ECE 471: Data Science Analytics using Probabilistic Graph Models

Undergrad – 3 hours – CRN: 77531 | Grad – 4 hours – CRN: 77533

Course Instructor

Ravishankar K Iyer, Professor, Electrical and Computer Engineering
255 Coordinated Science Lab MC 228
Class website: https://courses.grainger.illinois.edu/ece471/fa2024/
Office Hours: 4:00-5:00 p.m. Wednesday. In person or by Zoom.

About the Class

This class will introduce students to real-world problems of societal importance (e.g., Safety of AVs, Health Analytics, Large-scale Infrastructures) and their solutions (using e.g., Bayesian networks, Factor Graphs, and large language models (LLMs)). The course will use real data from several domains (including resiliency and healthcare science) to instill in the students following data-science and AI/ML expertise:

- Data management
- Feature engineering by identifying key data characteristics
- Problem formulation, machine learning models and validation
- Design, construction, and assessment of end-to-end application workflows by usingdata-driven insights
- Generating application domain insights that can improve the understanding of thedomain problem

Through this course, students will be able to develop expertise and the ability to form intuitions which can be more broadly applicable. Such an expertise will prepare students for data science and ML engineering roles.

Course Description

Many modern application domains require engineers and domain experts to collaborate in designing and analyzing heterogeneous datasets, often with the objective of automating decision-making (often referred to as actionable intelligence). Combining the right data with science/domain knowledge and Machine Learning (ML) to generate actionable intelligence from these datasets remains a compelling problem.

The course addresses this problem by allowing students to build analysis workflows that use advanced data measurement and management, combined with a variety of ML methods, ranging from Bayesian networks to large language models (LLMs) and their underlying principles. Students will work on real-world applications while interacting with invited domain experts.

This course instills the skillsets required for constructing end-to-end real-world analysis workflows through lectures, in-class group activities, homework, and mini-projects, which will allow students to derive domain insights.

The course will use real-world examples (measurement logs from supercomputers, data on autonomous vehicle safety, and clinical trials) together with ML methods ranging from Bayesian Networks to LLMs and their underlying principles. Students will gain hands-on implementation experience by completing two mini-projects requiring them to analyze high-fidelity real-world measurement data obtained from different application domains. While each workflow is end-to-end, the students will develop a deeper understanding of the methods as they progress through these projects. Additionally, students will learn to create quantifiable domain-specific metrics of interest (cost models) and use techniques to quantitatively assess their results.

Students will develop the expertise and an ability to form intuitions which are applicable in other domains.

This course will feature guest lectures from domain experts who will demonstrate how innovative analytic techniques have been transformative and, as a result, generated significant societal impact.

Recommended Text:

- – Class Notes and Lecture Slides
- – Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"

- Further Reading and Sample Problems:
  - – Daphne Koller and Nir Freidman, "Probabilistic Graphical Models: Principles and Techniques"
  - – Ravi Iyer's ECE 313 Class Notes (link)

Course Structure, Objectives and Learning Outcomes

Lectures
You will be introduced to real-world problems of societal importance (e.g., Safety of AVs, Health Analytics, Large-scale Infrastructures) and their solutions (using e.g., Bayesian networks, Factor Graphs, Large Language Models (LLMs))

Mini projects
There will be two mini-projects (for undergrads) and three mini projects for 4 credit hour students to provide hands-on experience with applications of data analytics/machine learning.

In-class activities
Students in small groups will work on realistic design and assessment problems including hands-on modeling with support from instructors and TA's. There will be 8 in-class activities expected to be completed by the group during class. Attendance is mandatory.

Homework assignments
You will be given theoretical questions, and small programming assignments to strengthen your understanding of the methods learned in lectures.

Course Outcomes.
At the end of the course, students will have learned:

- to handle complex high-dimensional data;
- how and when to apply ML-based solutions such as Bayesian Models (Hidden Markov Models/Factor Graphs) building up to LLMs (Role of Transformers, neural networks, reinforcement learning (RL) and finetuning); and
- how to derive insights by combining model solutions with domain knowledge.

Prerequisites
Basic probability and basic programming skills are essential. ECE 313 or CS 361 (Statistics and Probability), and exposure to Python.HW0 to test basic probability skills. Talk to the instructor if you find HW0 to be difficult.

Timeline
There are 45 hours lectures {30 hours classroom lectures, quizzes and presentations & 15 hours hands-on data analytics mini projects}, over 15 weeks in the fall semester.

Lecture Electronics Policy
During the lectures, and in class activities, cell phones or similar non-class use of electronics are NOT allowed. If, due to unforeseen circumstances, the student needs access to her/his cell phone, she/he shall inform the instructor in the beginning of the lecture and should sit in a way (typically furthest from the board) not to allow any students around them to be disturbed.

Attendance Policy
Attendance to all lectures, and in-class/group activities are required (15% of total grade). There will be in-class assignments and class participation is graded. Students are advised to contact both the TAs andthe instructor via a private post on Piazza (before the beginning of the lecture) if they are to miss a lecture due to unforeseen circumstances. Instructor and TAs reserve the right to take class attendance. Class attendance includes data analytics lab hours. Students can miss no morethan one group activity. Note that group activities are only tentatively scheduled.

DRES Accommodations

DRES requirements must be reported to instructor/TAs by the end of 1st week.

Evaluation
We will compute the final grade for undergraduates and graduate students based on this allocation:

Activity	Undergraduate Grade	Graduate Grade (4 credit hour students)
Mini-Projects 1, 2	30% (MP1 12%, MP2 18%)	30 points (MP1-12, MP2-18)
Midterm and Final	30%	30 points
Mini-Project 3 (Grads only)		25 points
In-class (design and assessment) activities	15%	15 points
Class Participation (includes office hours, discussions, and short end-of-lecture quizzes)	15%	15 points
Homework	10%	10 points
TOTAL	**100%**	**125 points prorated to 100%**

- There will be two mini-projects (for undergrads) and three mini projects for 4 credit hour students only to provide hands-on experience with applications of data analytics/machine learning
- Mini-projects will be handed out or posted on the class website
  - Full credit for submissions on time.
  - Late submission policy: 10% will be taken off for every day, prorated (up to 4 days max). 0 credit after that.
  - Groups Policy: Students will form groups (3 persons) for the project starting in week 1, otherwise TAs will form the groups for you
- While I encourage discussions, submitting identical material is not allowed and will incur appropriate penalties
- We will hold additional office hours/discussion sessions related to the mini-projects in progress. You are strongly encouraged to attend. Location: TBA 4pm-5pm on Fridays. No discussion in the first week.
- Project descriptions and due dates will be posted on the class website under Student Projects.
  - There will be a resources and tutorials sections for hints related to projects

Mini-Projects
Students work in groups and follow detailed instructions to build end-to-end workflows to solve real-world data science problems. Two mini-projects (three for graduate students) in high social impact domains ranging from autonomous vehicle safety to health analytics.

Graduate Project (one additional credit requirement for the graduate students; approximately 20% of the overall graduate credit). Graduate students work, in close collaboration with the instructor, on a semester long data- science team project of their choice. The project covers various aspects such as problem formulation, finding a dataset, pre-processing of the data and preliminary analysis, model building and validation and interpretation of results.

Schedule

Week	Date	Title	HW	MP
1	Aug 27	**Course outline** 1. Overview of key data analytics and ML concepts 2. Overview of Mini Projects • Autonomous Vehicle (AV) Safety Analytics • LLM interpretability and assessment in practice		
	Aug 29	**Probability Basics Overview**; Probability and Hypothesis Testing, p-value; fitting distributions (KS test, KL divergence); Introducing **Mini-project 1;** Demonstrating AV Simulator with Carla	HW0 release	
	Aug 30 (Friday)	Discussion session/open office hour on mini project 1 (optional discussion session, attendance is encouraged). 4 p.m./location TBD		
2	Sep 3	**In-Class Activity 1** (Probability concepts, hypothesis testing, jupyter notebook)		
	Sep 5	**Overview of unsupervised clustering algorithms**: K means, Gaussian Mixture Model (GMM), Expectation Maximization (EM) and similar methods.		
3	Sep 10	**Regression and dimensionality reduction techniques**: Distance Metrics, Principal Component Analysis (PCA), Factor analysis		
	Sep 12	**Real-world application of dimensionality reduction**		
4	Sep 17	**In Class Activity 2** on dimensionality reduction	HW1 release	
	Sep 19	**Probabilistic Graph Models (PGMs) and applications;** conditional independence, Naïve Bayes,		
5	Sep 24	**Bayesian Networks**		
	Sep 26	Bayesian networks/PGM's continue		
	Sep 27 (Friday)	Discussion/Open Office Hour on Bayesian Network (optional discussion session, attendance is encouraged) 4 p.m./location TBD		
6	Oct 1	Application of Bayesian Networks/PGMs to Health-care; **In Class Activity 3.**	Grad project draft proposal due	
	Oct 3	**Markov Models**: Data driven methods for building Markov Models for large-scale computer system (NCSA's Blue Waters) addressing performance and reliability; real example with data		
7	Oct 8	**Hidden Markov Models (HMM)**		
	Oct 10	**In-Class Activity 4:** Midterm review, sample problem-solving		
	Oct 11	Discussion session/open office hour on Midterm and Markov Models (optional discussion session, attendance is encouraged) 4 p.m./location TBD		

8	Oct 15	**MIDTERM EXAM – 60 minutes**		
	Oct 17	**In Class Activity 5** on HMM		
9	Oct 22	**Generalization of PGMs -Factor Graphs (FG)**		
	Oct 24	**Introducing Belief Propagation and its Applications in FGs**; Approximate solution methods		
	Oct 25 (Friday)	Discussion session/Open office hour on belief Propagation. (optional discussion session, attendance is encouraged) 4 p.m./location TBD		
10	Oct 29	**Approximate solution methods Cont'd** (Markov Chain Monte Carlo and Gibbs Sampling)		
	Oct 31	**In Class Activity 6** on factor graphs		
11	Nov 5	**Introduction to emergent LLMs** role in new generation of inference models		
	Nov 7	**Role of Neural Networks, Reinforcement Learning, and Reward Models in training LLMs**		
12	Nov 12	LLM Architecture and Training: Role of Transformers and RL, **Mini Project 2 on LLMs Intro**		
	Nov 14	**Addressing Hallucinations:** Retrieval Augmented Generation		
	Nov 15 (Friday)	Discussion session/Open Office Hour on mini project 2 and introduction to LLMs (optional discussion session, attendance is encouraged). 4 p.m./location TBD		
13	Nov 19	**In Class Activity 7 on LLMs**		
	Nov 21	LLM Model Finetuning: LoRA and Synthetic Data Generation		
14	Nov 26	***No classes – Fall break***		
	Nov 28	***No classes – Fall break***		
15	Dec 3	**Agentic Frameworks with LLMs**		
	Dec 5	**Modeling Uncertainties in LLM Training and Inference**		
	Dec 6 (Friday)	Discussion session/Open Office Hour on training of and inferring using LLMs and other problems. (optional discussion session, attendance is encouraged). 4 p.m./location TBD		
16	Dec 10	**Review and Practice Final Exam**		
	Dec 12	Reading Day (no classes)	Grad project final submission	
	TBD	**FINAL EXAM**		