class Location: WTRB 373 - except 2/5 3/4 in 371
Class Meetings: M W 12:30 to 2:00 pm.

Course Description

This course provides an introduction to algorithms used in data science with applications to biomedical and health data science. The goal of this course is to present an overview of many of the approaches used for big data focusing on analytical methods and algorithms. The course assumes that students have some knowledge of R. Students will be provided with 2 large data sets. Lectures on data reduction, classification, and optimization will request students complete homework for these datasets. Special attention will be given to students' active learning by programming in a statistical software package R.

Course Learning Outcomes

In this course, students will:

- Understand and apply penalized regression for variable selection and prediction
- Understand and apply classification, discriminant analysis and clustering methods
- Understand neural network and introduction to parallel computing.

Teaching Methods & Philosophy

Students are encouraged to read materials outside of class and ask questions in class.

Class Climate & Inclusivity

A teacher always learns from the students. All questions are a chance for both the questioner, other students and the teacher to increase their understanding.

Suggested Reading

- *Elements of Statistical leaning, Trevor Hastie Robert Tibshirani , Jerome Friedman. E book available.*

- *Machine Learning: A Probabilistic Perspective 3rd Edition, Kevin Murphy https://mitpress.mit.edu/books/machine-learning-1*

Assignments and Labs:

There will be weekly problem sets and lab exercises for each session. Problems will include hands on examples of real data. Students are expected to submit the assignments using an R script.

Grades breakdown

Exam (midterm, take home exam) - 35%

Week homework - 30% (assigned on Friday via canvas, due noon next Wednesday, submitted electronically through canvas).

Team Project (presented at the last class meeting) - 35%

Projects: There will be one project due at the end of term. This will involve a description, interpretation and application of a regression method to a dataset chosen by the student and approved by the instructor before the 5th week of class. The student will submit a written report and make a 10-15 minute presentation in class.

Course Schedule and Topics

Week	Date & Location	Topics	Readings
1	WTRB 373	**Introduction to Next Generation Sequencing data**	Shendure, J., Ji, H. Next-generation DNA sequencing. Nat Biotechnol 26, 1135–45 (2008) doi:10.1038/nbt1486
2	WTRB 373	**Linear and logistics regression.**	Machine Learning chapter 7
3	WTRB 373	**PCA for dimension reduction and visualization**	Elements of Statistical Learning, Chapter 14.5
4	WTRB 373	**Introduction to parallel computing**	NA

5	WTRB 373 & WTRB 371	**Penalized regressions.**	Elements of Statistical Learning, Chapter 3
6	WTRB 373	**Cluster analysis.**	Elements of Statistical Learning, Chapter 13, 14.3
7	WTRB 373	**Linear and quadratic discriminant analysis.**	Elements of Statistical Learning, Chapter 4.
8	WTRB 373	**Gaussian Graphic model and Ising model**	Machine Learning, chapter 19.
9	WTRB 373 & WTRB 371	**Introduction to Neural networks.**	Elements of Statistical Learning, Chapter 11.
10	WTRB 373	**Presentations of student projects.**	

Where class fits in terms of Data Science, Type and Applications?

Data Science		
Analytics	Algorithm	Statistical Inference
30	10	60
Theory vs Application		
Theory	Application	
25	75	