CS 459: Topics in Computer Science: Data Mining

G. Exoo

Fall, 2022

E-mail: cs457@cs.indstate.edu Web: **cs.indstate.edu/cs457**
Office Hours: Online (Zoom) MWF 12:30–1:00 and MW 3:00–3:30 Class Hours: MWF 11:00-11:50am
Class Room: A-017 Root Hall and Online (Zoom)

Course Description

This course deals with Data Mining. We will study algorithms used to extract useful information from massive datasets.

Programming Prerequisites

Python.

Learning Outcomes

Upon successful completion of this course, students should be able to:

- Use command line tools to preprocess large datasets (`grep`, `sed`, `awk`, etc.).
- Use statistical and graphical tools to summarized datasets.
- Perform basic data cleaning operations.
- Use basic data reduction methods (sampling, binning, SVD, etc.).
- Understand and use the basic regression, classification and clustering algorithms.

Online Texts/References

There is a wealth of information on Data Mining available online. Links will be provided on the class web page. If you need to review Python, you should go through the Standard Python Tutorial, shown below. Links where you can learn about a few other important Python modules are also shown.

The Standard Python Tutorial

NumPy for Beginners

Matplotlib: Basic Usage

10 Minutes to Pandas

Expected Amount of Work

If you take this class seriously and get what you should out of it, some weeks you will likely be spending an average of 15 hours/week or more on the class. The students who get A's in their CS courses and have an easy time finding jobs spend at least this much time. Not everyone will need to spend this much time and not all weeks will be the same, but you should plan on putting in whatever time it takes. Note - your classes should be more important than your part-time job.

Course Announcements

Announcements regarding the course will be made both during class, on the class web page, and to your cs email (which you should learn how to use). Communications related to grades will be sent to your @sycamores.indstate.edu email address. You should regularly check these locations.

Course Outline

Since this is a 600 level graduate class, students will be expect to remedy any deficiencies in SQL or Python on their own. This will be best accomplished by going through the tutorials listed above.

The main topics in the course are listed below. This is intended as a topical outline, not a timeline. Database and Machine Learning concepts will be developed together. A dynamic timeline for the class can be found at the end of this document.

Course Topics

- Introduction to Data Mining
- Map Reduce
  - Distributed File Systems
  - MapReduce Operation
    - Map
    - Reduce
    - Combine
  - MapReduce Algorithms
    - Relational Algebra Operations
    - Matrix Multiplication
  - Complexity of MapReduce Steps
- Similarity
  - The many types of hashing used in Data Mining
  - Nearest Neighbor Search
    - Set (Jaccard) Similarity
    - Document Similarity
  - Shingling
    - Shingle sizes
    - Hashing shingles
  - Minhashing
    - Minhashing and Jacard similarity
    - Minhash signatures
  - Locally sensitive hashing
  - Distance measurement
    - Euclidean distances
    - Jacard distance
    - Hamming distance
    - Edit distance
    - Cosine distance
  - Applications
- Mining Streams
  - The stream model vs. the RAM model.
  - Problems
    - Counting

        * Counting distinct elements
        * Heavy hitters
        * Moments
    - Algorithms
        * Bloom filters
        * Flajolet-Martin Algorithm
        * Alon-Matias-Szegedy Algorithm
        * Exact counts (DGIM Algorithm)
        * Decaying Windows

- Mining the Web
    - Page Rank
        * Structure of the Web
        * Dead ends
        * Spider traps
        * Efficient computation
        * Topic sensitive page rank
    - Spam
        * Spam farms
        * Trust rank
        * Hubs and authorities

- Frequency Measures
    - Market baskets
    - The a-priori algorithm
    - The Eclat algorithm
    - The FP-Growth algorithm

- Clustering
    - Geometry
    - The curse of dimensionality
    - Hierarchical clustering
    - k-Means
        * The *k*-Means algorithm
        * Intialization
        * Picking *k*
        * The Bradly, Fayyad, Reina Algorithm
    - Clustering in streams

- History of Data Models and Data Bases

  - Hierarchical Model
  - Network Model
  - Relational Model
    - Semantic Model
    - Object Oriented Model
  - Semi-structured Models
- Relational Model
  - Relational Algebra
  - SQL
    - Using SQL
    - Relational DBMS
      - SQLite
      - Postgres
      - DB2, Oracle, Sybase, MySQL, etc.
  - UML
- Current Trends
  - NOSQL
    - Key-Value Store
    - Document Store
      - XML
      - JSON
      - XPATH
      - XQUERY

Assignments

The students in this course have the following responsibilities: read assigned readings before lecture, attend lecture, complete homework assignments, take in-class quizzes, take exams, and complete all projects. In this class, your final grade will be based primarily on the projects, and to a lesser extent on the exams and quizzes.

Policies

Note that this course follows all standard CS course policies. In particular check the CS course policies related to - cheating/plagiarism, attendance, missing exams. See http://cs.indstate.edu/info/policies.html for details.

All assignments are posted in a pdf file on the class web page. Each such file will indicate the number of points, the due date and time, and the location where your assignment should be saved. Failure to save your work in the correct location will be viewed as equivalent to not doing the work.

Grade Components

Quizzes - 15

Midterm Exam - 15

Programming Projects - 55

Final Exam - 15

Late Assignments

The maximum points you will receive for late assignments will decay exponentially with time. If n is the number of points the assignment is worth, then a perfect assignment that is between $d-1$ and d days late will net at most $n/(2^d)$ points. [1]

We suggest attempting a homework assignment the day it is given, or the day after, so that if you do not understand how to do the assignment, you will have time to seek help. You may need to ask for help more than once, and you should certainly plan on spending a lot of time in the CS lab (A-015 Root Hall). Many of the homework assignments require thought and creative problem solving. Do not expect to solve the problems the first time you attempt them.

Grading Policy

We try to design homework assignments and exams so that a standard cutoff for grades will be close to what you deserve. After the first exam a grade will be created in Blackboard called **Letter Grade** that is intend to be your current grade in the class. The grades are generally based on the following table.

A	93-100
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	63-67
D-	60-63
F	0-60

Grades are intended to indicate your mastery of the course material. The following are offered as guidelines.

[1] Quiz: If an assignment is worth 100 points and you hand it in one week after it is due, what is the maximum number of points you could receive?

A

You can do all the assignments on your own.

B+/A-

The student understands almost everything, and should be able to use this knowledge in other courses or in a job.

B-/B

The student understands most, but not all, topics well.

C/C+

Learned enough and have the minimum skills to move on in the subject.

D+/C-

The student made some effort in, and understands some things at a high level, but hasn't mastered the details well enough to be able to use this knowledge in the future.

D-

Students will normally not get an F if - they attend 80% of the lectures, complete some of the assignments up through the end of the course, and get nearly half of the problems on the final exam correct.

F

Normally, students that get an F simply stopped doing the required work at some point.

Blackboard

The course has a blackboard site. You should see this course listed under your courses for the current term. The blackboard site is used only for giving you your grades (go to the course in blackboard, then click *My Tools*, and then *My Grades*.

Academic Integrity

Read CS course policies in terms of what is and is not allowed on assignments: http://cs.indstate.edu/info/policies.html. Please ask the instructor if you have doubts about what is considered cheating in this course.

Special Needs / Student Disabilities

Indiana State University recognizes that students with disabilities may have special needs that must be met to give them equal access to college programs and facilities. If you need course adaptations or accommodations because of a disability, please contact us as soon as possible in a confidential setting either after class or in my office. All conversations regarding your disability will be kept in strict confidence. Indiana State University's Student Support Services (SSS) office coordinates services for students with disabilities: documentation of a disability needs to be on file in that office before any accommodations can be provided. Student Support Services is located on the lower level of Normal Hall in the Center for Student Success and can be contacted at 812-237-2700, or you can visit the ISU website under A-Z, Disability Student Services and submit a Contact Form. Appointments to discuss accommodations with SSS staff members are encouraged. Once a faculty member is notified by Student Support Services that a student is qualified to receive academic accommodations, a faculty member is obligated to provide or allow a reasonable classroom accommodation under ADA.

Disclosures Regarding Sexual Misconduct

Indiana State University fosters a campus free of sexual misconduct including sexual harassment, sexual violence, intimate partner violence, and stalking and/or any form of sex or gender discrimination. If you disclose a potential violation of the sexual misconduct policy I will need to notify the Title IX Coordinator. Students who have experienced sexual misconduct are encouraged to contact confidential resources listed below. To make a report or the Title IX Coordinator, visit the Equal Opportunity and Title IX website: http://www.indstate.edu/equalopportunity-titleix/titleix.

The ISU Student Counseling Center

HMSU 7th Floor, 812-237-3939, www.indstate.edu/cns.

The ISU Victim Advocate

Trista Gibbons trista.gibbons@indstate.edu. HMSU 7th Floor, 812-237-3939 (office), 812-230-3803 (cell).

United Campus Ministries

321 N 7th St., Terre Haute, IN 47807 812-232-0186,

Covid-19 Issues

All things related to Covid-19

The ISU Victim Advocate: Leah Reynolds

HMSU Room 813

812-237-3829 (office)

812-243-7272

leah.reynolds@indstate.edu

http://www2.indstate.edu/sao/campusministries.htm

http://www.unitedcampusministries.org

mailto:ucmminister2@gmail.com.

For more information on your rights and available resources:
http://www.indstate.edu/equalopportunity-titleix/titleix.