UCLA Samueli
Computer Science

CS145: Introduction to Data Mining (Spring 2024)

Discussion 0: **Overview & Logistics**

Instructor: Dr. Ziniu Hu
Teaching Assistant: Yanqiao Zhu

The Scalable Analytics Institute (ScAI)
Department of Computer Science
University of California, Los Angeles (UCLA)

Logistics

- Course homepage:
  - https://bruinlearn.ucla.edu/courses/183561
  - https://piazza.com/ucla/spring2024/cs145/home

- Lectures:
  - Monday & Wednesday 12:00PM – 1:50PM
  - Boelter Hall 3400

The team

- Instructor: Dr. Ziniu Hu (acgbull@gmail.com)
  - Postdoc researcher at Caltech; CS PhD at UCLA (Class 2023)
  - Research interests: Neural-Symbolic AI, Large Language Models
  - Homepage: https://acbull.github.io

- Teaching assistant: Yanqiao Zhu (yzhu@cs.ucla.edu)
  - Second-year CS PhD student at UCLA
  - Research interests: Graph and Geometric Deep Learning, AI for Science
  - Homepage: https://web.cs.ucla.edu/~yzhu

Evaluation

- Evaluation scheme:
  - Assignments: 40%
    - Participation: +5% bonus
  - Project: 30%
    - Performance: +5% bonus
  - Exam: 30%

- Letter grade scheme:

Grade	Point Range
A+	[97, +∞)
A	[93, 97)
A-	[90, 93)
B+	[87, 90)
B	[83, 87)
B-	[80, 83)
C+	[77, 80)
C	[73, 77)
C-	[70, 73)
D+	[67, 70)
D	[60, 67)
F	[0, 60)

Assignments (40%)

- There will be a total of five weekly assignments; each will be given 10 days to finish
- Grading scheme: Lowest score dropped; each of the remaining four worth 10%
- Structure: Each assignment is structured as a Jupyter notebook
  - Most are practical, coding-based problems, involving completing code segments provided in the notebooks or developing models from scratch
- Submission: Through GradeScope in BruinLearn
  - No late submissions permitted; each student may request a **one-day** extension for one of the five assignments with no penalty, provided they inform the TA **before the deadline**

Participation bonus (5%)

- NO pop-up quizzes during lectures
- Active participation in answering other students’ questions on Piazza earns bonus points

Course project (30%)

- Team work: at most 6 students per group
- Scope: KDD Cup 2024 Open Academic Graph Challenge
  - WhoIsWho-IND: Given the paper assignments of each author and paper metadata, the goal is to detect paper assignment errors for each author
  - AQA: Given professional questions and a pool of candidate papers, the objective is to retrieve the most relevant papers to answer these questions
  - PST: Given the full texts of each paper, the goal is to automatically trace the most significant references that have inspired a given paper

  - Choose one of the above three tracks

Course project (30%)

- Large language model policy:
  - For all tracks, pre-trained models that have been open-sourced before the end of the competition are allowed to be used
  - WhoIsWho and IND allow the use of APIs
  - After a valid submission to the validation set, participating teams can obtain a free quota of 1 million tokens for the GLM-4 API
- Important deadlines:

Week 2	April 12	Team formation
Week 10	June 10 & 12	In-class project presentation
Week 10	June 12	Report & code submission

Course project (30%)

- Evaluation criteria:
  - Project presentation (5%)
  - Final report (10%)
  - Code repository (10%)
  - Peer review (5%)
  - Leaderboard performance (bonus, up to 5%)

- Detailed requirements can be found in the project guidelines

Course project (30%)

- Leaderboard performance (bonus, up to 5%)
  - A baseline performance will be set
  - Teams surpassing the baseline threshold in the leaderboard will be awarded bonus points
  - The leaderboard range between the baseline and the top 1 team will be divided into 5 equally spaced sections
  - Bonus points will be awarded based on the section the team falls into:
    - Section 1 (top 20%): 5% bonus
    - Section 2 (top 40%): 4% bonus
    - Section 3 (top 60%): 3% bonus
    - Section 4 (top 80%): 2% bonus
    - Section 5 (above baseline): 1% bonus

Course project (30%)

- Tutorials and project hackathon
  - Friday’s discussions will be dedicated to providing support for the course projects
  - Students can utilize this opportunity to collaborate with their teammates, seek guidance from the instructor and TA, and make progress on their projects
    - Each team is encouraged to work on their course projects together during the allocated discussion time
  - The instructor and TA will be available to answer questions, provide feedback, and assist with any challenges during the Friday sessions

Course project (30%)

- Team formation:
  - Team sign-up form: https://1drv.ms/x/s!AsVRzCssZoYOiZ8UBRIpK8g-i-A5sA?e=JqRjHb
  - Deadline: April 12, 11:59PM
- Looking for teammates?
  - Use Piazza to collaborate with other classmates and form your teams
  - Check the sign-up form and email the team leader to see if you are a good fit
  - Do NOT email TA regarding your team assignment

Course project (30%)

- Project deliverables:
  - Presentation: In-class presentation in Week 10
  - Final report: Up to 8 pages using the NeurIPS LaTeX template
  - References and appendices do not count towards page limitation
- Submissions of report and code will be made through GradeScope
  - One submission per team
  - Late submission is NOT allowed

Exams (30%)

- In-class exam on May 15 (Wednesday, Week 7)
- Exams are in-person only; no online or make-up exams offered
- Format:
  - Close-book but two letter-sized cheat sheets allowed
  - Simple calculators allowed
  - Internet access strictly prohibited
- Scope: all topics prior to Week 7

Lecture schedule

Week 1	4/1	Administrivia & Introduction	
	4/3	Linear Regression & Backpropagation	HW 1 out
Week 2	4/8	Logistic Regression, MLP, & Evaluation	
	4/10	Regularization & Lasso	HW 2 out
Week 3	4/15	Decision Trees, Bagging, & Random Forests	Team formation due
	4/17	Boosting, Ensemble Selection, & MoE	HW 3 out
Week 4	4/22	Clustering (K-Means, GMM) & Dimensionality Reduction	
	4/24	Latent Factor Models & Matrix Factorization	HW 4 out
Week 5	4/29	Discrete Representation Learning (VAE & VQ-VAE)	
	5/1	Graphs and Networks: Random Walks	HW 5 out

Lecture schedule

Week 6	5/6	Graphs and Networks: Label Propagation & Spectral Clustering	
	5/8	Language Models: Word Vectors & Seq2Seq Language Models	
Week 7	5/13	Language Models: Transformers	
	5/15	**In-class Exam**	
Week 8	5/20	Language Models: Pre-Training	
	5/22	Language Models: Post-Training (SFT, RLHF)	
Week 9	5/27	*Memorial Day (No Class)*	
	5/29	Language Models: Planning & Reasoning	
Week 10	6/3	**Project Presentation**	
	6/5	**Project Presentation**	