PQHS 471: Machine Learning /Data Mining (Spring 2018) (DRAFT)

Instructor: Chun Li

Time: Tuesday/Thursday 2:30 – 3:45 pm
Locatio: Wood Building WG-86
Office hour: Available through contact; Wolstein Research Building 2528 <cxl791@case.edu>

General description: This course aims to introduce concepts and major methods in statistical learning, machine learning, and data mining, emphasizing on the statistical aspects of various approaches and on biomedical applications. Specifically, we will cover prediction model building, model regularization (shrinkage, lasso), classification (logistic regression, discriminant analysis, k-nearest neighbors), trees; ensemble methods (random forests, boosting), support vector machines, artificial neural networks (backpropagation, deep learning, CNN, RNN); association rules, k-means and hierarchical clustering, GANs. Basic techniques that are applicable to many of the areas, such as cross-validation, the bootstrap, dimensionality reduction, and splines, will be explained and used repeatedly. Minimum requirements are calculus, linear algebra, and some exposure to statistics (EPBI 431).

Textbooks:
ISLR: James et al. (2013) *An Introduction to Statistical Learning, with Applications in R*. (8th printing) (http://www-bcf.usc.edu/~gareth/ISL/) (available at https://link.springer.com/ from any Case IP)
NNDL: Nielsen (2015) *Neural Networks and Deep Learning* (http://neuralnetworksanddeeplearning.com/)

Other books:
ESL: Hastie et al. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (12th printing) Springer (http://www.stanford.edu/~hastie/ElemStatLearn/)
CASI: Efron and Hastie (2016) *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press (https://web.stanford.edu/~hastie/CASI/index.html)
DL: Goodfellow et al. (2016) *Deep Learning*, (http://www.deeplearningbook.org/)
MMDS: Leskovec et al. (2014) *Mining of Massive Datasets*, 2nd ed. (http://www.mmds.org/)
HOML: Géron (2017) *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly (http://proquest.safaribooksonline.com/9781491962282?uicode=ohlink)

Course style: Lecture + Discussion

1. Students should read the material to be covered before each lecture. I will randomly call on students to briefly (<1 minute) summarize the material: what is this section about (big picture, methods in general). It is okay if you do not understand some technical details.
2. Students are strongly encouraged to raise questions and participate in discussions.

Course grade: 25% each for
1) homework,
2) midterm on **March 8**,
3) final exam in the week of **April 30**, and
4) participation (summarize materials and participate in discussions)

PQHS 471 tentative schedule (Spring 2018):

Week	Date	HW/exam	Topic
1	1/16		Introduction; challenges from AI, big data; data science; R/Python; statistical learning in general (ISLR 1-2)
2	1/22		linear regression; curse of dimensionality (ISLR 3)
3	1/29		classification, LDA/QDA, ROC, etc. (ISLR 4)
4	2/5	HW1 due	cross-validation, bootstrap, subset selection (ISLR 5-6)
5	2/12		ridge, lasso, splines (ISLR 6-7)
6	2/19		local regression, GAMs, trees (ISLR 7-8)
7	2/26	HW2 due	random forests, boosting (ISLR 8)
8	3/5	*Midterm*	support vector machines (ISLR 9)
	3/12	Spring break	
9	3/19		neural networks, backpropagation, model tuning (NNDL 1-3)
10	3/26		deep learning, CNN, RNN (NNDL 6)
11	4/2		unsupervised learning, PCA, clustering (ISLR 10)
12	4/9	HW3 due	association rules, SOM, MDS (ESL 14.2, 14.4, 14.8-9)
13	4/16		GANs, additional topics
14	4/23		Examples and review
15	4/30	*Final exam*	