CS37300 Data Mining and Machine Learning

Section 1

Instructor: Steve Hanneke, Assistant Professor of Computer Science

Email: hanneke@purdue.edu

Lecture: Monday, Wednesday, and Friday 10:30-11:20am @ KRAN 140

Office Hours: Thursdays 9:00-10:00am and 6:00-7:00pm @ DSAI 1100

Section 2

Instructor: Tianyi Zhang, Assistant Professor of Computer Science

Email: tianyi@purdue.edu

Lecture: Monday, Wednesday, and Friday 12:30-1:20pm @ KRAN 140

Office Hours: Monday and Wednesday 1:30-2:30pm @ LWSN 3154H

Teaching Assistants:
Prerit Gupta gupta596@purdue.edu
Office Hours: Thurs 4:00 - 5:00 pm @ DSAI B063

Bonan Kou koub@purdue.edu
Office Hours: Thurs 9:30-10:30 am @ DSAI B055

Wei-Hao Chen chen4129@purdue.edu
Office Hours: Fri 11:30-12:30 pm @ DSAI B063

Daniel P. de Mello ddemello@purdue.edu
Office Hours: Wed 1:00 - 2:00 pm @ HAAS G072

Zachary J. Lee lee5230@purdue.edu
Office Hours: Tues 4:00 - 5:00 pm @ DSAI B063

Nikolaos Papagiannis npapagia@purdue.edu
Office Hours: Mon 2:30 - 3:30 pm @ DSAI B061

Amit Roy roy206@purdue.edu
Office Hours: Wed 2:00 - 3:00 pm @ DSAI B061

Juexiao Wang wang5360@purdue.edu
Office Hours: Fri 3:30 - 4:30 pm @DSAI B061

Instructional Modality: Face-to-Face

Course Credits: 3.0

Prerequisites: The formal prerequisites are CS 18200: Foundations of Computer Science, and CS 25100: Data Structures and Algorithms. You also must have either taken or be taking STAT 35000: Introduction to Statistics, or STAT 51100: Statistical Methods. If you have comparable courses, such as ECE 36800, please contact the instructor.

Course Description

This course will introduce students to the field of data mining and machine learning, which sits at the interface between statistics and computer science. Data mining and machine learning focuses on developing algorithms to automatically discover patterns and learn models of large datasets. This course introduces students to the process and main techniques in data mining and machine learning, including exploratory data analysis, predictive modeling, descriptive modeling, and evaluation.

The course will primarily be taught through lectures, supplemented with reading. The written theory and programming assignments are also a significant component of the learning experience. We will be using Gradescope to turn in and comment on assignments.

Brightspace will be used for recording and distributing grades, as well as for any other non-public information about the course. Lecture slides will also be posted on Brightspace.

We will use Ed Discussion to make announcements, send reminders, ask & answer questions, etc. Please join our course on Ed Discussion via this link. You are encouraged to make your question visible to everyone, since other students may have similar questions. You are also encouraged to hold discussions with other students. Please keep the collaboration guidelines in the Intellectual Honesty Policy in mind. Purdue has paid licenses for WebEx and Zoom, if you wish to meet remotely with other students.

Course Registration Policy

This course is anticipated to be oversubscribed, and as such registration is initially limited to CS students. If you have been unable to register, please see the CS department Course Access & Request Policy. Please do not ask the instructor for an override, we have been told that if the course is shown as full, the registrar will not allow registration even with a form 23 signed by the instructor, so you would just be wasting your time and ours. Please follow the process above or consult with your advisor.

Learning Outcomes

After successful completion of this course, a student will be able to:

- Identify key elements of data mining and machine learning algorithms.
- Understand how algorithmic elements interact to impact performance.
- Understand how to choose algorithms for different analysis tasks.

- Analyze data in both an exploratory and targeted manner.
- Implement and apply basic algorithms for supervised and unsupervised learning.
- Accurately evaluate the performance of algorithms, as well as formulate and test hypotheses.

Textbooks, Learning Resources, and Technology

The texts below are recommended but not required. Reading materials will be distributed as necessary through Ed Discussion. Please check regularly.

- D. Hand, H. Mannila, P. Smyth (2001). *Principles of Data Mining*. MIT Press. ISBN 026208290X.
  This book is available for free within the Purdue domain.
- Christopher M. Bishop (2006), *Pattern Recognition and Machine Learning*.
  This is a very detailed and thorough book on the foundations of machine learning.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (2018), *Foundations of Machine Learning*.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press.
  This is a detailed e-book on deep learning.
- Christopher M. Bishop, and Hugh Bishop (2024). *Deep Learning: Foundations and Concepts*. Springer.
  This is a more up-to-date textbook on deep learning. It is available for free within the Purdue domain.

The following are also useful:

- Kevin P. Murphy (2022), *Probabilistic Machine Learning: An Introduction*.
- Hal Daume III, A Course in Machine Learning.
  This is a good book with important practical guidelines.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*.
  This is an excellent reference book.

Grading

Midterm Exam: 20%

Final Exam: 30%

Theory and programming assignments: 45% (5% for HW0 and 8% for HW1-5)

Take-home Quizzes: 5%

This course follows the generic letter-based grading scheme on Purdue Brightspace. Depending on the performance of most students in this course, the instructor will decide whether to curve or

not. If a student's final percentage grade is within 0.5% to reach a higher letter grade, the student's participation in the class and office hours will be taken into consideration to decide whether to assign a higher grade or not.

Percentage Grade (%)	Letter Grade
[97, 100]	A+
[93, 97)	A
[90, 93)	A-
[87, 90)	B+
[83, 87)	B
[80, 83)	B-
[77, 80)	C+
[73, 77)	C
[70, 73)	C-
[67, 70)	D+
[63, 67)	D
[60, 63)	D-
[0, 60)	F

Course Schedule

	Lecture Topic	Reading	Homework
Jan 13 (M)	Course Overview and Introduction	Principles of Data Mining, Chapter 1 Pattern Recognition and Machine Learning, Chapter 1	
Jan 15 (W)	Linear Algebra Review	Deep Learning, Chapter 2	

Jan 17 (F)	Python Basics Tutorial by TAs		Release HW0
Jan 20 (M)	**MLK Day (No Class)**		
Jan 22 (W)	Probability basics: Independence, expectation/variance, distributions, MLE	Pattern Recognition and Machine Learning, Chapter 1.2, 2 Principles of Data Mining, Chapter 4.1-4.5	
Jan 24 (F)			
Jan 27 (M)	Nearest Neighbors	Principles of Data Mining, Chapter 10.6	
Jan 29 (W)	Decision Trees	Principles of Data Mining, Chapter 10.5	HW0 Due HW1 Release
Jan 31 (F)	Decision Trees		
Feb 3 (M)	Naive Bayes	Principles of Data Mining, Chapter 10.8	
Feb 5 (W)	Linear Classifiers, Perceptron	Pattern Recognition and Machine Learning, Chapter 4 Principles of Data Mining, Chapter 10.3	
Feb 7 (F)	Hard-margin SVM	Pattern Recognition and Machine Learning, Chapter 7	
Feb 10 (M)	Soft-SVM / Logistic Regression		HW1 Theory Due
Feb 12 (W)	Kernel Methods	Pattern Recognition and Machine Learning, Chapter 6	
Feb 14 (F)	Gradient Descent	Principles of Data Mining, Chapter 8.3	HW1 Programming Due HW2 Release
Feb 17 (M)	Linear Models for Regression	Principles of Data Mining, Chapter 11.3	

Feb 19 (W)	Gaussian Processes	Pattern Recognition and Machine Learning, Chapter 6.4	
Feb 21 (F)	Sequential Data Modeling	Pattern Recognition and Machine Learning, Chapter 13	
Feb 24 (M)	Ensemble Methods: Bagging, Boosting	Pattern Recognition and Machine Learning, Chapters 14.2 and 14.3	
Feb 26 (W)	Evaluation for Supervised Learning	Principles of Data Mining, Chapter 7.4 Principles of Data Mining, Chapter 10.10	HW2 Theory Due
Feb 28 (F)	**Overfitting and Bias/Variance**		
Mar 3 (M)	Midterm Review Session		HW2 Programming Due HW3 Release
Mar 5 (W)	Midterm		
Mar 7 (F)	Learning Theory I	Foundation of Machine Learning, Page 19 and Chapter 3.3	
Mar 10 (M)	Learning Theory II		
Mar 12 (W)	Introduction to NN	Pattern Recognition and Machine Learning, Chapter 5	
Mar 14 (F)	NN Backpropagation		HW3 Theory Due
Mar 17-22	**Spring Break (No Class)**		
Mar 24 (M)	Neural Network Architectures	Deep Learning, Chapters 9, 10	HW3 Programming Due HW4 Release

Mar 26 (W)	Neural Network Architectures	Transformers for Machine Learning: A Deep Dive, Chapter 2	
Mar 28 (F)	NN Training, Inference, Regularization, Optimization	Deep Learning, Chapters 7, 8	
Mar 31 (M)	Pattern Mining, Association Rules	Principles of Data Mining, Chapter 13	
Apr 2 (W)	Partition-based Clustering	Principles of Data Mining, Chapter 9.4	
Apr 4 (F)	Mixture Models	Principles of Data Mining, Chapter 9.2, 9.6	HW4 Theory Due
Apr 7 (M)	Expectation Maximization	Pattern Recognition and Machine Learning, Chapters 9.2.2, 9.3	
Apr 9 (W)	Hierarchical Clustering	Principles of Data Mining, Chapter 9.5	
Apr 11 (F)	Principal Component Analysis	Principles of Data Mining, Chapter 3.6	HW4 Programming Due HW5 Release
Apr 14 (M)	Density-based Clustering Methods	DBSCAN (KDD'96) OPTICS (SIGMOD'99) DENCLUE (KDD'98) CLIQUE (SIGMOD'98)	
Apr 16 (W)	Evaluation for Unsupervised Learning	Introduction to Data Mining, Chapter 8.5	
Apr 18 (F)	Semi-supervised Learning	Semi-supervised learning literature survey	
Apr 21 (M)	Large Language Models	A survey of large language models	HW5 Theory Due

Apr 23 (W)	Generative Modeling: Variational Autoencoders	Deep Learning: Foundations and Concepts, Chapter 19.2	
Apr 25 (F)	Generative Modeling: Score-based/Diffusion models	Deep Learning: Foundations and Concepts, Chapter 20	HW5 Programming Due
Apr 28 (M)	Reliable ML: Model Uncertainty and Robustness	Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods	
Apr 30 (W)	ML Interpretability, Privacy, and Fairness	Interpretable Machine Learning, Christoph Molnar When Machine Learning Meets Privacy: A Survey and Outlook A Survey on Bias and Fairness in Machine Learning	
Mar 2 (F)	Final Exam Review		

Theory and Programming Assignments

This course includes six homework assignments, which takes 45% of your final grade. Each homework (except HW0) has theory questions and programming questions. To help students manage the workload, we set two deadlines for each homework---one for the theory questions and one for the programming questions. Please check the release date and due date for each homework in the course schedule above.

Homework 0:

Programming:

- Python Basics using numpy, pandas and matplotlib
- Loading & analyzing dataset

Homework 1:

Theory:

- Probability Review, Bayes Theorem, Probability Distribution
- KNN & Decision Trees

Programming:

- KNN & Decision Trees

Homework 2:

Theory:
- Naïve Bayes Classifier
- Perceptron Algorithm
- Logistic Regression

Programming:
- Naïve Bayes Classification
- Logistic Regression

Homework 3:

Theory:
- Bias-Variance tradeoff & Cross-validation
- Ensemble Methods
- Support Vector Machines

Programming:
- Cross-validation
- Bagging & Boosting

Homework 4:

Theory:
- Multi-Layered Perceptron
- Activation Functions
- Convolutional Neural Network

Programming:
- Neural Networks
- Training & Testing Analysis

Homework 5:

Theory:
- Clustering (K-Means, Hierarchical)
- Mixture Models

Programming:
- PCA Implementation & Evaluation

Exams

This course includes a midterm exam and a final exam. The midterm takes 20% of your final grade and the final takes 30% of your final grade.

The midterm exam is scheduled to take place during the lecture time on Oct 13, Friday. For Section 1, it will be 9:30-10:20am at MATH 175. For Section 2, it will be 12:30-1:20pm @ WTHR 172. The midterm covers the content from Aug 21 to Oct 6. We will release practice questions and have a review session for the midterm on Oct 11. Since the midterm for Section 2 is two hours after the midterm for Section 1, we will design different exam questions with the same level of difficulty to avoid potential question leakage. Furthermore, students from Section 1 are not allowed to discuss the exam questions with students from Section 2 until the midterm of Section 2 is concluded. The violation of this requirement is considered as academic dishonesty.

We will have one single final exam for both sections at the same time and classroom. The time and location of the final exam will be announced in the middle of the semester. The final covers all content taught in this course. We will release practice questions and have a review session for the final on Dec 8.

Take-Home Quizzes

We will have five take-home quizzes, which takes 5% of your final grade. These quizzes are designed to be lightweight, including only a couple of questions based on some important and/difficult concepts from the lectures. The main purpose of these quizzes is to strengthen students' understanding about these concepts. Each quiz is announced at the end of a lecture and students need to submit their solutions on Gradescope before the next lecture. In the next lecture, the instructors will go over the solutions and have in-class discussions to review the concepts covered by the quiz.

Policies

Attendance

This course follows Purdue's academic regulations regarding attendance, which states that students are expected to be present for every meeting of the classes in which they are enrolled. While we will not check attendance in each class, we will use other ways such as class discussion to check your attendance and participation in the class. Please come to the class continuously and participate in discussions.

If you feel sick, have any symptoms associated with COVID-19, or suspect you have been exposed to the virus, you should stay home and contact the Protect Purdue Health Center. Please also notify the instructor so that the instructor can arrange remote participation for you. If you miss classes because of COVID-related reasons, your final grade will not be affected by your absence of classes. For more guidance on class attendance related to COVID-19 are outlined in the Protect Purdue Pledge for Fall 2021 on the Protect Purdue website.

For other conflicts or absences, such as for many University-sponsored activities and religious observations, the student should inform the instructor of the situation as far in advance as possible. When the student is unable to make direct contact with the instructor and is unable to leave word with the instructor's department because of circumstances beyond the student's control, and in cases falling under excused absence regulations, the student or the student's representative should contact or go to the Office of the Dean of Students website to complete appropriate forms for instructor notification. Under academic regulations, excused absences may be granted for cases of grief/bereavement, military service, jury duty, and parenting leave. For details, see the Academic Regulations & Student Conduct section of the University Catalog website.

Missed or Late Work

Late work will not be accepted, except as follows. You are allowed five extension days, to be used at your discretion throughout the semester (illness, job interviews, etc.) You may use at most two days on each assignment (so that we can get solution sets out in a timely manner.) Fractional use is not allowed -- each late day is a 24-hour extension. Late submission will not be accepted more than 48 hours after the due date, or after you have used your late days.

Intellectual Honesty

Please read the departmental academic integrity policy. This will be followed unless we provide written documentation of exceptions. You should also be familiar with the Purdue University Code of Honor and Academic Integrity Guide for Students. You may also find Professor Spafford's course policy useful - while we do not apply it verbatim, it contains detail and some good examples that may help to clarify the policies above and those mentioned below.

In particular, we encourage interaction: you should feel free to discuss the course with other students. However, unless otherwise noted work turned in should reflect your own efforts and knowledge.

For example, if you are discussing an assignment with another student, and you feel you know the material better than the other student, think of yourself as a teacher. Your goal is to make sure that after your discussion, the student is capable of doing similar work independently; their turned-in assignment should reflect this capability. If you need to work through details, try to work on a related, but different, problem.

If you feel you may have overstepped these bounds, or are not sure, please come talk to us and/or note on what you turn in that it represents collaborative effort (the same holds for information obtained from other sources that provided substantial portions of the solution.) If we feel you have gone beyond acceptable limits, we will let you know, and if necessary we will find an alternative way of ensuring you know the material. Help you receive in such a borderline case, if cited and not part of a pattern of egregious behavior, is not in our opinion academic dishonesty, and will at most result in a requirement that you demonstrate your knowledge in some alternate manner.

Use of Copyrighted Materials

Students are expected, within the context of the Regulations Governing Student Conduct and other applicable University policies, to act responsibly and ethically by applying the appropriate exception under the Copyright Act to the use of copyrighted works in their activities and studies. The University does not assume legal responsibility for violations of copyright law by students who are not employees of the University.

A Copyrightable Work created by any person subject to this policy primarily to express and preserve scholarship as evidence of academic advancement or academic accomplishment. Such works may include, but are not limited to, scholarly publications, journal articles, research bulletins, monographs, books, plays, poems, musical compositions and other works of artistic imagination, and works of students created in the course of their education, such as exams, projects, theses or dissertations, papers and articles.

Grief Absence Policy for Students

Purdue University recognizes that a time of bereavement is very difficult for a student. The University therefore provides the following rights to students facing the loss of a family member through the Grief Absence Policy for Students (GAPS). GAPS Policy: Students will be excused for funeral leave and given the opportunity to earn equivalent credit and to demonstrate evidence of meeting the learning outcomes for misses assignments or assessments in the event of the death of a member of the student's family.

Violent Behavior Policy

Purdue University is committed to providing a safe and secure campus environment for members of the university community. Purdue strives to create an educational environment for students and a work environment for employees that promote educational and career goals. Violent Behavior impedes such goals. Therefore, Violent Behavior is prohibited in or on any University Facility or while participating in any university activity.

Emergencies

In the event of a major campus emergency, course requirements, deadlines and grading percentages are subject to changes that may be necessitated by a revised semester calendar or other circumstances beyond the instructor's control. Relevant changes to this course will be posted onto the course website or can be obtained by contacting the instructors or TAs via email or phone. You are expected to read your @purdue.edu email on a frequent basis.

Accessibility and Accommodations

Purdue University strives to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability, you are welcome to let me know so that we can discuss options. You are also encouraged to contact the Disability Resource Center at: drc@purdue.edu or by phone: 765-494-1247.

Nondiscrimination

Purdue University is committed to maintaining a community which recognizes and values the inherent worth and dignity of every person; fosters tolerance, sensitivity, understanding, and mutual respect among its members; and encourages each individual to strive to reach his or her own potential. In pursuit of its goal of academic excellence, the University seeks to develop and nurture diversity. The University believes that diversity among its many members strengthens the institution, stimulates creativity, promotes the exchange of ideas, and enriches campus life.

Purdue University prohibits discrimination against any member of the University community on the basis of race, religion, color, sex, age, national origin or ancestry, genetic information, marital status, parental status, sexual orientation, gender identity and expression, disability, or status as a veteran. The University will conduct its programs, services and activities consistent with applicable federal, state and local laws, regulations and orders and in conformance with the

procedures and limitations as set forth in Executive Memorandum No. D-1, which provides specific contractual rights and remedies. Any student who believes they have been discriminated against may visit University's website (www.purdue.edu/report-hate) to submit a complaint to the Office of Institutional Equity. Information may be reported anonymously.