Data Science CSE 40647/60647

Tuesday and Thursday, 2:00 pm to 3:15 pm, 140 DeBartolo

Instructor:

Dr. Meng Jiang, mjiang2@nd.edu
Office: 326C Cushing Hall Phone: (574) 631-7454

Teaching Assistant (TA):

Qi Li, qli8@nd.edu
Office: 247 Fitzpatrick Hall

Office hours:

Instructor: Thursday, 3:30 pm to 4:30 pm, 326C Cushing Hall
TA: Tuesday, 3:30 pm to 4:30 pm, 247 Fitzpatrick Hall
Piazza: https://piazza.com/class/j6dmfs52c6d5ov
Please email me (mjiang2@nd.edu) to schedule any appointment outside of the office hours.

Course webpage: http://www.meng-jiang.com/teaching-csex0647.html

Text Book (not mandatory to have it):

- Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers.
- We provide lecture notes from the 2nd edition of the text book on course website.

Prerequisites:

Programming with Python
Data structures and algorithms

As a prerequisite for:

CSE 40625/60625: Machine Learning

Course Goals:

At the end of the course, students will be able to:

- Process raw data: data cleaning, data integration, data reduction, dimension reduction
- Describe data warehouse, OLAP, data cube concepts and technology that work on multi-dimensional data
- Use Apriori and FP-Growth for frequent pattern mining
- Describe diverse patterns, sequential patterns, graph patterns
- Use Decision Tree, Naïve Bayes, Ensembles for classification

- Describe SVMs and Neural Networks for classification
- Use K-Partitioning Methods (K-Means, etc.) for clustering
- Describe Kernel-based Clustering and Density-based Clustering
- Use appropriate measures to evaluate results of different functionalities

Grading:

- Homework assignments: 25% = 5%*5
- Class project: 25%
- Mid-term: 20%
- Final: 30%

Any discrepancy in the grade should be brought to the attention of the instructor within **7** days of grade assignment.

Assignments:

- The assignments will require individual effort.
- The assignments will comprise of paper-based and computer exercises.

Class Project:

- The assignments will require individual effort.
- The students will be required to write a term paper.
- The progress will be monitored in homework assignments.
- The students will volunteer to make a class presentation on their project and be graded by classmates and the instructor. Students that do not present will be graded by the instructor.
- The term paper will go through a peer review process among the classmates.
- Programming language will be **Python**.

Course Policies:

- Collaboration Policy

Unless instructed otherwise, students must turn in work that is their own. Students must write their own code, run their own data analyses, and write up their own results and answers to assignment questions.

- Assignments

There will be several assignments. Unless announced otherwise, assignments will be due at 11:59pm ET on the provided submission date.

- Quizzes and Exams

There will be no quizzes. There will be one mid-term exam and one final exam. Make-up exams will be allowed as per the du Lac Class Absence Policy. Whenever possible, students are expected to provide advance notice if they will be unable to take an exam. Make-up exams for travel to academic events will be provided at the discretion of the instructor.

Students are not allowed to use any electronic equipment (laptop, calculator, etc.). Students are allowed to have two-page references for the exams but not more than one letter-size piece of paper.

- Course Project

For the course project, students will be expected to use the provided dataset, follow the instructions, write a term paper, and volunteer to present in class.

- Late Policy

Assignments submitted after the submission deadline but within the next day are counted as one day late. The next 24 hours will be counted as two days late, and so on. Each day late contributes a 33% penalty to the original assignment value.

- Office Hours

Students are encouraged to take advantage of the instructor and TA's office hours, and schedule additional time as needed.

Academic Dishonesty:

- The CSE and du lac honor code will be strictly followed.
- All assignments are individual unless instructed. You can discuss the assignment at a high level, but you should independently and individually write down the answers and/or the program. The sharing and copying of homework solutions or programs or functions or exams will be considered cheating.
- All the references and sources should be carefully provided and cited.
- Entering Notre Dame you were required to study the on-line edition of the Academic Code of Honor, to pass a quiz on it, and to sign a pledge to abide by it. The full Code and a Student Guide to the Academic code of Honor are available at: http://honorcode.nd.edu.
- Perhaps the most fundamental sentence is the beginning of section IV-B: "The pledge to uphold the Academic Code of Honor includes an understanding that a student's submitted work, graded or ungraded – examinations, draft copies, papers, homework assignments, extra credit work, etc. - must be his or her own."

08-22T	**Introduction**	10-12R	Classification: Naïve Bayes
08-24R	Data description	10-24T	Classification: Evaluation
08-29T	Data visualization	10-26R	Classification: Ensembles
08-31R	**Project introduction**	10-31T	Classification: SVMs
09-05T	Data cleaning and data integration	11-02R	Classification: Neural networks
09-07R	Data reduction and dimension reduction	11-07T	Clustering: Concepts
09-12T	Data cube: Concepts and operations	11-09R	Clustering: Partitioning methods
09-14R	Data cube: Data warehouse and OLAP	11-14T	Clustering: Kernel-based
09-19T	Frequent pattern mining: Apriori	11-16R	Clustering: Density-based
09-21R	Frequent pattern mining: FP-Growth	11-21T	Clustering: Evaluation
09-26T	Frequent pattern mining: Evaluation	11-28T	**Course review 2**
09-28R	Frequent pattern mining: Beyond itemset	11-30R	**Course review 3**
10-03T	**Course review 1**	12-05T	**Project presentation 1**
10-05R	**Mid-term**	12-07R	**Project presentation 2**
10-10T	Classification: Decision tree induction	12-12T	**Final**

08-22T	Introduction	10-12R	**HW4 out**
08-24R	Data processing	10-24T	
08-29T	**HW1 out**	10-26R	
08-31R	Project introduction **Project out**	10-31T	
09-05T		11-02R	
09-07R		11-07T	Clustering
09-12T	Data cube **HW1 due, HW2 out**	11-09R	**HW4 due, HW5 out**
09-14R		11-14T	
09-19T	Frequent pattern mining	11-16R	
09-21R	**HW2 due, HW3 out**	11-21T	
09-26T		11-28T	Course review 2 **HW5 due**
09-28R		11-30R	Course review 3 **Project due**
10-03T	Course review 1 **HW3 due**	12-05T	Project presentation 1
10-05R	Mid-term	12-07R	Project presentation 2
10-10T	Classification	12-12T	Final