COMP 540 : STATISTICAL MACHINE LEARNING
Spring 2016
4 credits
Time: MWF 10:00 to 10:50
Venue: HRZ 212

COMP 540 is about learning models from data. The course is designed to give you a foundational understanding of modern algorithms in learning and data mining, as well as hands-on experience with its applications in science and engineering. The course assumes prior background in discrete mathematics, algorithms, linear algebra, convex optimization, probability and statistics as well a facility with programming. COMP 540 can be taken as grounding for future research in machine learning or to gain familiarity with machine learning methods for application in other fields.

Instructor

Devika Subramanian
Office: 3094 Duncan Hall; Hours: after class TTh and by appt.
Email: devika@rice.edu Telephone: 713-348-5661

Course Policies

Grades will be based on the term project (30%), six homework assignments (20%), two exams (20% each), and class participation (10%).

Course workload and attendance

You must attend the MWF lectures at 10:00 am. You should expect to work about 8-10 hours per week for this class, including lecture time.

Term Project

The term project gives you an opportunity to work on a real-world prediction problem. You will compete in a Kaggle-in-class

project on an image recognition task. It is an Honor code violation to make the dataset available to anyone else. You will implement the algorithms and analysis techniques taught in class in the context of this prediction problem. Term projects will be done in groups of two. You will present your final project as a poster in the final week of class and submit a final project report on the last day of class. More details about the project are here.

Homework assignments

Homework assignments will be posted on Owlspace. Please watch Owlspace for these postings. The mathematical component of the assignments reinforces lecture material and will typically involve extending algorithms covered in class. Please typeset (LaTeX) this portion of your homework to make it easier for us to grade your work. The purpose of the programming component of the assignments (in Matlab and Python) is to build an operational level of understanding of the algorithms covered in class. Please turn assignments in before their due dates on Owlspace. Only one submission needs to be made per group. My TAs and I will grade them and return feedback on Owlspace within a week of submission. No submissions will be accepted a week after the due date.

Examinations

There will be two in-class examinations. The first examination will be held on February 25, 2016 from 7 pm to 10 pm. If you have a conflict with that time or day, you can take the exam earlier on February 24 during a three-hour period between 9 am and 4 pm. The second exam will be held at a time and place scheduled by the registrar during the finals period. It will be a three-hour examination.

Late policies

Because each of you will probably come upon some time during the semester where so much work piles up that you need a little

extra time, every student begins the semester with two free late days. After your two late days are exhausted, assignments that come in late (up to a maximum of three days) will be assessed a late penalty of 10% of your score per late day. You should think of these free late days as extensions you have been granted ahead of time, and use them when you might have otherwise tried to ask for an extension. As a result, **getting an extension beyond the two free late days will generally not be granted**. In very special circumstances for which you can provide official documentation (primarily extended medical problems or other emergencies), extensions may be granted beyond the late days. All extension requests must be directed to me (devika@rice.edu), no later than 24 hours **before** the assignment is due.

Grading and Re-grading
Your performance statistics will be posted on Owlspace. If you believe we have made an error in grading your assignments or exams, please bring the matter to our attention within one week of when we return your work. No makeup exams will be given.

Piazza
All course related discussions and questions should be posted on Piazza. Please do not send personal email to me or to the TA — Piazza is the fastest way to get a response from us and from the class community. Piazza allows you to ask questions in private as well as anonymously. We request you not to post code or answers to assignments on Piazza. If you are having difficulty with an assignment and need to show the teaching assistants your code, please make a private Piazza post.

Academic Integrity
The work you submit for this class is expected to be the result of your own work and that of your teammate. You are free to discuss course material and approaches with your other classmates, the TA and me, but you should never misrepresent

someone else's work as your own. It is also your responsibility to protect your work from unauthorized access. I expect you to follow the Honor Code in this course.

Accommodation for Disability
If you have a documented disability that will impact your work in this class, please contact me (devika@rice.edu) to discuss your needs. Additionally, you will need to register with the Disability Support Services Office in the Ley Student Center.

Textbook
The textbook I will follow is Machine Learning: a probabilistic perspective by Kevin Murphy, published by the MIT Press, 2014. I like this book for two reasons: it is a well-written textbook that covers both theoretical (algorithmic and statistical) and practical aspects of modern machine learning methods. The probabilistic perspective provides a common mathematical framework for understanding supervised and unsupervised learning algorithms.

Other textbooks
Other resources for the class, covering background reading, as well as alternative viewpoints on the core material, include

- Pattern recognition and machine learning by Christopher M. Bishop published by Springer, 2006.
- Convex optimization by Boyd and Vandenberghe by Cambridge University Press.
- An introduction to statistical learning by James, Witten, Hastie and Tibshirani, Springer, 2014.

Readings for the class
Here is the sequence of topics we will cover during the 14 week term (all chapter references are from Murphy's book). You will get the most out of the lectures, if you can read the material for each week ahead of class. The table below is a tentative outline of topics to be covered. Assignment due dates are also

indicated. You will have the opportunity to explore several real-world data sets through the assignments and through the Kaggle project.

Week	Topic	Reading	Assignment due dates
Jan 11, 13, 15	Introduction to machine learning, Linear models for regression: regression as optimization, gradient descent	Chapter 1.1 - 1.4 Chapter 7.1-7.3	
Jan 18, 20, 22	Linear models for regression: MLE parameter estimation, basis function expansion, regularization to control overfitting, crossvalidation to select regularization parameters, ridge regression, statistical properties of parameter estimators, global and local regression models	Chapter 7.3-7.6, Chapter 6.2, 6.4, 6.5	Assignment 1 due January 22 at 8 pm on Owlspace
Jan 25, 27, 29	Linear models for classification: discriminative models, logistic regression, MLE estimation of logistic regression model, stochastic gradient descent, Newton's method, IRLS method, L1 and L2 regularization, lasso, ridge and subset selection, coordinate descent algorithms, gradient projection, multi-class logistic regression.	Chapter 8.1,8.2, 8.3, 8.4, 8.6, Chapter 13.1, 13.2, 13.3, 13.4, 13.5	
Feb 1,3,5	Linear models for classification: generative models, MLE estimation of model parameters: Gaussian discriminant analysis, Naive Bayes (Bernoulli and Gaussian), controlling overfitting.	Chapter 3.1-3.5, Chapter 4.1-4.3, 4.6	Assignment 2 due on Feb 5 at 8 pm on Owlspace

Feb 8,10,12	Linear models for classification: online learning and the perceptron algorithm, properties of the perceptron algorithm. Non-linear models for classification: kernel-based methods, fitting kernel models, Mercer's theorem, the kernel trick — kernelizing learning algorithms.	Chapter 8.5, Chapter 14.1, 14.2, 14.4, 14.7	
Feb 15, 17, 19	Sparse kernel methods for classification: Support vector machines, maximum margin classification, kernelizing support vector machines, extension to inseparable data, the hinge loss function, support vector regression.	Chapter 14.5	Assignment 3 due on Feb 19 at 8 pm on Owlspace.
Feb 22, 24, 26	Practical advice for Kaggle project, feature selection, feature engineering, bias and variance, model selection, crossvalidation to select hyper parameters, A/B testing, tools for machine learning, brief introduction to convolutional neural networks.	Chapter 16.5, 16.7, 16.8, Chapter 28.1, 28.2, 28.3	Midterm examination on Feb 25 from 7-10 pm in TBA location.
	SPRING BREAK		
Mar 7,9,11	Non-linear models for classification and regression: neural networks as stacked logistic models, prediction by forward propagation and estimating parameters by back propagation, deep neural networks	Chapter 16.5, 16.6.1, Chapter 28.3	Assignment 4 due on Mar 11 at 8 pm on Owlspace
Mar 14, 16, 18	Non-linear models for classification and regression: more adaptive basis function	Chapter 16.2, 16.3	

	models: CART, GAMs, random forests		
Mar 21, 23, 25	Non-linear models for classification and regression: ensemble models, bagging, boosting	Chapter 16.4, 16.6	Assignment 5 due on Mar 25 at 8 pm on Owlspace
Mar 28, 30	Latent variable models: k-means clustering, mixture models, parameter estimation for mixture models, the EM algorithm, model selection for latent variable models, PCA	Chapter 11.1, 11.2, 11.3, 11.4, 11.5, Chapter 12.2, 12.3, 12.4	
Apr 4,6,8	Latent variable models for discrete data: LDA, RBM. Deep generative models: deep Boltzmann machines, deep belief networks	Chapter 27.1, 27.3, 27.7, Chapter 28.3	Assignment 6 due on Apr 8 at 8 pm on Owlspace
Apr 11, 13, 15	Directed graphical models: Hidden Markov models Undirected graphical models: Markov random fields	Chapter 17.1-17.5, Chapter 19.1-19.6	
Apr 18, 20, 22	Poster presentations and course wrap up		Poster due Apr 18 at 8 AM on Owlspace. Project report due Apr 22 at 8 pm on Owlspace. Final exam scheduled by registrar.