

**COURSE TITLE/SECTION:** Data Mining (COSC 4335)

January 10, 2015

**TIME:** TT 2:30-4p**FACULTY:** Christoph F. Eick**E-mail:** ceick@uh.edu**OFFICE HOURS:** TU 4-5p TH 11:30a-12:30p**Phone:** 33345 (use e-mail!!)**FAX:** 33335

## **I. Course *Data Mining (COSC 4335)***

### **A. Catalog Description**

Data mining overview, data quality, data preprocessing, OLAP, and statistics on one variable; techniques: classification, regression, clustering, dimensionality reduction, association rules; scoring, post-processing, and data mining case studies.

*Prerequisite:* **COSC 3380** and **MATH 3336** .

Taking COSC 3380 and MATH 3336 concurrently with COSC 4335 is acceptable!

### **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. After defining what knowledge discovery and data mining is, the course will give a basic introduction to data analysis. Next, data mining tasks such classification, clustering, and association analysis will be discussed in detail. Moreover, basic techniques how to preprocess a data set for a data mining task will be discussed and basic visualization techniques and statistical methods will be introduced. 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 a sound foundation on how to analyze data
3. will have sound knowledge of popular classification techniques, such as decision trees, support vector machines and nearest-neighbor approaches.
4. will know the most important association analysis techniques
5. will have detailed knowledge of popular clustering algorithms such as K-means, DBSCAN, and hierarchical clustering
6. will conduct projects in which data mining is applied to real world data sets. They will obtain valuable experience in learning how to interpret data mining results, how to select parameters of data mining tools, and how to make sense out of data.
7. will learn on how to use the currently most popular popular data mining programming environment **R**.
8. will get a basic introduction to R programming, in particular how to provide novel data analysis and data mining capabilities on the top of **R**

## III. Course Content<sup>1</sup>

1. Introduction to Data Mining
2. Data
3. Introduction to R
4. Exploratory Data Analysis—how to Visualize and Compute Basic Statistics for Datasets
5. Using R for Data Analysis and Assignment1
6. Introduction to Similarity Assessment and Clustering
7. Writing Programs in R and Assignment2
8. Introduction to Supervised Learning: Basic Concepts and Decision Trees
9. More on Supervised Learning: Instance-based Learning, Support Vector Machines and Regression
10. R-libraries for Classification and Regression and Assignment3
11. Association Analysis —Mining Rules, Sequences and Graphs
12. Association Rules in R and Assignment4
13. Data Preprocessing
14. Top Ten Data Mining Algorithms and Course Summary

---

<sup>1</sup> As Dr. Eick offers this course the first time, what is exactly taught in the course is subject to revision!

## IV. Course Structure

23 lectures  
2 exams  
4 assignments  
1 student presentation

## V. Textbooks

### Recommended Text:

P.-N. Tang, M. Steinback, and V. Kumar *Introduction to Data Mining*, Addison Wesley, 2006.

### Other Text Useful for the Course

Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* Morgan Kaufman Publishers, third edition, 2012.

## VI. Course Requirements

There will be 2 (maybe 3) exams in Spring 2015. There will be 4 assignments:  
Assignment 1: Data Analysis and Using R for Statistical Analysis and Visualization  
Assignment 2: Cluster Analysis and Writing New Functions in R  
Assignment 3: Learning Classification Models and Making Sense of Data  
Assignment 4: Likely, Association Analysis centering on Association Rules

Most likely, Assignments 1 and 3 will be group projects.

## VII. Evaluation and Grading

Assignments: 44-50%  
Exams: 49-54%  
Class Participation: 1%

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 problem-solving 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 3/5 of the course work.

## VIII. Consultation

Instructor: [Dr. Christoph F. Eick](#)  
office hours (573 PGH): TU 11:30a-12:30p and TH 4-5p  
e-mail: ceick@uh.edu  
class meets: TU/TH 2:30-4p

## 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.

**Addendum:** Whenever possible, and in accordance with 504/ADA guidelines, the University of Houston will attempt to provide reasonable academic accommodations to students who request and require them. Please call 713-743-5400 for more assistance.