EE 660 **Course Syllabus** Fall 2022

V1.0_SoC, 8/21/2022

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 are set up to be attended in person; they can also be viewed remotely by live streaming, or by archived video after the live event. You are encouraged to attend (remotely or in person) every lecture and discussion session live.

Course Contact Information

Professor

B. Keith Jenkins
Office: EEB 404A
Email: jenkins@sipi.usc.edu [please include "EE 660" in the subject line]
Phone: (213) 740-4149
Office hours: Tu Th 5:00-6:00 PM PT. Office hours will be available in person (please wear a mask) and over zoom, unless announced otherwise. Zoom link will be provided on the D2L Course Calendar, and location will be announced.

Teaching Assistants

Thanos Rompokos
Email: rompokos@usc.edu [please include "EE 660" in the subject line]
Office hours and location: TBA

Zhiruo Zhou
Email: zhiruozh@usc.edu [please include "EE 660" in the subject line]
Office hours and location: TBA

Yao Zhu
Email: yaozhu@usc.edu [please include "EE 660" in the subject line]
Office hours and location: TBA

Graders

Name: Hardik Prajapati
Email: hprajapa@usc.edu [please include "EE 660" in the subject line]

Name: TBD
Email: TBA [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

Machine learning (ML) has had almost explosive growth, and now is being used in almost all aspects of life, including for example space exploration; better human-computer interfaces and more human-like computers; engineering innovative and improved devices, autonomous and efficient transportation; making advances in the hard sciences, medicine, and psychology; and improving finance and business. This class will focus on (i) theory that has been developed to understand ML better, and to use ML more effectively; (ii) ML methods that can help make this happen; and (iii) different modes and scenarios of ML so you can use it in a wide assortment of problem types.

Specifically, topics will include the following. 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, with an emphasis on graphical techniques. 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 realm adapts or learns to work in a different realm), will also be covered. Definitions of, and techniques for, human interpretation of machine learning systems (to understand why or how the ML system decides on its output predictions) 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 class 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 (2021-2022 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 (2021-2022 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.

Textbooks, reading materials, and other resources

Required textbooks and reading materials

Assigned readings (and some exercises) are selected from the reference sources listed below. Most of these are available for free download or viewing; [2] must be purchased, but is very reasonably priced. We will first use [1], and then start using [2] in Week 4. [3]-[7] (and occasionally [1] again) will be used after that.

1. Kevin P. Murphy, *Probabilistic Machine Learning: An Introduction* (MIT Press, Cambridge, 2022); https://probml.github.io/pml-book/book1.html . [In short, "Murphy book 1", or just "Murphy"] Now available for purchase in hardcopy form; also the latest pre-print is available for download (as of 8/19/2022). (This book has some of the topics and required readings for EE 660; it also covers many of the EE 559 topics so can provide a good review or reference.)
2. Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin, L*earning From Data* (AMLbook.com, 2012). [In short, "AML"] (Available from Amazon, and will be available from USC bookstore.)
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*, Second Edition, https://christophm.github.io/interpretable-ml-book/ , 2022.
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). 12th printing available for download at https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12.pdf (Excellent book as a second resource, especially for graphical/tree based techniques in supervised learning.)

ii. Kevin P. Murphy, *Probabilistic Machine Learning: Advanced Topics* (MIT Press, Cambridge, 2023); https://probml.github.io/pml-book/book2.html . Not yet published; preprint is available for free download. (Covers a plethora of ML topics for further exploration if you're interested. It does include a few of the topics of EE 660; however the required EE 660 readings are in other references.)

iii. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective (MIT Press, Cambridge, 2012). (Forerunner of Murphy's second edition books; this first edition often has more detail of the topics, although tends to be less pedagogical.)

iv. M Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, second edition (MIT Press, Cambridge, 2018). (Good for a more theoretical viewpoint, and has extensions of treatments in the AML book [2].)

v. C. M. Bishop, "Pattern Recognition and Machine Learning" (Springer, 2006). (EE 559 textbook.)

vi. R. O. Duda, P. E. Hart, and D. G. Stork, , *Pattern Classification,* Second Edition (Wiley-Interscience, John Wiley and Sons, Inc., New York, 2001) (Classical book on pattern classification, with excellent treatment of many of its topics.)

vii. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (MIT Press, Cambridge, 2016). (Good introduction of deep learning, but missing some of the most recent techniques.)

Comment: The Murphy books [1] and [ii] have accompanying software notebooks in Python, as well as some tutorials, at: https://github.com/probml/pyprobml/tree/master/notebooks. Caveat: most are essentially beta versions so may still have "rough edges". You may find the notebooks useful if you would like to try running variants of the algorithms featured in figures in the books.

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 and discussion notes, lecture and discussion 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 lectures, discussion sessions, and office hours.
- Grade book, showing your scores on assignments to date.
- Dropboxes for uploading your completed assignments, and links for viewing and retrieving your graded assignments.

Description and Assessment of Assignments

Homework assignments[1]

There will be approximately one homework assignment per week, although some weeks will have no homework assignment. 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. All coding for homework problems will be done in Python.

Course project[1]

Overview: There will be two types of projects, and you can choose either one.

For Type 1 projects, you will choose a dataset and develop a problem topic of interest to you. Each student or team will define their project goals and their approach. The purpose is to apply techniques and theory learned in EE 660 on a realistic problem, from initial dataset through learning and prediction, analysis of results, and understanding of what your system does. It may involve solving a supervised learning problem, and/or transferring the ML system from one realm to another, and/or using some labeled and some unlabeled data, or other learning modes.

For Type 2 projects, you will design an experiment that you will do on your computer, to test or learn about some aspect of machine learning. It would typically involve using

machine learning theory to hypothesize what will happen, then try it by computer simulation on synthetic data you generate (or perhaps on real data), and see if the result verifies (or disproves) your hypothesis.

For both project types, it is hoped you will find a problem that you're enthusiastic about, so you will enjoy working on the project while you "learn by doing".

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 you, for tasks that are specific to your project, and use of available libraries or packages, for tasks that can use well established algorithms.

Participants: Each project can be done by an individual student or a team of 2 students. The workload and problem difficulty of the project will be graded accordingly (i.e., a team project es expected to accomplish more work, and/or solve a more difficult problem, than an individual project).

Deliverables: A project proposal describing the chosen topic, dataset(s), goals, and plan of approach, will be submitted as a homework assignment, mid-way through the semester. We will provide feedback and guidance based on your proposal.

Your main outputs to us will be a typewritten final report describing your project work and results, and a file with all your code. Requirements and tips will be posted. Suggested length of the report is less than (or equal to) 15 pages, using a readable font size (typically 12 point), and including figures and tables.

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); System 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 the proposal, which will be assigned shortly after the midterm assignment. The final written report will be due on or about the last day of classes (12/2/2022). There will be no oral presentations.

Student work and grading[1]

Grading breakdown

Assignment	% of Grade
Homework	20
Course project	24
Exam 1: midterm assignment	24
Exam 2: in-class exam	24
Class participation*	8
TOTAL	100

*Class participation includes: real-time participation (in person or over Webex, during lectures), and piazza participation (questions, comments, and answers on piazza). For both real-time and piazza, "good" questions, comments, and/or answers get more credit.

Exam dates

Exam 1 will be a take-home midterm assignment, and will be a 5-7 day assignment in which each student does their own work without collaboration. Dates TBD.

Exam 2 will be on campus (in person) for all non-DEN students (unless COVID conditions warrant another mode). It will be given per the university's official schedule for final exams: Thursday, 12/8/2022, 2:00-4:00 PM PST.

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 or final report instructions. 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 [2]

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 [4.5]

3. Probabilistic machine learning for regression [2]

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

4. Logistic regression [0.5]

   Cross-entropy error; maximum likelihood.

5. Graphical techniques for supervised learning [2]

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

Learning theory and its implications [7]

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 [2.5]

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

8. Practical applications [2]

   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]

Domain adaption and transfer learning [3.5]

10. Introduction and theory [1]

    Examples and realms;
    Cross-domain generalization error bounds.

11. Approaches and methods [2.5]

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

Semi-supervised learning [3.5]

12. Introduction and methods 1 [2]

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

13. Methods 2 [1.5]

    Co-training; graphical techniques; SS SVM.

Unsupervised learning [3]

14. Statistical techniques [1.5]

    Maximum likelihood, expectation maximization.

15. Nonstatistical techniques [1]

    Similarity measures; hierarchical and graphical techniques.

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

Human interpretability [2.5]

17. Human interpretability overview [2.5]

    Using interpretable models; taxonomy of methods;
    Intrinsic interpretable models; model agnostic methods; example-based explanations.

Course conclusions [1]

18. Course wrap-up and review for final exam [1]

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