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

Instructor: Chun Li

Time	Tuesday/Thursday 2:30 – 3:45 pm
Location	Wood Building WG-73
Office hour	Available through contact; Wolstein Research Building 2528; cxl791@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 at https://link.springer.com/) http://www-bcf.usc.edu/~gareth/ISL/
HOML	Géron (2017) *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* O'Reilly. http://proquest.safaribooksonline.com/9781491962282?uicode=ohlink
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. Springer. (12th printing) 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/
DL	Goodfellow et al. (2016) *Deep Learning.* MIT Press. http://www.deeplearningbook.org/
MMDS	Leskovec et al. (2014) *Mining of Massive Datasets*, 2nd ed. Cambridge University Press. http://www.mmds.org/
R4DS	Grolemund and Wickham (2017) *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly. http://r4ds.had.co.nz

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 on March 8),
3) final exam (in the week of April 30), 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 2018):

Week	Date	HW due	exam	Chapters	Topic
1	1/16			ISLR 1	introduction; data science (AI, big data); R/git/Python
1	1/18			ISLR 2, HOML 1	statistical learning and machine learning in general
2	1/23			ISLR 3.1–3.2	linear regression; demonstration of a model building process
2	1/25			ISLR 3.3–3.5, HOML 2	linear regression; curse of dimensionality
3	1/30			ISLR 4.1–4.3, HOML 3	classification, logistic regression
3	2/1			ISLR 4.3–4.4, HOML 3	LDA/QDA, ROC, etc.
4	2/6	HW1		ISLR 4.5, 5	cross-validation, bootstrap
4	2/8			ISLR 6	subset selection; C_p, AIC, BIC; ridge regression
5	2/13			ISLR 6.2–6.3, HOML 4	ridge, lasso, PCR, PLS
5	2/15			ISLR 6.4, 7.1–7.4, HOML 4	splines
6	2/20			ISLR 7.5–7.7	smoothing spline; local regression, GAMs
6	2/22			ISLR 8, HOML 6	trees
7	2/27	HW2		ISLR 8, HOML 7	random forests, boosting
7	3/1			ISLR 9	support vector machines
8	3/6			ISLR 9, HOML 5	support vector machines
8	3/8		Midterm	ISLR 9	"doughnut" data demonstration
-	3/12			– Spring break –	
9	3/20			NNDL 1, HOML 10	artificial neural networks, backpropagation
9	3/22			NNDL 2-3, HOML 10	ANN model tuning
10	3/27			NNDL 6, HOML 13	deep learning, CNN
10	3/29			NNDL 6, HOML 14	RNN
11	4/3			ISLR 10.1–10.2, HOML 8	unsupervised learning, PCA
11	4/5			ISLR 10.3	hierarchical clustering
12	4/10	HW3		ESL 14.2	MBA, association rules
12	4/12			ESL 14.4	SOM
13	4/17			ESL 14.8–14.9	multidimensional scaling
13	4/19				additional topics if time permits
14	4/24				additional topics if time permits
14	4/26				additional topics if time permits
15	4/30		Final		

HOML 9 (TensorFlow) ?