INF 552: Machine Learning for Data Science (Summer 2020)

Units:	4
Instructor:	Mohammad Reza Rajati, PhD PHE 414 rajati@usc.edu – Include INF 552 in subject
Office Hours:	Tuesday 5:30 pm –6:30 pm
TA(s):	Nripsuta Saxena nsaxena@usc.edu – Include INF 552 in subject
Office Hours:	TBD
	Jiao Sun jiaosun@usc.edu – Include INF 552 in subject
Office Hours:	TBD
Course Producer(s):	TBD @usc.edu – Include INF 552 in subject
Lecture:	Thursday, Tuesday, 3:00 pm –5:30 pm online
Webpages:	Piazza Class Page for everything except grades and USC Blackboard Class Page for grades and GitHub for code submission – All HWs, handouts, solutions will be posted in PDF format – *Student has the responsibility to stay current with webpage material*
Prerequisite:	Prior courses in multivariate calculus, linear algebra, probability, and statistics. – This course is a prerequisite to INF 558.
Other Requirements:	Computer programming skills. Using Python is mandatory. Students must know Python or must be willing to learn it.
Tentative Grading:	Assignments 50% Midterm Exam 20% Final Exam 30% Participation on Piazza* 5%

Letter Grade Distribution:

$\geq$ 93.00	A	73.00 - 76.99	C
90.00 - 92.99	A-	70.00 - 72.99	C-
87.00 - 89.99	B+	67.00 - 69.99	D+
83.00 - 86.99	B	63.00 - 66.99	D
80.00 - 82.99	B-	60.00 - 62.99	D-
77.00 - 79.99	C+	$\leq$ 59.99	F

Disclaimer: Although the instructor does not expect this syllabus to drastically change, he reserves every right to change this syllabus any time in the semester.

Note on e-mail vs. Piazza: If you have a question about the material or logistics of the class and wish to ask it electronically, please post it on the piazza page (not e-mail). You may post it anonymously if you wish. Often times, if one student has a question/comment, other also have a similar question/comment. Use e-mail with the professor, TA, graders only for issues that are specific to you individually (e.g., a scheduling issue or grade issue).

Catalogue Description: Practical applications of machine learning techniques to real-world problems. Uses in data mining and recommendation systems and for building adaptive user interfaces.

Course Description: This is a foundational course with the primary application to data analytics, but is intended to be accessible both to students from technical backgrounds such as computer science, computer engineering, electrical engineering, or mathematics; and to students from less technical backgrounds such as business administration, communication, accounting, various medical specializations including preventative medicine and personalized medicine, genomics, and management information systems. A basic understanding of engineering and/or technology principles is needed, as well as basic programming skills, sufficient mathematical background in probability, statistics, and linear algebra.

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 hidden Markov models and graphical models.
- Understand feedforward and recurrent neural networks and deep learning.

Exam Dates:

- **Midterm Exam:** Friday June 26, 3-4:50 PM.
- **Final Exam:** Tuesday, July 28, 3:00 - 4:50 AM

Textbooks:

- **Required Textbook:**

  1. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013. (ISLR)
     Available at `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf`

- **Recommended Textbooks:**

  1. *Applied Predictive Modeling*, 1st Edition
     Authors: Max Kuhn and Kjell Johnson; Springer; 2016. **ISBN-13:** 978-1-4614-6848-6
  2. *Machine Learning: A Concise Introduction*, 1st Edition
     Author: Steven W. Knox; Wiley; 2018. **ISBN-13:** 978-1-119-43919-6
  3. *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
  4. *Machine Learning: An Algorithmic Perspective*, 2nd Edition
     Author: Stephen Marsland; CRC Press; 2014. **ISBN-13:** 978-1-4614-7137-0
  5. *Deep Learning*, 1st Edition
     Authors: Ian Goodfellow, Yoshua Bengio, and Aaron Courville; MIT Press; 2016. (DL) **ISBN-13:** 978-0262035613
  6. *Neural Networks and Learning Machines*, 3rd Edition
     Author: Simon Haykin; Pearson; 2008. **ISBN-13:** 978-0131471399
  7. *Neural Networks and Deep Learning: A Textbook*, 1st Edition
     Authors: Charu Aggrawal; Springer; 2018. **ISBN-13:** 978-3319944623
  8. *Introduction to Machine Learning*, 2nd Edition
     Author: Ethem Alpaydine; MIT Press; 2010. (AL) **ISBN-13:** 978-8120350786
  9. *Machine Learning*, 1st Edition
     Authors: Tom M. Mitchell; McGraw-Hill Education; 1997. **ISBN-13:** 978-0070428072

Grading Policies:

- The letter grade distribution table guarantees the *minimum* grade each student will receive based on their final score. When appropriate, relative performance measures will be used to assign the final grade, at the discretion of the instructor.
  - Final grades are non-negotiable and are assigned at the discretion of the instructor. If you cannot accept this condition, you should not enroll in this course.
- Your lowest homework grade will be dropped from the final grade.
- *Participation on Piazza has up to 5% extra credit, which is granted on a competitive basis *at the discretion of the instructor*.

- **Homework Policy**
  - Homework is assigned on an approximately weekly basis. A one-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.*
  - In case of *documented illness* or *grave family* situations, exceptions can be made to the late submission policy.
  - 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 instructor to make individual exceptions.
  - Homework solutions and simulation results should be typed or *scanned* using scanners or mobile scanner applications like CamScan and uploaded on blackboard (photos taken by cell-phone cameras and in formats other than pdf will NOT be accepted). Programs and simulation results have to be uploaded on GitHub as well.
  - 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.
- **Exam Policy**
  - **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. In the case of a required business trip or a medical emergency, a signed letter from your manager or physician has to be submitted. This letter must include the contact of your physician or manager.
  - Midterm and final exams will be closed book and notes. Calculators are allowed but computers and cell-phones or any devices that have internet capability are not allowed. One letter size cheat sheet (back and front) is allowed for the midterm. Two letter size cheat sheets (back and front) are allowed for the final.
  - All exams are cumulative, with an emphasis on material presented since the last 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.

Important Notes:

- Textbooks are secondary to the lecture notes and homework assignments.
- Handouts and course material will be distributed.
- Please use your USC email to register on Piazza and to contact the instructor and TAs.

Tentative Course Outline

Thursday	Tuesday
May 21st **1** **Introduction to Statistical Learning** (ISLR Chs.1,2, ESL Chs.1,2) Motivation: Big Data Supervised vs. Unsupervised Learning	26th **2** **Introduction to Statistical Learning** (ISLR Chs.1,2, ESL Chs.1,2) Regression, Classification The Regression Function Nearest Neighbors Model Assessment The Bias Variance Trade-off No Free Lunch Theorem
28th **3** **Linear Regression** (ISLR Ch.3, ESL Ch. 3) Estimating Coefficients Estimating the Accuracy of Coefficients Variable Selection and Hypothesis Testing Multiple Regression **Homework 0 Due**	June 2nd **4** **Linear Regression** (ISLR Ch.3, ESL Ch. 3) Analysis of Variance and The F-test Qualitative Variables Comparison with K-Nearest Neighbors
4th **5** **Classification** (ISLR Ch. 4, ESL Ch. 4) Multi-class and Multi-label Classification Logistic Regression Hypothesis Testing and Variable Selection Class Imbalance Subsampling and Upsampling SMOTE Multinomial Regression **Homework 1 Due**	9th **6** **Classification** (ISLR Ch. 4, ESL Ch. 4) Bayesian Linear Discriminant Analysis Measures for Evaluating Classifiers Quadratic Discriminant Analysis* Comparison with K-Nearest Neighbors
11th **7** **Classification** (ISLR Ch. 4, ESL Ch. 4) The Naïve Bayes' Classifier Text Classification Feature Creation for Text Data Handling Missing Data **Homework 2 Due**	16th **8** **Resampling Methods** (ISLR Ch. 5, ESL Ch. 7) Model Assessment Validation Set Approach Cross-Validation The Bias-Variance Trade-off for Cross The Bootstrap Bootstrap Confidence Intervals

Thursday	Tuesday
18th **9** **Linear Model Selection and Regularization** (ISLR Ch.6, ESL Ch. 3) Subset Selection AIC, BIC, and Adjusted R^2 Shrinkage Methods Ridge Regression The LASSO Elastic net Dimension Reduction Methods*	23rd **10** **Tree-based Methods** (ISLR Ch. 8, ESL Chs. 9, 10) Regression and Classification Trees Bagging, Random Forests, and Boosting
25th **11** **Support Vector Machines** (ISLR Ch. 9, ESL Ch. 12) Maximal Margin Classifier Support Vector Classifiers The Kernel Trick Support Vector Machines L1 Regularized SVMs Multi-class and Multilabel Classification The Vapnik-Chervonenkis Dimension Support Vector Regression **Homework 3 Due**	30th **12** **Unsupervised Learning** (ISLR Ch. 10, ESL Ch. 14) K-Means Clustering Hierarchical Clustering
$\boxed{\text{July 2nd}}$ **13** **Unsupervised Learning** (ISLR Ch. 10, ESL Ch. 14) Practical Issues in Clustering Principal Component Analysis Association Rules* Gaussian Mixtures and Soft K-Means* **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 **Homework 4 Due**	7th **14** **Neural Networks and Deep Learning** (ESL Ch. 11, DL Ch. 6) The Perceptron Feedforward Neural Networks

Thursday	Tuesday
9th **15** **Neural Networks and Deep Learning** (ESL Ch. 11, DL Ch. 6) Feedforward Neural Networks Backpropagation and Gradient Descent Overfitting Autoencoders and Deep Feedforward Neural Networks Regularization Early Stopping and Dropout **Homework 5 Due**	14th **16** **Neural Networks and Deep Learning** (DL Chs. 6, 7, 9, 10) Adversarial Training* Convolutional Neural Networks Sequence Modeling Recurrent Neural Networks Sequence-to-Sequence Modeling* Long Short Term Memory (LSTM) Neural Networks
16th **17** **Hidden Markov Models** (AL Ch. 15) Principles The Viterbi Algorithm Applications of HMMs **Homework 6 Due**	21st **18** **Graphical Models*** (ESL Ch. 17, AL Ch. 14) Markov Graphs and Their Properties Bayesian Networks Belief propagation* Restricted Boltzmann Machines **Reinforcement Learning*** Definitions Task-Reward-Policy Formulation Total Discounted Future Reward Optimal Policy Value Function Q-Function The Bellman Equation Q-Learning Exploration- Exploitation Temporal Difference Learning Extensions to Stochastic Environments and Rewards Deep Reinforcement Learning
23rd **19** **Fuzzy Systems*** Fuzzy Sets Set Operations T-norms, T-conorms, and Fuzzy complements Cylindrical Extensions and Fuzzy Relations Fuzzy If-Then Rules as Association Rules Inference from Fuzzy Rules Fuzzification and Defuzzification Learning Fuzzy Rules from Examples The Wang-Mendel Algorithm **Homework 7 Due**	28th **20** **Final Exam**

Notes:

- Items marked by * will be covered only if time permits.

Statement on Academic Integrity: USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. SCampus, the Student Guidebook, contains the University Student Conduct Code (see University Governance, Section 11.00), while the recommended sanctions are located in Appendix A. See: http://scampus.usc.edu.

Emergency Preparedness/Course Continuity in a Crisis In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies. See the university's site on Campus Safety and Emergency Preparedness: http://preparedness.usc.edu

Statement for Students with Disabilities: Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. Website: http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html

(213) 740-0776 (Phone), (213) 740-6948 (TDD only), (213) 740-8216 (FAX) ability@usc.edu.