Bioinformatics Online Course: BINF620 - Big Data Analytics in Healthcare

Instructor: Zugui Zhang, PhD, FAHA

Prerequisites: Pre-requirements: prerequisites Advanced statistics (calculus and probability-based), linear algebra, epidemiology and data analysis, basic computational skills in R/RStudio. Credits: 3.0 hours.

Course Description: Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. This course will introduce students to detect risk factors, find patterns and reason about data, make causal inference and decision about health care and precision medicine.

Course Goals: The main objective of the course is to instruct students how to conduct big data analysis in health care via machine learning methodologies. Machine learning in healthcare is becoming more acceptable in health care worldwide. Today, machine learning is helping with administrative processes in hospitals, identification and treatment of diseases and personalized medical treatments, among others. This course will introduce students to design studies in health care, obtain health care big data, pre-processing big-data set, detect rick factors, find patterns and reason about data, make causal inference and decision about health care and precision medicine.

Computational tools
R/RStudio packages for statistical analysis and visualization; Python is optioal.
Microsoft Excel for manipulation of data files

Course format: all course materials, including presentation slides, video presentations, assignments and projects, are on line.

If you register for **BINF620-010,** which is in hybrid format, you will study on-line and we will meet in Person each week **Friday: 1:25PM-2:15PM, Robinson Hall Room 203**, reviewing course materials and answering questions.

If you can't meet in person on Fridays, you should register for course **BINF620-194, and you study on line.**

Office hours and location:
Friday 1:25PM-2:15PM, Robinson Hall Room 203

Or via Zoom meeting or Appointment

meeting ID: 470 914 5196

password: 200610

Required Textbook

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani: ***An Introduction to Statistical Learning--with Applications in R.*** **Springer (2015).** A copy of PDF will be provided (Purchase not necessary).

Course Requirements

1. **Participating.** Students are required to take all course activities. Be sure to finish reading assignments and watch lecture presentations, and complete the assigned tasks by due dates.
2. **Submitting Homework.** Students will be required to complete 7 - 8 homework and submit each homework on time.
3. **Data processing work (short project).** Each Student will submit a detail preparation work using methods of pre-processing big data analysis.
4. **Applied Project.** Each of student will required to submit a final research report. Each student will choose individual data set from public accessible database, design his/her own study, analyzing the big data via machine learning, and prepare your report.

Applied project should be submitted in the format of academic article, consisting of the following parts:

- Title
- Abstract
- Background
- Study Design: Aims, Population selection; Data and materials
- Statistical analysis: Methods; Response Variable, Variable Selection/Model Selection.
- Results: Model used; model evaluation metrics; major findings. Tables and figures/plots.
- Discussion: summary of methods and results, strength of your study; limitation of your study.
- Conclusions
- References
- A supplementary document: R file with R codes for your study

Evaluation: Homework assignments (up to 8), one data processing work, one final applied project.

1.	Participation	8%
2.	Homework:	32%
2.	Data processing work:	20%
3.	Final Research Proposal:	40%

Grade Cutoffs

>90%:	A
80~89%:	B
70~80%:	C
<70%:	D

Statistical and Machine Learning Topics

1. Basic modelling methods.
2. Linear regression and generalized linear regression.
3. Classification.
4. Resampling methods.
5. Model and feature selection.
6. Tree-based methods
7. Support vector machines
8. Unsupervised machine learning
9. Machine Learning for Survival Analysis*

*if time permts

Healthcare Topics

1. Risk factor identification and model section. Using correlation matrix and dimension reduction methods to effectively select variable and parsimony model, to improve predictive capability. Bootstrap technique, cross-validation, k-fold cross-validation , and nested cross-validation for model selection and model evaluation.
2. Population-Level Risk Stratification. Based on administrative, utilization, and clinical data, applying machine learning to find surrogates for risk factors that would otherwise be missing and perform risk stratification at the large patients population.
3. Reveal disease progression model and burden of disease. Using clustering (k-means algorithm to discover disease subtypes and stages. Explore disease progression models from electronic health record and the burden of chronic disease, to derive a meaningful characterization of disease progression and stages, to identify the progression trajectory of individual patients, to provide decision support for early intervention, and develop data-driven guidelines for care plan management.
4. Assessing the treatment effects from medical observational studies. Observational data has become an important complementary of clinical trial. Average Treatment Effect (ATE) will be assessed via propensity score approach. Propensity scores for the analysis of observational data are typically estimated using logistic regression. Neural networks, support vector machines, decision trees (CART), and meta-classifiers with boost can generate propensity score with fewer assumptions or greater accuracy.
5. Causal inference in health-care data. A major challenge in causal inference from observational studies is how to control or adjust for the confounding factors. Random forests, Bayesian trees, Gaussian processes, Causal Lasso are the useful techniques to make causal inference in the observational health care data.
6. Predicting disease onsets from longitudinal data. Convolutional Neural Networks (CNN) for disease identification, detection, and prediction.
7. Handling missing data and address selection bias in medical studies. Stochastic Regression Imputation, random forests, and other machine learning algorithms are the methods to deal with missing data, in medical records.
8. Predicting and preventing medical cost. Due to the non-linear structure of health care data, nonlinear models including support vector regression and neural networks are better at predicting hospital prices than linear regression and regression trees.