PQHS 471: Machine Learning /Data Mining (Spring 2019)

Instructor: Chun Li

Time	Tuesday/Thursday 2:30 – 3:55 pm
Location	Wood Building WG-73
Office hour	Available through contact; Wolstein Research Building 2528; cxl791@case.edu
TA	Youjun Li; yxl1469@case.edu
Course site	https://github.com/cxl791/PQHS471

General description: We introduce concepts and major methods in machine learning and data mining. The goals are to understand the models, intuition, statistical underpinnings, strengths and weaknesses, assumptions and trade-offs of various approaches. Technical details such as optimization algorithms and theoretical properties are not of primary interest. 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. R and Python will be used. Minimum prerequisites are calculus, linear algebra, and some exposure to statistics (PQHS 431).

Books	Book title and webpage
ISLR	James et al. (2013) *An Introduction to Statistical Learning, with Applications in R*. Springer. (8th printing) Book website: http://www-bcf.usc.edu/~gareth/ISL/
DLwR	Chollet and Allaire (2018) *Deep Learning with R.* Manning.
HOML	Géron (2017) *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* O'Reilly.
NNDL	Nielsen (2015) *Neural Networks and Deep Learning.*
ESL	Hastie et al. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer. (12th printing)
CASI	Efron and Hastie (2016) *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press.
DL	Goodfellow et al. (2016) *Deep Learning*. MIT Press.
MMDS	Leskovec et al. (2014) *Mining of Massive Datasets*, 2nd ed. Cambridge University Press.
R4DS	Grolemund and Wickham (2017) *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly.

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 (due by the end of Friday, March 8),
3) final exam (in the week of May 2), and
4) participation (summarize materials and participate in discussions)

For HW/exams, turn in: (1) A PDF/Word file for answers and figures, and a text file (.r, .py) for code with comments, or (2) Everything in a notebook format (.Rmd, .ipynb). Via github (preferred) or email.

PQHS 471 *tentative* schedule (Spring 2019):

Week	Date	HW/exam due	Chapters	Topic
1	1/15		ISLR 1	introduction; data science; R/RStudio; Git; Python/Anaconda
1	1/17		ISLR 2, HOML 1	statistical learning and machine learning in general
2	1/22		ISLR 3.1–3.3	linear regression; demonstration of a model building process
2	1/24		ISLR 3.3–3.5, 4.1–4.3	linear regression; curse of dimensionality; logistic regression
3	1/29		ISLR 4.3–4.4	interaction; Simpson's paradox; LDA/QDA
3	1/31		ISLR 4.5, 5.1	confusion matrix; ROC; cross-validation
4	2/5	HW1	ISLR 5.2, 6.1	bootstrap; subset selection; C_p, AIC, BIC
4	2/7		ISLR 6.2	ridge regression, lasso
5	2/12		ISLR 6.3–6.4, 7.1–7.4	PCR, PLS; splines
5	2/14		ISLR 7.5–7.7	smoothing splines; local regression, GAMs
6	2/19		ISLR 8	trees
6	2/21		ISLR 8	random forests; boosting
7	2/26	HW2		boosting
7	2/28		ISLR 9	support vector machines (SVMs)
8	3/5		ISLR 9	support vector machines; "doughnut" data demonstration
8	3/7		– ?	– ?
-	3/8	Midterm		
-	-		– Spring break –	
9	3/19		NNDL 1, HOML 10	artificial neural networks (ANNs)
9	3/21		–	–
10	3/26		–	–
10	3/28		–	–
11	4/2			deep neural networks; TensorFlow?
11	4/4			various issues when fitting neural network models
12	4/9	HW3		convolutional neural networks (CNNs)
12	4/11			recurrent neural networks (RNNs)
13	4/16		ISLR 10.1–10.2, HOML 8	unsupervised learning, PCA; generative adversarial networks (GANs)
13	4/18		ISLR 10.3	hierarchical clustering
14	4/23		ESL 14.2	market basket analysis (MBA), association rules
14	4/25		ESL 14.8, 14.4	multidimensional scaling (MDS); self-organizing maps (SOM)
-	-	Final		