EE 660 **Course Syllabus** Fall 2021

V2.1_SoC, 9/1/2021

Machine Learning II: Mathematical Foundations and Methods

Administrative information

Times and days

Lecture: Tu Th 2:00 – 3:50 PM, OHE 122, online, and through DEN@Viterbi.

Discussion session: Friday 3:30 – 4:20 PM, OHE 122, online and through DEN@Viterbi.

All times are given in Pacific Time (PT), including Pacific Daylight Time (PDT) or Pacific Standard Time (PST), whichever is active on the given date.

Lectures and discussion sessions can be attended in person for students able to do so; they can also be viewed by live streaming and by archived video any time after the live event, for students unable to attend live. The live streaming provides for you optionally to interact via audio, video, text, and other software-provided actions.

Course Contact Information

Professor

B. Keith Jenkins
Office: EEB 404A
Email: jenkins@sipi.usc.edu [please include "EE 660" in the subject line]
Office hours: Tu Th 5:00-6:00 PM PT (Zoom link is provided in D2L, in the calendar entry)
Phone: (213) 740-4149

Teaching Assistants

Fernando Valladares Monteiro
fvallada@usc.edu [please include "EE 660" in the subject line]
Office hours: W 1:00-2:30 PM (Zoom link is provided in D2L, in the calendar entry)

Zhiruo Zhou
Email: zhiruozh@usc.edu [please include "EE 660" in the subject line]
Office hours: Fr 1:30-3:00 PM (Zoom link is provided in D2L, in the calendar entry)

Grader

Name: Qirui Sun
Email: qiruisun@usc.edu [please include "EE 660" in the subject line]

Catalogue description

Supervised, semi-supervised, and unsupervised machine learning; domain adaptation and transfer learning. Feasibility of learning, model complexity, and performance (error) on unseen data..

Extended course description

Foundations of machine learning, which apply to many or all algorithmic approaches, will be studied. These will include feasibility of learning; complexity of hypothesis sets; bias/variance tradeoff; regularization, overfitting and underfitting of models to data; model selection and assessment; prediction of performance on unseen data. Particular methods that are key to machine learning will also be covered, and include linear and nonlinear techniques for regression and classification. Methods studied for classification by semi-supervised learning (using some labeled data and some unlabeled data for training), and for clustering by unsupervised learning (using only unlabeled data), will include statistical and distribution-free approaches. Techniques for domain adaptation and transfer learning (in which a system trained in one domain adapts or learns to work in a different domain), will also be covered. Definitions of, and techniques for, human interpretable machine learning systems will be introduced. Students will be exposed to examples of techniques run on both synthetic and real-word data, through examples in lectures and the reading, as well as in homework problems and in the course project. This course is intended for MS and PhD students in ECE and related Viterbi departments, who have an interest (and some prior coursework) in machine learning.

Learning Objectives

After successfully completed this course, the student will:

(1) Have a solid foundation in machine learning principles and theory, and the capability to apply them to problems.

(2) Have intuition grounded in theory for different machine learning realms.

(3) Understand and be able to use common and successful methods (techniques and algorithms) in machine learning.

(4) Have sufficient foundation and knowledge to be able to learn about many of the plethora of machine learning techniques that exist and that are being created, on his or her own as needed.

(5) Be able to adapt existing algorithms, and create new algorithms, to problems and domains that aren't yet well served by existing approaches.

Preparation

Prerequisites: EE 503, EE 510, and EE 559.

Recommended Preparation: Experience with Python 3 at the level of EE 541 or EE 559 (2021 versions), including the use of modules, functions, classes, and OOP. Familiarity with general machine learning methods including regression and classification, and with computational complexity, at the level of EE 559 (2021 versions).

Computer Software Requirements:

Students are required to use Python 3 for all homework computer problems. For the class project, students may use Python and/or C/C+. (To use other languages for the class project, check first with the TA or instructor.) All students will be responsible for installing and maintaining their own Python distribution (e.g., from https://www.anaconda.com/products/individual).

Python software packages will be used for some of the homework computer problems, including numpy, pandas, scipy, matplotlib, and scikit-learn.

The Murphy textbook has an accompanying software package in Python, available to use for free at: https://github.com/probml/pyprobml/ ; caveat: it is essentially a beta version and so may still have bugs. We recommend the standard Python packages instead for most of your computer work based on libraries.

Textbooks, reading materials, and other resources

Required textbooks and reading materials

Selected portions of books 1 and 2 below will be used for the class; you are required to have access to a copy. Please note that the total cost of these two books is approximately the same as the cost of one typical textbook in a graduate-level EE class. The other books and publications (3-7)[1] that include required reading are available for free viewing or download.

1. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective (MIT Press, Cambridge, 2012). [In short, "Murphy"] (Available at USC bookstore and online sellers)
2. Yasir S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin, Learning From Data (AMLbook.com, 2012). [In short, "AML"] (Available from USC bookstore and Amazon)
3. Wouter M. Kouw and Marco Loog, "An introduction to domain adaptation and transfer learning," Technical Report, Delft University of Technology, arXiv:1812.11806v2, 14 Jan 2019. https://arxiv.org/abs/1812.11806
4. Xiaojin Zhu and Andrew B. Goldberg, Introduction to Semi-Supervised Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers, 2009). Available for download through USC Library.

5. Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms", IEEE Trans. Neural Networks, Vol. 16, No. 3 (May 2005). A link will be provided on the course web site.
6. Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/index.html, 2019, Ch. 4-6.
7. Mengnan Du, Ninghao Liu, Xia Hu, "Techniques for Interpretable Machine Learning", arXiv:1808.00033v3 [cs.LG] 19 May 2019: https://arxiv.org/abs/1808.00033.
8. Instructor-provided notes and materials that will be posted on the course web site.

Supplementary books for your information

i. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition (Springer, 2009)

ii. Kevin P. Murphy, *Probabilistic Machine Learning: An Introduction* (MIT Press, Cambridge, 2022); complete version not yet available, but a free draft version is available at: https://probml.github.io/pml-book/book1.html .

iii. M Mohri, A. Rostamizadeh, and A. Talwalker, *Foundations of Machine Learning*, second edition (MIT Press, Cambridge, 2018)

iv. C. M. Bishop, "Pattern Recognition and Machine Learning" (Springer, 2006)

v. R. O. Duda, P. E. Hart, and D. G. Stork, , *Pattern Classification,* Second Edition (Wiley-Interscience, John Wiley and Sons, Inc., New York, 2001)

vi. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (MIT Press, Cambridge, 2016).

Course web site (Desire2Learn system)

courses.uscden.net

The site includes:

- Links to online lectures, discussion sessions, and office hours.
- Course materials (handouts, homework assignments, lecture notes, lecture videos, etc.), which will be posted as we progress through the semester.
- Link to our discussion forum (piazza) (to be set up).
- Course calendar, showing events and deadlines.
- Grade book, showing your scores on assignments to date.
- Dropboxes for uploading your completed assignments and exams, and links for viewing and retrieving your graded assignments and exams.

Description and Assessment of Assignments

Homework assignments[1]

There will be approximately one homework assignment per week. Assignments will generally include some pencil-and-paper problems, some computer problems, and some reading.

Overall approximately 50% of your homework time will be devoted to computer problems. Note that for some homework computer problems, you will be required to code up the problem yourself, without the use of libraries or software packages. For other homework computer problems, you will be encouraged to use libraries or packages. Each homework problem will specify what packages are recommended and allowed.

Homework assignments will generally be due on Mondays, and will be posted on D2L 7-10 days prior to the due date.

Course project[1]

Programming languages: The project must be coded using Python and/or C/C++. (If there are special reasons to use another language, be sure to check with the TA or instructor first.) The code will typically be a combination of code written by the project authors, and use of available libraries or packages.

Project content: Each student or team will find and choose their own dataset(s) and problem to work on, subject to certain conditions and guidelines that will be given in the project assignment. The problem assignment will also detail a set of required elements that every project must in include. Briefly, the set of required elements includes (i) use of real-world data (exceptions are allowed for projects based on exploratory machine-learning experiments that require synthetic data); (ii) complexity analysis, including hypothesis-set and sample complexity; (iii) estimation of out-of-sample (generalization) error; (iv) baseline system to compare with; (v) description of how the data is used; (vi) description of overall procedure followed; (vii) minimum of 50% of the project work using methods and techniques covered in EE 660; (viii) citation of others where appropriate; (ix) the student's or team's work, which only includes work done specifically and solely for this course.

Typical project: For the typical project, the student or team will choose one or two set(s) of real-world data, and define a set of goals. For example, one of the goals may be to use regression or classification techniques to predict the output attribute(s) y as well as possible. Additional goals could include, for example, understanding what the limitations of the final system are caused by; investigating the attributes that are most predictive (e.g., how important each feature is in predicting the outcome), and assessing why; assessing how well the trained system can adapt to another domain by using transfer learning; using unsupervised learning for feature discovery or reduction, or semi-supervised learning to evaluate effects of adding unlabeled data to the learning process; etc. There will typically be other issues to address as well, such as number of data points N not being ideal, missing or noisy data, imbalance of data set, categorical feature values, preprocessing steps, etc. A list of online resources, and tips for choosing datasets, will be provided with the project assignment.

Participants: Each project can be done by an individual student or a team of 2 students. The workload of the project will be graded accordingly; *e.g.,* 2 students should accomplish about twice the work of one student (or solve a problem that is an appropriate factor more difficult).

Deliverables: A project proposal describing the chosen topic, dataset(s), goals, plan of approach, as well as any potential pitfalls and ideas on how to overcome them, will be submitted as a homework assignment. Feedback and guidance based on the proposal will be provided. Project final report requirements and tips will also be posted; it must be typewritten, maximum suggested length of 15 pages, using a readable font, figures, and tables. A file of all code written and used must also be submitted in computer readable form; a set of code files that are runnable and can reproduce your results, may also be required.

Grading criteria. Your Final Project will be graded by the following criteria, each weighted approximately equally: Workload (difficulty of problem, amount of work); Technical approach and execution (appropriate goals, correct approach); Analysis (understanding and interpretation); Data handling (correctness and appropriateness); Performance (correctly estimated or evaluated; comparison with baseline system and work of other people if available); Report write-up (clarity, completeness, conciseness).

Dates: The project will start with a proposal homework that is assigned shortly after the midterm exam. The final written report will be due on the last day of classes (12/3/2021). There will be no oral presentations.

Student work and grading[1]

Grading breakdown

Assignment	% of Grade
Homework	25
Midterm exam	20
Course Project	25
Final exam	25
Online participation	5
TOTAL	100

Exam dates

Midterm exam: Tuesday 10/12/2021, 2:00-4:00 PM PDT[2].

Final exam (per university's official schedule): Thursday, 12/9/2021, 2:00-4:00 PM PST[2].

[2] Everyone that is located in a time zone compatible with this time slot, must be available for these exam periods. If we have students in non-compatible time zones, then we will consider also having a second time slot for them, or changing the midterm time for everyone if there is a time that all students can attend.

Policy on Collaboration and Individual Work in this Class

Collaboration on techniques for solving homework assignments and computer problems is allowed, and can be helpful; however, each student is expected to work out, code, and write up his or her own solution. Use of other solutions to homework assignments or computer problems, from any source including other students, before the assignment is turned in, is not permitted.

For class projects, general collaboration to resolve issues, or to clarify technical material, is allowed. Use of internet as well as journal and conference literature is encouraged. However, each student (or team) does their own work and writes up their own report. The author(s) of the report are presenting themselves as having done the work described in the report. Any reported work, explanations, information, or code that is obtained from others must be cited as such; instructions for doing this will be given with the project assignment. Including such work in the report without citing it amounts to plagiarism.

Of course, collaboration on exams is not permitted.

Please also see below for additional policies that apply to all USC classes.

Course Outline[1]

[x] = approximate number of lectures

Introduction [1.5]

1. Course introduction

   Learning modes (supervised, semi-supervised, unsupervised, transfer, interpretable);
   Learning theory and applications;
   Course administrative info and syllabus

2. Key issues and concepts

Supervised Learning [6]

3. Probabilistic machine learning for regression [2]

   Maximum-likelihood and MAP estimation;
   Regularizers and interpretation from Lagrangian optimization
   Ridge, lasso, and bridge regression.

4. Logistic regression [1]

   Cross-entropy error; maximum likelihood

5. Graphical techniques for supervised learning [3]

   Classification and regression trees (CART);
   Bagging and boosting;
   Random forest;
   Adaboost

Learning theory and its implications [6]

6. Learning theory. feasibility of learning. PAC learning [2]

   Hypothesis sets and their complexity, growth function and VC dimension;

7. Implications and extensions of learning theory [2]

   PAC learning bounds, generalization-error bounds
   Multiclass problems, regression problems;
   Error measures; regularization

8. Practical applications [1.5]

   Study of overfitting;
   Dataset methodology, validation and test;
   Dataset size and complexity; effects on generalization error bounds

9. Concluding remarks – a few principles to be aware of [0.5]

   Occam's Razor, Axiom of Non-Falsifiability, Data snooping, Sampling bias

Midterm [1.5]

10. Review for midterm; midterm exam

Domain adaption and transfer learning [3]

11. Introduction and theory [1]

    Examples and realms;
    Cross-domain generalization error bounds;

12. Approaches and methods [2]

    Domain/data shifts: prior shift, covariate shift;
    Importance weighting; subspace mapping;, domain-invariant spaces;

Semi-supervised learning [3.5]

13. Introduction and methods 1 [2]

    Overview, assumptions;
    Self-training algorithms;
    Mixture models and expectation maximization

14. Methods 2 [1.5]

    Co-training; graphical techniques; SS SVM

Unsupervised learning [3]

15. Statistical techniques [1.5]

    Maximum likelihood, expectation maximization

16. Nonstatistical techniques [1]

    Similarity measures; hierarchical and graphical techniques

17. Evaluating cluster quality and choosing *K* [0.5]

Human interpretability and course conclusions [2.5]

18. Human interpretability overview [2]

    Using interpretable models; model agnostic methods; example-based explanations; examples using deep neural networks

19. Course wrap-up and review for final exam [0.5]

[1] Instructor reserves the right to make changes as he deems appropriate during the semester, to accommodate student needs, semester timing, and changes in COVID incidents.

Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems

Student Health Counseling Services - (213) 740-7711 – 24/7 on call
engemannshc.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-4900 – 24/7 on call
engemannshc.usc.edu/rsvp
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) | Title IX - (213) 740-5086
equity.usc.edu, titleix.usc.edu
Information about how to get help or help a survivor of harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following protected characteristics: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations.

Bias Assessment Response and Support - (213) 740-2421
studentaffairs.usc.edu/bias-assessment-response-support

Avenue to report incidents of bias, hate crimes, and microaggressions for appropriate investigation and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
studentaffairs.usc.edu/ssa
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.