CSE 40647/60647 Data Science

Tuesday and Thursday, 2:00 - 3:15 PM, 126 DeBartolo

Description:

Data mining uses methods from multiple fields for scientific studies. The focus of this course will primarily be on **machine learning** concepts and methods, with relevant inclusions and references from probability, statistics, pattern recognition, databases, information theory, and visualization. The course will give students an opportunity to implement and experiment with some of the concepts (e.g., **classification, clustering, frequent pattern mining**), and also apply them to the real-world data sets.

Instructor:

Meng JIANG, Ph.D., mjiang2@nd.edu
Office: 326C Cushing Hall Office hours: 3:30 - 4:45 PM **Tuesday** (except 8/27 and 11/19)

** Please email me to schedule any appointment outside of the office hours.

Graduate Teaching Assistants (GTAs):

Wenhao YU, wyu1@nd.edu Tong ZHAO, tzhao2@nd.edu
** GTAs do not offer office hours. Please email them for appointments or *tutorials* (e.g., offering instructions of using Anaconda, Pandas, etc., if necessary)

Sakai: (Assignments, Test, Quizzes, Chat Room, Forums, Polls, Calendar, etc.)

https://sakailogin.nd.edu/portal/site/FA19-CSE-40647-CX-01

Text Book (not mandatory):

- Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers. (3rd edition) ***PDF available on Sakai***

Prerequisites:

Programming with Python
Data structures and algorithms
Basic knowledge in Probability & Stats

Courtesy Policy:

You are expected to contribute to promoting a healthy learning environment. Please make sure what you do is not disruptive to others in the classroom.

Classroom Recording Notification:

This course will be recorded using Panopto. This system allows us to automatically record and distribute lectures in a secure environment. You can watch these recordings anytime, anywhere, on any device. In Sakai, look for the "Panopto" tool on the left hand side of the course. Because we will be recording in the classroom, your questions and comments may be recorded. Recordings typically only capture the front of the classroom, but if you have any concerns about your voice or image being recorded please speak to me to discuss your concerns. Except for faculty and staff who require access, no content will be shared with individuals outside of your course without your permission. These recordings are jointly copyrighted by the University of Notre Dame and your instructor. Posting them to other

websites (including YouTube, Facebook, SnapChat, etc.) or elsewhere without express, written permission may result in disciplinary action and possible civil prosecution.

Course Goal and Objectives:

Goal: To *introduce* students to several fundamentals of data science, including data preprocessing, classification, clustering, and frequent pattern mining.

Objectives: By the end of the course, students will be able to:

- **Use** data **preprocessing** techniques to understand data: data description, data visualization, cleaning, integration, data reduction, and dimension reduction.
- **Use** Decision Trees, Naïve Bayes, and SVMs for **classification**.
- **Describe** Ensembles and Neural Networks models for **classification**.
- **Use** K-Partitioning methods for **clustering**.
- **Describe** hierarchical, kernel-based, and density-based **clustering**.
- **Use** Apriori and FP-Growth for **frequent pattern mining** and association mining.
- **Describe** diverse patterns, sequential patterns, graph patterns.
- **Use** appropriate measures to **evaluate** results of different functionalities.

** "Use": implement with Python or Python libraries (e.g., scikit-learn).
** "Describe": explain the ideas and methodologies.

Grading Policy:

- Written & programming assignments (30%): typed and submitted to Sakai. Each chapter has written/programming assignment. So according to the following course schedule (4 chapters), we have 4 times of homework in the semester.
- In-class exercises (quizzes) (15%): 6 in the semester, random lectures, at the end of the lectures (3:05 - 3:20 PM), 6 single-choice questions each quiz on Sakai. The highest five will be counted for final score (the lowest is dropped).
- Course project (20%): team project, 3-4 person a team, submitted to Sakai.
- Mid-term exam (15%): October 10, 2019
- Final exam (20%): December 19, 2019

** It is important that you contact the instructor if there is any time conflict.

Letter Grades:

- A: [93, +∞)
- A-: [90, 93)
- B+: [87, 90)
- B: [84, 87)
- B-: [81, 84)
- C+: [78, 81)
- C: [75, 78)
- C-: [72, 75)
- D: [60, 72)
- F: (-∞, 60)

Course Schedule: (29 lectures)

Tuesday	Thursday	
8/27: L01 Introduction	8/29: L02 Description	**Data Preprocessing**
9/3: L03 Visualization	9/5: L04 Cleaning	
9/10: L05 Reduction	9/12: L06 Project Proposal	
9/17: L07 Decision Trees	9/19: L08 Naïve Bayes	**Classification (supervised machine learning)**
9/24: L09 SVM and Kernels	9/26: L10 Class. Evaluation	
10/1: L11 Ensembles	10/3: L12 Neural Networks	
10/8: L13 Course Review I	**10/10: L14 Mid-term Exam**	
10/15: L15 Proj. Milestone I	10/17: L16 Proj. Milestone II	
Mid-term Break		
10/29: L17 Programming!	10/31: L18 K-Means	**Clustering (unsupervised machine learning)**
11/5: L19 DBSCAN	11/7: L20 PCA & SVD	
11/12: L21 Clus. Evaluation	11/14: L22 Apriori	
11/19: L23 FP-Growth	11/21: L24 FP Evaluation	**Frequent pattern mining (unsupervised machine learning)**
11/26: L25 Diverse Patterns	Thanksgiving break	
12/3: L26 Project Final I	12/5: L27 Project Final II	
12/10: L28 Project Final III	12/12: L29 Course Review II	
12/19: Final Exam (10:30am-12:30pm)		

Assignment Policies:

Submission and dues

- The assignments and dues will be announced on Sakai.
- Only soft copy submissions at Sakai are accepted.
  - Texts and simple formulae must be typed.
  - Figures/tables/complicated formulae may be hand-drawn.
  - You must submit one zipped folder which includes one PDF file and multiple code files (e.g., .py).
- Student's full name should appear on the first page.

**There will be -10% penalty if any of the above rules is violated.

Deadlines

Every assignment is due at midnight (11:55pm) with some grace period.

- After the due time, submissions will be accepted with -33% penalty until the solution is posted.
- The instructor will post the solution at any time after 7am the next day.
- No more submission will be accepted after the solution is posted.

Regrading

Request of regrading must be emailed to the grader within one week after the scores are released.

- Explanation of the issues and reasons behind the request.
- Original submission with grading to be reconsidered highlighted (e.g., in screenshots).
- Please cc the instructor (mjiang2@nd.edu) if the regrading is requested to the GTAs.

Release of grades will be announced by email. Failure to see the announcement is not a valid excuse for late requests.

Project Policies:

** More details will be given in *Project Instruction* document.

Grading distribution

Proposal (10%), Milestone (30%)

Final paper & Presentation (40%), Code/Data (20%), Discussion bonus (+5%)

Teaming

- Each team should have three to four students. A mixture of undergraduates and graduates is allowed. The teams will be evaluated uniformly no matter they have undergraduates or graduates.

Required items and dues

- Proposal
  - Paper (PDF): title, problem definition, potential solutions, data sources, proposed evaluation methods, project plan/timeline **due 9/9**

- Milestone
  - Paper (PDF): introduction, related work, problem definition, one or more method that has been used, some other potential solutions, data and experiment settings, evaluation methods, preliminary results and experimental analysis, timeline **due 10/18**
  - Presentation (PPTX) for selective projects: **due 10/15, 10/17** in class
- Final
  - Paper (PDF): introduction, related work, problem definition, solutions and methods that have been used, data and experiment settings, evaluation methods, experimental results and analysis (tables and figures), discussion and future work, conclusions **due 12/13**
  - Presentation (PPTX): **due 12/3, 12/5, 12/10** in class.
- Data and code package (ZIP): **due 12/13**

The papers (PDF) should be formatted according to the new Standard ACM Conference Proceedings Template: https://www.acm.org/publications/proceedings-template

There will be -33% penalty if the paper is not formatted correctly.

Deadlines

Every project required item is due at midnight (11:55pm) with some grace period. There will be -33% penalty for each 24 hours past the deadline.

Requesting Excused Absences:

There will be in-class exercises (quizzes), *randomly distributed in the semester*, and 0 will be recorded if a student misses a class. This will be excused if a valid excuse is requested and approved by the instructor ***before the class***, then the student can *do the exercise in office hours*. Some examples of the excuses are:

- University-related absences.
- Sickness.
- On-site interviews.

If excuses are requested before missing the class, they will be processed immediately, and you don't need to provide evidences unless the requests are repetitive. If excuses are requested after missing the class, they will be processed by the instructor's discretion.

Honesty Policies:

"As a member of the Notre Dame community, I will not participate in or tolerate academic dishonesty." – The Honor Code Section II.

"Students should familiarize themselves with the directives given by the instructor in each class concerning what is and is not permitted, especially in matters of group projects, lab reports, written papers and the attribution of research to sources (footnoting), including the Internet." – The Honor Code Section IV. A.3.

The general rule is described in the following table: The CSE Guide to the Honor Code (http://cse.nd.edu/undergraduates/honor-code) discusses the boundaries between resources and solutions and between consulting and copying.

Peer discussion

For both written and programming assignments, discussions are highly encouraged. However, source codes/solutions cannot be directly seen or shared under any circumstance. For example, you may discuss how you arrived at your solution/source codes, but you are not allowed to show yours to other peers directly or send your solution/source codes to a classmate.

Referencing online resources

You are encouraged to study with online resources with the following rules:

- If your learned from sth., you do not need to cite the source.
- If you copied from sth., you have to cite the source by giving its full URL.
    - Direct copy & paste is allowed as long as what you copied is marked with double quotation marks.
    - In case you are re-using others' source codes, you may use comments to declare it.

  There will be absolutely no penalty for re-using someone else's idea/codes/answers.

Explicitly citing all sources is tedious and sometimes even troublesome. However, negligence in doing so often constitutes plagiarism which is serious academic dishonesty. Learning how to build upon others' intellectual works is an important part of the education of Computer Science and Engineering students.

Plagiarism

The following behavior constitutes plagiarism.

- Copying & pasting whole or part of the sentences with more than six consecutive words without citation.
- Rephrasing whole or part of the sentences with more than six consecutive words without citation.
- Copying & pasting whole or part of diagrams/figures without citation.
- Redrawing whole or part of diagrams/figures without citation.

On the other hand, all above are allowed if explicit citation is provided with a full URL.

More references to check

For more information about the honest code of Notre Dame and/or CSE, please refer to following links:

- https://honorcode.nd.edu/the-honor-code/
- http://cse.nd.edu/academics/honor-code

If you discover any inconsistency between this document and the university or department honest code, please contact the instructor. Sometimes the course policy overrides and sometimes the university/department policy overrides.