Department of Civil and Environmental Engineering

University of Massachusetts Amherst

CEE 697M: Data Mining and Machine Learning for Engineers

Spring 2023 COURSE SYLLABUS

Contents

1 Personnel and Logistics

1.1 Venue and Class Time

Days & Times:
Lectures: W and F 8:30am–9:45am at E-Lab 307
Office Hours: M 11:00am–12:00pm; W 10:00am–11:00am at Marston 214D (or Zoom)

1.2 Instructor

Name: Jimi Oke[1]
Email: jboke@umass.edu[2]
Office: 214D Marston Hall

2 Course Information

2.1 Description

This course covers core concepts in machine learning (models and algorithms) from a **probabilistic perspective**. Key topics include:

- linear methods for regression and classification (including flexible functional forms)
- deep neural networks for structured data, sequences and images
- nonparametric methods: kernels, support vector machines, decision trees
- unsupervised learning (dimensionality reduction, clustering)

Applications to various subdisciplines will be highlighted, especially in transportation, environmental, structural and industrial engineering. Hands-on programming in Python (R will also be supported) throughout the course will enable students to analyze and train models on real-world datasets. Through this course, students will understand the potential of machine learning in civil, environmental and industrial engineering, among other disciplines, as well as learn to create and train models from data to solve challenging problems.

Given the current and future directions of this course, it will be permanently renamed "**Probabilistic Machine Learning [for Engineers]**" starting next spring. Advanced topics, such as VI, MCMC, BNNs, generative models (AR, GAN, etc), HMMs, SSMs, and more, will be covered in a 700-level future course titled "**Advanced Probabilistic Machine Learning [for Engineers]**."

2.2 Objectives

- Understand the theory behind fundamental ML models and algorithms and apply them to engineering problems
- Develop and train ML models for various problems in engineering and beyond
- Learn to use Python or similar programming language (e.g. R) to execute ML models

[1] Approximate pronunciation of last name in IPA: /ɔˈkɛ or "aw-KEH"
[2] Please allow up to 48 hours for a response to your email. Be sure to put "CEE697M" in the subject to ensure a prompt response.

2.3 Texts

The primary texts for this course are:

- Murphy, K. (2022). *Probabilistic Machine Learning: An Introduction.* MIT Press. (This text is freely available at https://probml.github.io/pml-book/book1.html. Abbreviated as **PMLI** in the schedule and lecture ntoes.)
- Goulet, J.-A. (2020) *Probabilistic Machine Learning for Civil Engineers*, MIT Press. (This text is freely available at http://profs.polymtl.ca/jagoulet/Site/Goulet_web_page_BOOK.html. Abbreviated as **PMLCE** in the schedule and lecture notes.)
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York, NY. Second Edition. (This text, among other resources, is also freely available from the authors at https://web.stanford.edu/~hastie/ElemStatLearn/. Abbreviated as **ESL** in the schedule and lecture notes.)

Supplementary texts:

- Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*, MIT Press. (This text is freely available at https://www.deeplearningbook.org/. Abbreviated as **DL** in the schedule and lecture notes.)
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* Springer, New York. (The text, among other resources is freely available from the author at https://www.statlearning.com/. Abbreviated as **ISL** in the schedule and lecture notes.)

Any other recommended or required reading will be provided on Moodle.

2.4 Prerequisites

College-level knowledge of probability, statistics, linear algebra and calculus. Some programming experience in any language is helpful, but you should be ready to get up to speed with any necessary technical skills. Familiarity with Python/R is encouraged.

3 Policies and Values

I will use slides in the classroom, and annotate them electronically when possible. These slides will be available to you prior to the lecture. I will endeavor to foster an equitable and inclusive learning environment that will spark your curiosity and challenge you learn actively. I strongly urge you to come to class prepared, having done the reading, ready to reflect on your homework or problem set and engage with new material. I will ask frequent questions of you, and will also expect you to ask as many questions as possible. Further specifics on class policies and values are as follows.

3.1 Assessments

There will be no grading on a curve. Consistent with this, after drop date, students who remain in this class are not in jeopardy of seeing their grades change due to the change in class composition. The breakdown is provided in Table 1.

Table 1 Course components and grade breakdown

Assessment	**Value (%)**
Problem Sets	50
Midterm Exam 1	15
Midterm Exam 2	15
Project	20

3.2 Problem sets

Five problem sets will be assigned. Submission will be online (PDFs and other supporting code; or Jupyter notebooks) via Moodle. Each will be worth 10% of your total grade. *Late problem sets will automatically attract a 25% penalty and will not be accepted more than 4 days beyond the due date* (excepting prior permission).

3.3 Midterms

There will be 2 take-home midterms, which will be open-resource. Previous exams may be available for practice.

3.4 Programming

Some lectures will incorporate engineering applications of machine learning concepts using Python. Problem sets will also involve some coding in Python. I recommend installing JupyterLab. You are welcome to use other languages/platforms such as R/RStudio or Matlab for your assignments. However, I cannot guarantee the same level of support for Matlab in particular.

3.5 Computing resource

Having a laptop is not a requirement for this course. However, if you own one and are able to bring it to the classroom, it may improve your learning experience during the programming segments of the lecture.

3.6 Project

The term project will be worth 20% of your total grade. You are encouraged to start thinking about the concepts and methods you would like to investigate further in a real-world setting. I will ask you to submit a project proposal (individually or with a partner or two of your choice) that applies two of the modeling approaches covered in class to a relevant problem. This may be related to your own research as well. Further guidance will be provided midway through the semester. The final exam time will be devoted to in-class presentations of each project.

3.7 Attendance and participation

You are expected to show up to every class (either virtually or in-person), in the absence of any emergencies or illness (please email me ahead of time if any situations arise).

3.8 Academic Honesty Policy Statement

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst. Academic dishonesty including but not limited to cheating, fabrication, plagiarism, and facilitating dishonesty, is prohibited in all programs of the University. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. The procedures outlined below are intended to provide an efficient and orderly process by which action may be taken if it appears that academic dishonesty has occurred and by which students may appeal such actions. Since student are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent. For more information about what constitutes academic dishonesty, please see the Dean of Students' website: `https://www.umass.edu/honesty/`

3.9 Disability Statement

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier-free campus. If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. It is your responsibility to initiate these services and to communicate with faculty ahead of time to manage accommodations in a timely manner. For more information, consult the Disability Services website at `http://www.umass.edu/disability/`.

4 Schedule

This course is broadly organized around 5 modules. The schedule (see Table 2 on the next page) may be adapted over the duration of the semester to suit the needs of the class. Readings will be provided in lecture notes and on Moodle.

4.1 Important dates

- No class on the following dates:
  - Wednesday, February 22 (Monday schedule followed)
  - Wednesday, March 15 and Friday, March 17 (Spring Break)
  - Wednesday, March 29 (Exam I)
  - Wednesday, May 17 (Exam II)
- Last day of class: Wednesday, May 12
- Project presentations: Wednesday, May 24 (8:00–11:00am)

Table 2 Course schedule

Day	Date	L/N	Topic	Assignments
Module 1: Foundations				
W	Feb 8	1a	Introduction	
F	Feb 10	1b	Probability	PS1 assigned
W	Feb 15	1c	Statistics	
F	Feb 17	1d	Decision theory; Information theory	
F	Feb 24	1e	Linear Algebra	
W	Mar 1	1f	Optimization	
Module 2: Linear Methods				
F	Mar 3	2a	Linear discriminant analysis	PS1 due; PS2 assigned
W	Mar 8	2b	Logistic regression	
F	Mar 10	2c	Linear regression (OLS, WLS)	
W	Mar 22	2d	Ridge and Lasso regression	
F	Mar 24	2e	Splines and generalized additive models (GAMs)	
W	Mar 29	2f	Generalized linear models (GLMs)	
F	Mar 31	E1	Exam I (take-home; no class)	PS2 due
Module 3: Deep Neural Networks (DNNs)				
W	Apr 5	3a	NNs for structured data I (MLP, backpropagation)	PS3 assigned
F	Apr 7	3b	NNs for structured data II (training, regularization)	
W	Apr 12	3c	NNs for images (CNNs)	
F	Apr 14	3d	NNs for sequences (RNNs)	
Module 4: Nonparametric Methods				
W	Apr 19	4a	Exemplar-based methods (KNN, KDE, LOESS)	PS3 due; PS4 assigned
F	Apr 21	4b	Gaussian processes	Project proposal assigned
W	Apr 26	4c	Support vector machines	
F	Apr 28	5d	Trees and ensemble methods	
Module 5: Unsupervised Learning				
W	May 3	5a	Principal components analysis (PCA)	PS4 due; PS5 assigned
F	May 5	5b	Factor analysis (FA)	Proposal due
W	May 10	5c	Autoencoders (AEs, VAEs)	
F	May 12	5d	Clustering (HAC, KMeans, MM)	
W	May 17	E2	Exam II (take-home; no class)	PS5 due
W	May 24		Project presentations	