UNIVERSITY of HOUSTON

COLLEGE OF NATURAL SCIENCES & MATHEMATICS HTTP://NSM.UH.EDU

COURSE TITLE/SECTION: Data Mining (COSC 6335) *August 20, 2020*

TIME: TT 2:30-4p

FACULTY: Christoph F. Eick **OFFICE HOURS**: TU 4-5p TH 11:30a-12:30p

E-mail: ceick@uh.edu **Phone: 33345 (use e-mail!!)** **FAX: 33335**

I. Course *Data Mining (COSC 6335)*

A. Catalog Description

Goals and objectives of data mining, data quality, data preprocessing, OLAP and data warehousing, exploratory data analysis, classification and prediction, similarity assessment, cluster and outlier analysis, association analysis, post processing techniques, data mining methodologies, data mining case studies.

B. Purpose

Data mining centers on finding novel, interesting, and potentially useful patterns in data. It aims at transforming a large amount of data into a *well of knowledge*. Data mining has become a very important field in industry as well as academia. The course covers most of the important data mining techniques and provides background knowledge on how to conduct a data mining project. Topics covered in the course include exploratory data analysis, classification and prediction, clustering and similarity assessment, association analysis, outlier and anomaly detection, and interpreting and interpreting data analysis/data mining results. In the first 9 weeks a very basic introduction to data mining will be given. After defining what knowledge discovery and data mining is, data mining tasks such classification, clustering, and association analysis will be discussed in detail. Also basic visualization techniques and statistical methods will be introduced. Moreover, hands on data mining experience will be provided in three problem sets. Finally, you will learn on how to use and do programming in the popular statistics, visualization, and data mining environment *R*.

II. Course Objectives

Upon completion of this course, students

1. will know what the goals and objectives of data mining are and how to conduct a data mining project
2. will have sound knowledge of popular classification techniques, such as decision trees, support vector machines and neural networks.
3. will know the most important association analysis techniques
4. will have detailed knowledge of popular clustering algorithms such as K-means, density-based, graph-based, and hierarchical clustering.
5. will obtain some basic knowledge about popular outlier detection techniques
6. will conduct small and medium-sized projects in which data mining is applied to real world data sets. They will obtain valuable experience in learning how to interpret and evaluate data mining results, how to select parameters of data mining tools, and how to make sense out of data.
7. will get some practical experience in evaluating data mining results of other students in the course as well as data mining publications. Kritik (https://www.kritik.io/) will be used for some evaluation tasks of the course.
8. will obtain practical experience in designing and implementing data mining algorithms
9. will learn on how to use popular data mining programming environment **R**.

III. Course Content

I. Introduction to Data Mining
II. Exploratory Data Analysis
III. A Short Introduction to R
IV. Introduction to Classification: Basic Concepts and Decision Trees, Support Vector Machines and Neural Networks.
V. Association Analysis —Rule, Sequence, Graph and Collocation Mining
VI. Outlier and Anomaly Detection
VII. Introduction to Clustering and Similarity Assessment
VIII. More on Clustering: Hierarchical, Density-based, and Graph-based Clustering.
IX. Spatial Data Mining *only if enough time*
X. Data Storytelling *only if enough time*
XI. Data Preprocessing

IV. Course Structure

23 lectures
2 exams
3 problem sets
1 student presentation
2 40-minute review sessions

V. Problem Sets

Problem Sets contain paper and pencil tasks which review your understanding of basic data mining concepts and algorithms, tasks which use data mining tools, and small and medium sized data analysis/data mining projects, and tasks in which you evaluate data mining results of other students and data mining publications. Some tasks will be group tasks. There will be three Problem Sets in Fall 2020:

Problem Set1: Exploratory Data Analysis, Classification, and Evaluating Data Mining Results
Problem Set2: Association Analysis and Outlier Detection
Problem Set3: Clustering and Data Mining Paper Reviewing

VI. Textbooks

Highly Recommended Text:

P.-N. Tang, M. Steinback, and V. Kumar *Introduction to Data Mining*,
Addison Wesley, Second Edition.

Recommended Text:

Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*
Morgan Kaufman Publishers, Third Edition

VII. Evaluation and Grading

Problem Set1: 12% (+ 4% evaluation)
Problem Set2: 14% (+ 2% evaluation)
Problem Set3: 12% (+4% evaluation)
Evaluation: 10-12%
Spontaneous Online Credit: 2-4%
Midterm Exam: 20%
Final Exam: 26%
Attendance: 2%

Students will be responsible for material covered in the lectures and assigned in the readings..

Translation number to letter grades:
A:100-90 A-:90-86 B+:86-82 B:82-77 B-:77-74 C+:74-70
C: 70-66 C-:66-62 D+:62-58 D:58-54 D-:54-50 F: 50-0

Students may discuss course material and homeworks, but must take special care to discern the difference between **collaborating** in order to increase understanding of course materials and collaborating on the homework / course project itself. We encourage students to help each other understand course material to clarify the meaning of homework problems or to discuss problemsolving strategies, but it is **not** permissible for one student to help or be helped by another student in working through homework problems and in the course project. If, in discussing course materials and problems, students believe that their like-mindedness from such discussions could be construed as collaboration on their assignments, students must cite each other, briefly explaining the extent of their collaboration. Any assistance that is not given proper citation may be considered a violation of the Honor Code, and might result in obtaining a grade of F in the course, and in further prosecution.

Policy on grades of I (Incomplete): A grade of 'I' will only be given in extreme emergency situations and only if the student completed more than 2/3 of the course work.

VIII. Course and UH Policies

Excused Absence Policy

Regular class attendance, participation, and engagement in coursework are important contributors to student success. Absences may be excused as provided in the University of Houston Undergraduate Excused Absence Policy and Graduate Excused Absence Policy for reasons including: medical illness of student or close relative, death of a close family member, legal or government proceeding that a student is obligated to attend, recognized professional and educational activities where the student is presenting, and University-sponsored activity or athletic competition. Additional policies address absences related to military service, religious holy days, pregnancy and related conditions, and disability.

Recording of Class

Students may not record all or part of class, livestream all or part of class, or make/distribute screen captures, without advanced written consent of the instructor. If you have or think you may have a disability such that you need to record class-related activities, please contact the Center for Students with DisABILITIES. If you have an accommodation to record class-related activities, those recordings may not be shared with any other student, whether in this course or not, or with any other person or on any other platform. Classes may be recorded by the instructor. Students may use instructor's recordings for their own studying and notetaking. Instructor's recordings are not authorized to be shared with *anyone* without the prior written approval of the instructor. Failure to comply with requirements regarding recordings will result in a disciplinary referral to the Dean of Students Office and may result in disciplinary action.

Syllabus Changes

Due to the changing nature of the COVID-19 pandemic, please note that the instructor may need to make modifications to the course syllabus and may do so at any time. Notice of such changes will be announced as quickly as possible through (*specify how students will be notified of changes*).

UH Email

Email communications related to this course will be sent to your Exchange email account which each University of Houston student receives. The Exchange mail server can be accessed via Outlook, which provides a single location for organizing and managing day-to-day information, from email and calendars to contacts and task lists. Exchange email accounts can be accessed by logging into Office 365 with your Cougarnet credentials or through Acccess UH. They can also be configured on IOS and Android mobile devices. Additional assistance can be found at the Get Help page.

Course Delivery and Final Exams

This course is being offered in the Synchronous Online format. Synchronous online class meetings will take place according to the class schedule. There is no face-to-face component to this course. In between synchronous class meetings, there may also be asynchronous activities to complete (e.g., discussion forums and assignments). This course will have a final exam per the University schedule. The exam will be delivered in the synchronous online format, and the specified date and time will be announced during the course. Prior to the exam, descriptive information, such as the number and types of exam questions, resources and collaborations that are allowed and disallowed in the process of completing the exam, and procedures to follow if connectivity or other resource obstacles are encountered during the exam period, may be provided.

IX. Bibliography

The course textbook contains a detailed data mining bibliography. Moreover, the following conferences center on data mining and related areas:

1. Data mining and KDD
   - Conference proceedings: ICDM, KDD, PKDD, PAKDD, SDM, MLDM etc.
   - Journal: Data Mining and Knowledge Discovery
2. Database field (SIGMOD member CD ROM):
   - Conference proceedings: VLDB, ICDE, ACM-SIGMOD, CIKM
   - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.

3. AI and Machine Learning:
    - Conference proceedings: ICML, AAAI, IJCAI, etc.
    - Journals: Machine Learning, Artificial Intelligence, etc.
4. Statistics:
    - Conference proceedings: Joint Stat. Meeting, etc.
    - Journals: Annals of statistics, etc.
5. Visualization:
    - Conference proceedings: CHI, etc.
    - Journals: IEEE Trans. visualization and computer graphics, etc.

X. Helpful Links

COVID-19 Updates: https://uh.edu/covid-19/

Coogs Care: https://www.uh.edu/dsaes/coogscare/

Laptop Checkout Requests: https://www.uh.edu/infotech/about/planning/off-campus/index.php#do-you-need-a-laptop

Health FAQs: https://uh.edu/covid-19/faq/health-wellness-prevention-faqs/

Student Health Center: https://uh.edu/class/english/lcc/current-students/student-health-center/index.php