USC Viterbi School of Engineering

# Data Science 352:
# Applied Machine Learning and Data Mining
# Spring 2024

**Units:** 4

**Instructor:** Fred Morstatter, Ph.D.
fred@isi.edu

**Office Hours:** One hour after class, PHE 514
By appointment (Zoom or ISI 930)

**TA:** Myrl Marmarelis
myrlm@isi.edu

**Lecture:** Monday/Wednesday, 4:00-5:50 PM DMC 101

**Webpages:** Piazza for everything except grades
USC Blackboard for grades
Vocareum for code submission

**Prerequisites:** DSCI 250 and MATH 208

**Other requirements:** Computer programming skills.
Students must know Python or must be willing to learn it.
All HWs, handouts, solutions will be posted in PDF.

**Tentative Grading:**

55% Programming Assignments (Labs)
25% Problem Sets
5% Topic Presentation
10% Midterm Exam
10% Final Exam

**Letter Grade Distribution:**

A ≥ 93
93 > A− ≥ 90
90 > B+ ≥ 87
87 > B ≥ 83
83 > B− ≥ 80
80 > C+ ≥ 77
77 > C ≥ 73
73 > C− ≥ 70
70 > D+ ≥ 67
67 > D ≥ 63
63 > D− ≥ 60
60 > F

**Disclaimer:** The instructional team 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). Often, if one student has a question/comment, others also have a similar question/comment. Private Piazza posts should be used to contact the professor, TA, graders only for issues that are specific to you individually (e.g., a scheduling issue or grade issue). Please use your USC email to register on Piazza and to contact the instructor and TAs.

**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 massive data 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: Wednesday, February 21, 4:00-5:50 PM.
- Final Exam: Last Class Meeting (April 24, 2024, 4:00-5:50 PM).

**Textbooks:**

- Required Textbooks:

1. Alpaydin, Ethem (2020). *Introduction to Machine Learning*. MIT Press, Fourth Edition. **[IML]**
2. Leskovec, Jure; Rajaraman, Anand; & Ullman, Jeff D. (2020). *Mining of Massive Data Sets*. Cambridge University Press, Third Edition. **[MMDS]**

- Recommended Textbooks:
  1. James, Gareth; Witten, Daniela; Hastie; Trevor, & Tibshirani; Robert (2021). *An Introduction to Statistical Learning with Applications in R*, Springer, Second Edition. **[ISLR]**

**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 grade in problem sets and your lowest grade in programming assignments (Labs) will be dropped from the final grade. Lab 0 will not be graded.

**Homework Policy:**

- Homework is assigned on an approximately weekly basis. A one-day grace period can be used for each homework with a 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 before 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 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.
- 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 recommending 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 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. Taking the exam, waiting for the grade, and then mentioning that you were sick is not acceptable.
- Midterm and final exams will be closed book, closed notes, and closed devices. 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 considerable emphasis on material presented since the last exam.

**Student Topic Presentation:**

- On the week of April 15, you will give a 5 minute presentation. In this presentation, you will identify a real-world example of one of the machine learning topics discussed in class being used to address a problem of your choice.
- Example topics include machine learning applied to: sports forecasting (e.g., sabermetrics), election prediction, video compression, anti-cheat methods in video games. These are just examples to get you thinking. Any topic is in play as long as it pertains to a machine learning concept discussed in class.
- Presentation structure:
    - What problem is being addressed? Why is this important? *(~2 minutes)*
    - How was this problem addressed? What class concept was used, and ***how***? *(~3 minutes)*
- Students will be required to create one slide to aid their presentation. At a minimum, the slide must include a pointer to the work that is being presented. The slide will be a Google Slides presentation, and will be due April 8.

**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:**

1. January 8
   Introduction to Machine Learning (IML Ch. 1)

2. January 15 – Martin Luther King, Jr. Day
   January 10, 17
   Supervised Learning (IML Ch. 2)

3. January 22
   Supervised Learning (cont'd)
   Bayesian Decision Theory (IML Ch. 3)

4. January 29
   Parametric Methods (IML Ch. 4)
   Multivariate Methods (IML Ch. 5)

5. February 5
   Clustering (IML Ch. 7)
   Nonparametric Methods (IML Ch. 8)

6. February 12
   Decision Trees (IML Ch. 9)
   Linear Discrimination (IML Ch. 10)

7. February 19 – Presidents Day
   **February 21 – Midterm Exam**

8. February 26
   Multilayer Perceptrons (IML Ch. 11)
   Deep Learning, pt. 1 (IML Ch. 12)

9. March 4
   Deep Learning, pt. 2 (IML Ch. 12)
   Local Models (IML Ch. 13)

10. March 11
    Spring Break (no class held this week)

11. March 18
    Kernel Machines (IML Ch. 14)

12. March 25
    Graphical Models (IML Ch. 15)
    Hidden Markov Models (IML Ch. 16)

13. April 1
    Data Mining (MMDS Ch. 1)
    Map-Reduce and the New Software Stack (MMDS Ch. 2)

| |
|---|
| 14. April 8<br>Design and Analysis of Machine Learning Experiments (Ch. 20) |
| 15. April 15<br>Student Topic Presentations |
| 16. April 22 – Final Review<br>**April 24 – Final Exam** |

**Deadlines for Assignments**

| **Deliverable** | **Deadline** |
|---|---|
| Lab 0 | January 27, 2024 |
| Problem Set 1 | February 3, 2024 |
| Lab 1 | February 10, 2024 |
| Problem Set 2 | February 17, 2024 |
| Lab 2 | February 24, 2024 |
| Problem Set 3 | March 2, 2024 |
| Lab 3 | March 9, 2024 |
| Problem Set 4 | ~~March 16, 2024~~ March 23, 2024 |
| Lab 4 | ~~March 23, 2024~~ March 30, 2024 |
| Problem Set 5 | ~~March 30, 2024~~ April 6, 2024 |
| Lab 5 | ~~April 6, 2024~~ April 13, 2024 |
| Problem Set 6 | ~~April 13, 2024~~ April 20, 2024 |
| Lab 6 | ~~April 20, 2024~~ April 27, 2024 |