ITCS6162/ITCS8162

Knowledge Discovery in Databases - KDD

Prerequisites: ITCS6160, full graduate standing or content of the department.
Textbook: "Introduction to Data Mining", by Pang-Ning Tan, Michael Steinbauch, Vipin Kumar, Addison Wesley.

Course Syllabus

- Data preprocessing
- Association and classification rules
- Decision trees
- Granular computing
- Data discretization
- Data reduction (preserving knowledge hidden in data)
- Mining incomplete and unbalanced data
- Data actionability (action rules and meta-actions)
- Null value imputation and Chase
- Evaluation methods (confusion matrix, ...)
- Scalability
- Cluster analysis
- Text analytics, folksonomy
- Mining temporal data
- Applications in health (Clinical pathways analysis, risk prediction, ...)
- Business recommender systems (NPS score, customer churn, ...)

Lectures

😊 Association Rules

😊 Association Rules (Video by L. Powell)

😊 Classification Trees

😊 Classification Trees (Video by L. Powell)

😊 LERS/ERID
Granular Computing
Reducts and Discretization
Reducts (Video by L. Powell)
Discretization (Video by L. Powell)
Mining Incomplete Data
Action Rules and Meta-Actions
Chase Algorithms
Sample Problems (Midterm Exam)
Midterm Exam
Solutions to Midterm Exam
Action Rules Extraction Using Action Reducts
Clustering Methods
TV Trees
Clustering (Video)
Clustering - Sample problems
Evaluation Methods
Chase II Algorithm (for incomplete datasets)
Collaborative Query Processing
Sample Problems
Data Sanitization
Example-Data Sanitization
Temporal DB Mining
Business Analytics
Sample Problems I

Sample Problems II with Solutions

Sample Problems III

Sample Problems (Final Exam)

Project

Project and LISp-Miner

You should submit the project report and the dataset you created by email to: Yuehua Duan at [yduan2@uncc.edu] and Aileen Benedict at [abenedi3@uncc.edu].
Deadline to submit: April 30 (Thursday), 2020

Midterm: March 12
Final (WebEx): May 7 (Thursday), 8:00-11:00am
Points: 30 points Test, 30 points Final, 40 points Project

Grades: A [90-100], B [80-89], C [65-79]. Final grades B, C can be replaced by Pass grade.

Class Location: Woodward 140
Meeting Time: Thursday, 8:30-11:15am

Instructor: Zbigniew W. Ras

Office:

Location: Woodward Hall 430C
Telephone: 704-687-8574
Office Hours (Woodward 430C): Thursday: 11:30am-1:00pm
e-mail: ras@uncc.edu

GTA: Yuehua Duan

Office:
Location: KDD Lab. (Woodward Hall 402)
Telephone: 704-687-8546
Office Hours (Woodward 402): Tuesday, Thursday: 1pm-3pm
e-mail: yduan2@uncc.edu

- **Lisp Miner** (by Jan Rauch)
- **Lisp Miner Manual** (by Jan Rauch's Student)
- **Rough Set Exploration System (RSES)**
- **Bratko's ORANGE**
- **Random Forests**
- **WEKA**
- **More software for data mining**
- **Repository of large datasets**
- **LERS vs ERID**
- **Extracting Rules from Incomplete Table**
- **Lance & Williams Distance**
- **Sample Problems for Midterm Exam**

July 2, Thursday, 2020