COMS W4721 Machine Learning for Data Science

Columbia University, Spring 2017

Instructor: John Paisley
Location: 501 Schermerhorn Hall
Time: T/Th 7:40pm - 8:55pm
Office hours: Monday 11am-12pm @ 422 Mudd Building

TA's:	Ghazal Fazelnia	gf2293@columbia.edu	CVN office hours via email (no fixed time)
	Tianhao Lu	tl2710@columbia.edu	Tue/Thu 1pm - 2pm @ CS TA room, Mudd 122A (1st floor)
	Dheeraj Kalmekolan	drk2143@columbia.edu	Wed 4:30pm - 6:30pm @ CS TA room, Mudd 122A (1st floor)
	Ashutosh Nanda	an2655@columbia.edu	Fri 1:30pm - 3:30pm @ CS TA room, Mudd 122A (1st floor)
	Avinash Bukkittu	ab4377@columbia.edu	Tues 10am - 12pm @ CS TA room, Mudd 122A (1st floor)
	Yuhao Zhang	yz3044@columbia.edu	Tues 9pm - 11pm @ CS TA room, Mudd 122A (1st floor)
	Jiefu Ying	jy2799@columbia.edu	Fri 4pm - 6pm @ CS TA room, Mudd 122A (1st floor)
	George Yu	gy2206@columbia.edu	Tues 4pm - 5pm & Wed 12pm-1pm @ CS TA room, Mudd 122A (1st floor)
	Peng Wu	pw2393@columbia.edu	Mon 6:30pm - 8:30pm @ CS TA room, Mudd 122A (1st floor)

Synopsis: This course provides an introduction to supervised and unsupervised techniques for machine learning. We will cover both probabilistic and non-probabilistic approaches to machine learning. Focus will be on classification and regression models, clustering methods, matrix factorization and sequential models. Methods covered in class include linear and logistic regression, support vector machines, boosting, K-means clustering, mixture models, expectation-maximization algorithm, hidden Markov models, among others. We will cover algorithmic techniques for optimization, such as gradient and coordinate descent methods, as the need arises.

Prerequisites: Basic linear algebra and calculus, introductory-level courses in probability and statistics. Comfort with a programming language (e.g., Matlab) will be essential for completing the homework assignments. Not open to students who have taken COMS 4771, STATS 4400 or IEOR 4525.

Text: There is no required text for the course. Suggested readings for each class will be given from the textbooks below. These readings are meant to be general pointers and may contain more material than we cover in class.

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning, Second Edition,* Springer. [link]
C. Bishop, *Pattern Recognition and Machine Learning*, Springer. [link]
H. Daume, *A Course in Machine Learning,* Draft. [link]

Grading: 5 homework assignments (50%), midterm exam (25%), final in-class exam (25%). Each homework assignment will have a programming component that will count significantly toward the final homework grade. The final in-class exam will focus on material from the second half of the course.

	Date	Slides	Topics covered	Suggested readings	Additional Information
Week 1	1/17/2017	[PDF]	Introduction, maximum likelihood estimation	ESL Ch. 1-2; PRML Ch. 2.1-2.3	
	1/19/2017	[PDF]	linear regression, least squares, geometric view	ESL Ch. 3.1-3.2; PRML Ch. 1.1, 3.1	
Week 2	1/24/2017	[PDF]	ridge regression, probabilistic views of linear regression	ESL Ch. 3.3-3.4; PRML Ch. 3.1-3.2	Homework 1 out (see Courseworks)
	1/26/2017	[PDF]	bias-variance, Bayes rule, maximum a posteriori	ESL Ch. 7.1-7.3, 7.10; PRML Ch 2.3	Due February 5 by 11:59pm
Week 3	1/31/2017	[PDF]	Bayesian linear regression	PRML 3.3-3.5	
	2/2/2017	[PDF]	sparsity, subset selection for linear regression	ESL Ch. 3.3-3.8	
Week 4	2/7/2017	[PDF]	nearest neighbor classification, Bayes classifiers	ESL Ch. 13.3-13.5; CML Ch. 2, 7	
	2/9/2017		No class (University-wide)		
Week 5	2/14/2017	[PDF]	linear classifiers, perceptron	ESL Ch. 4.5; CML 3	Homework 2 out (see Courseworks)
	2/16/2017	[PDF]	logistic regression, Laplace approximation	ESL Ch. 4.4; PRML Ch. 4.3-4.5	Due February 26 by 11:59pm
Week 6	2/21/2017	[PDF]	kernel methods, Gaussian processes	ESL Ch. 6; PRML Ch. 6; CML Ch. 9	
	2/23/2017	[PDF]	maximum margin, support vector machines	ESL Ch. 12.1-12.3; PRML Ch. 7.1	
Week 7	2/28/2017	[PDF]	trees, random forests	ESL Ch. 9.2, 15; CML Ch. 1	
	3/2/2017	[PDF]	boosting	ESL Ch. 10; CML Ch. 11	
Week 8	3/7/2017		Midterm exam (covers material through Week 7)		
	3/9/2017		no class		
Week 9	3/14/2017		no class (Spring break)		Homework 3 out (see Courseworks)
	3/16/2017		no class (Spring break)		Due March 26 by 11:59pm
Week 10	3/21/2017	[PDF]	clustering, k-means	ESL Ch. 14.3; PRML Ch. 9.1; CML Ch. 13	
	3/23/2017	[PDF]	EM algorithm, missing data	ESL Ch. 8.5; PRML Ch. 9.3-9.4	
Week 11	3/28/2017	[PDF]	mixtures of Gaussians	PRML Ch. 9.2; CML Ch. 14	Homework 4 out (see Courseworks)
	3/30/2017	[PDF]	matrix factorization	Review article	Due April 12 by 11:59pm
Week 12	4/4/2017	[PDF]	non-negative matrix factorization	ESL Ch. 14.6; Review article	
	4/6/2017	[PDF]	latent factor models, PCA and variations	ESL Ch. 14.5; PRML Ch. 12.1-12.3	
Week 13	4/11/2017	[PDF]	Markov models	PRML Ch. 13.1	Homework 5 out (see Courseworks)
	4/13/2017	[PDF]	hidden Markov models	PRML Ch. 13.2	Due April 23 by 11:59pm
Week 14	4/18/2017	[PDF]	continuous state-space models	PRML Ch. 13.3	
	4/20/2017	[PDF]	association analysis	ESL Ch. 14.2; Book chapter	
Week 15	4/25/2017	[PDF]	model selection	SL Ch. 7.5-7.7; PRML Ch. 4.4	
	4/27/2017		Final in-class exam (covers material starting from Week 10)		