DSCI 352: Applied Machine Learning and Data Mining (Spring 2022)

Units:	4
Instructor:	Jim Blythe, PhD blythe@isi.edu (include DSCI 352 in the subject line)
Office hours:	By appointment (zoom only)
TA:	Rujun Han
Office hours:	Fridays 1-4pm
Lecture:	Tuesday, Thursday 4:00 pm - 5:50 pm
Web pages:	Piazza class page for everything except grades & code (https://piazza.com/usc/spring2022/dsci352) USC Blackboard class page for grades GitHub for code submission
Prerequisites:	DSCI 250 and Math 208
Other requirements:	Students must be able to code in Python or be willing to learn.
Tentative grading:	Programming assignments (labs) 55% Problem sets 25% Midterm exam 10% Final exam 10% Participation on piazza 5%

Catalog description: Foundational course focusing on the understanding, application, and evaluation of machine learning and data mining approaches in data intensive scenarios.

Course description: This is an introductory undergraduate course on Machine Learning and Data Mining with a focus on applications. The primary approach of instruction in this course is *Learning by Doing*. The focus of the course is to provide the students with basic understanding of Machine Learning and Data Mining algorithms and to make them use the algorithms to analyze datasets and convert them into information for decision-making.

Course objectives: Upon successful completion of this course a student will

- Broadly understand major algorithms used in machine learning.
- Understand supervised and unsupervised learning techniques.
- Understand regression methods.
- Understand resampling methods, including cross-validation and bootstrap.
- Understand decision trees, dimensionality reduction, regularization, clustering, and kernel methods.
- Understand feedforward neural networks and deep learning.
- Understand map reduce and its use in mining massive data.
- Understand methods for mining association rules.
- Understand how recommender systems work.

Exam dates:

- **Midterm exam:** Thursday March 10th, 4:00-5:50 pm.
- **Final exam:** as set by the university.

Textbooks:

- Required:
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013. (ISLR) Available at https://statlearning.com
  - Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ulman, Mining Massive Data Sets, 3rd Edition, Cambridge University Press, 2014. (MMDS) Available at https://www.mmds.org
- Recommended:

– The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition Authors: Trevor Hastie, Robert Tibshirani, and Jerome Friedman; Springer; 2008. (ESL) ISBN-13: 978-0387848570

Homework policy:

- Homework is assigned on an approximately weekly basis. A two-day grace period can be used for each homework with 10% penalty. *Absolutely no late homework will be accepted after the grace period. A late assignment results in a zero grade.* The only exception is a medical or family emergency.
  Important Note: If you have emergencies, you should state them the homework deadline, not at the end of the semester.

- Homework solutions should be typed or scanned using scanners or mobile scanner applications like CamScanner and uploaded (photos taken by cellphone cameras and in formats other than pdf will NOT be accepted). Programs and simulation results have to be uploaded on github as well.

- Poor internet connection, failing to upload properly, or similar issues are NOT acceptable reasons for late submissions. If you want to make sure that you do not have such problems, submit homework eight hours earlier than the deadline. Please do not ask the instructors to make individual exceptions.

- Students are encouraged to discuss homework problems with one another, but each student must do their own work and submit individual solutions written/ coded in their own hand. Copying the solutions or submitting identical homework sets is written evidence of cheating. The penalty ranges from F on the homework or exam, to an F in the course, to recommended expulsion.

- Posting the homework assignments and their solutions to online forums or sharing them with other students is strictly prohibited and infringes the copyright of the instructor. Instances will be reported to USC officials as academic dishonesty for disciplinary action.

Make-up Exams: No make-up exams will be given. If you cannot make the above dates due to a class schedule conflict or personal matter, you must drop the class or accept a zero grad in the exam. In the case of a required business trip or a medical or family emergency, a signed letter from your manager or counselor or physician has to be submitted. This letter must include the contact of your physician or counselor or manager.
Important Note: If you have emergencies, you should state them before taking the exam.

Attendance: Students are required to attend all the lectures and discussion sessions and actively participate in class discussions. Use of cellphones and laptops is prohibited in the classroom. If you need your electronic devices to take notes, you should discuss with the instructor at the beginning of the semester.

Tentative Course Outline

Tuesday	Thursday
Jan 11th 1 **Introduction to Statistical Learning** (ISLR Chs. 1,2, ESL Chs. 1,2) Motivation: Big Data Supervised vs. Unsupervised Learning	13th 2 **Introduction to Statistical Learning** (ISLR Chs. 1,2, ESL Chs. 1,2) Regression, Classification The Regression Function Nearest Neighbors
18th 3 **Introduction to Statistical Learning** (ISLR Chs.1,2, ESL Chs.1,2) Model Assessment The Bias-Variance Trade-off No Free Lunch Theorem	20th 4 **Linear Regression** (ISLR Ch.3, ESL Ch. 3) Estimating Coefficients Estimating the Accuracy of Coefficients
25th 5 **Linear Regression** (ISLR Ch.3, ESL Ch. 3) Variable Selection and Hypothesis Testing Multiple Regression Analysis of Variance and the F Test Lab 0 Due (Not Graded)	27th 6 **Linear Regression** (ISLR Ch.3, ESL Ch. 3) Stepwise Variable Selection Qualitative Variables
Feb 1 7 **Classification** (ISLR Ch. 4, ESL Ch. 4) Multi-class and Multi-label Classification Logistic Regression Class Imbalance Hypothesis Testing and Variable Selection Lab 1 Due	3rd 8 **Classification** (ISLR Ch. 4, ESL Ch. 4) Subsampling and Upsampling SMOTE Multinomial Regression PS 1 Due
8th 9 Classification (ISLR Ch. 4, ESL Ch. 4) Bayesian Linear Discriminant Analysis	10th 10 **Classification** (ISLR Ch. 4, ESL Ch. 4) Measures for Evaluating Classifiers Quadratic Discriminant Analysis* Comparison with K-Nearest Neighbors The Naive Bayes' Classifier Text Classification Feature Creation for Text Data Handling Missing Data Lab 2 Due
15th 11 **Resampling Methods** (ISLR Ch. 5, ESL Ch. 7) Model Assessment Validation Set Approach Cross-Validation The Bias-Variance Trade-off for Cross-validation The Bootstrap Bootstrap Confidence Intervals	17th 12 **Linear Model Selection and Regularization** (ISLR Ch.6, ESL Ch. 3) Subset Selection AIC, BIC, and Adjusted R^2 Shrinkage Methods Ridge Regression PS 2 Due
22nd 13 **Linear Model Selection and Regularization** (ISLR Ch.6, ESL Ch. 3) The LASSO Elastic Net Dimension Reduction Methods*	24th 14 **Tree-based Methods** (ISLR Ch. 8, ESL Chs. 9, 10) Regression and Classification Trees Cost Complexity Pruning PS 3 Due

March 1st 15 **Tree-based Methods** (ISLR Ch. 8, ESL Chs. 9, 10, 16) Bagging, Random Forests, and Boosting*	3rd 16 **Support Vector Machines** (ISLR Ch. 9, ESL Ch. 12) Maximal Margin Classifier Support Vector Classifiers
8th 17 **Support Vector Machines** (ISLR Ch. 9, ESL Ch. 12) The Kernel Trick L1 Regularized SVMs Multi-class and Multilabel Classification The Vapnik-Chervonenkis Dimension* Support Vector Regression Lab 3 Due	10th 18 Midterm
13th recess	17th recess
22nd 19 **Neural Networks and Deep Learning** (ESL Ch. 11, DL Ch. 6) The Perceptron Feedforward Neural Networks Backpropagation and Gradient Descent Overfitting	24th 20 **Neural Networks and Deep Learning** (DL Chs. 6, 7) Regularization Early Stopping and Dropout Convolutional Neural Networks* PS 5 Due
29th 21 **Unsupervised Learning** (ISLR Ch. 10, ESL Ch. 14) K-Means Clustering Hierarchical Clustering	31st 22 **Unsupervised Learning** (ISLR Ch. 10, ESL Ch. 14) Practical Issues in Clustering Lab 4 due
April 5th 23 **Unsupervised Learning** (ISLR Ch. 10, ESL Ch. 14) Principal Component Analysis* Anomaly Detection* PS 6 Due	7th 24 **Active and Semi-Supervised Learning** Semi-Supervised Learning Self-Training Co-Training Yarowsky Algorithm Refinements Active vs. Passive Learning Stream-Based vs. Pool-Based Active Learning Query Selection Strategies
12th 25 **Introduction to Data Mining** (MMDS Ch. 1) Motivations Relationship with Machine Learning Summarization Bonferroni Correction PS 7 Due	14th 26 **Map Reduce and New Stack Software** (MMDS Ch. 2) Distributed Computing Distributed File Systems Map Reduce for Word Counting

19th 27 **Frequent Itemsets and Association Rules** (MMDS Ch. 6, IDM Ch. 6) The Market-Basket Model Applications Association Rules High-Confidence Rules Lab 5 Due	21st 28 **Frequent Itemsets and Association Rules** (MMDS Ch. 6, IDM Ch. 7) Algorithms for Rule-Mining
26th 29 **Recommendation Systems*** (MMDS Ch. 9) Content-Based Recommendation	28th 30 **Recommendation Systems*** (MMDS Ch. 9) Collaborative Filtering

Items marked with * will be covered only if time permits.