COMP SCI 5402: Introduction to Data Mining

CS Building 202, TuTh 9:30AM - 10:45AM, Fall 2018

Instructor: Dr. Yanjie Fu, Assistant Professor of Computer Science
E-mail: fuyan@mst.edu
Phone: (573)341-4991
Blackboard: CS5402@Canvas
Office: 313, Computer Science Building

Text Book: "Introduction to Data Mining", by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley, ISBN: 0-321-32136-7, 2005.

Recommended Readings:

- "Mining of Massive Datasets 2nd Edition", by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, ISBN: 1-107-07723-0, 2014.
- "Pattern Recognition and Machine Learning", by Christopher M. Bishop, Springer, ISBN: 0-387-31073-8

Course Description and Objectives:

(1) to introduce data mining tasks including regression, classification, clustering, ranking, and outlier detection, and (2) to discuss emerging data mining topics such as recommender systems, topic modeling, representation learning, trajectory computing.

Programming Languages:

R and Python are two major languages for data science. To learn Python, check out: https://www.programiz.com/python-programming/first-program To learn R, check out: https://www.statmethods.net.You are suggested to use R or/and Python to complete the homework assignments.

Grading Policy:

3-6 homework assignments:	40%
Exam	20%
Attendance and participation in discussion:	10%
Project/Presentation/Paper	30%

All grading will be determined based on a scale of: A: [90-100%], B: [80-90%), C: [70-80%), D: [60- 70%), F: [0-60%). Final grades in the course may be curved at the instructor's discretion.

Schedule (non-restrictive):

1. Introduction

   - What is data mining?
   - Introduction to Data Mining Tasks (Classification, Clustering, Association Analysis, Anomaly Detection)
2. Data and Data Exploration
   - Understanding of Data
   - Data Cleaning and Preprocessing
   - Feature Engineering
3. Classification
   - Rule-based Classifiers
   - Decision Trees
   - Nearest Neighbor Based Classifiers
   - Naïve Bayesian Classifiers
   - Logistic and Ridge Classifiers
   - Classification Model Selection and Evaluation
4. Clustering
   - Types of Clusters and Clustering
   - K-Means Clustering
   - Hierarchical Clustering
   - Density-based Clustering
   - Fuzzy Clustering
   - GMM
   - Cluster Validation
5. Recommender Systems
   - What are Recommender Systems?
   - Content-based Recommender Systems
   - User-based Collaborative Filtering
   - Item-based Collaborative Filtering
   - Matrix Factorization
   - Evaluation of Recommender Systems
6. Ranking
   - What is Learning To Rank?
   - Point-wise Learning To Rank
   - Pairwise-wise Learning To Rank
   - Listwise-wise Learning To Rank
   - Evaluation of Ranking
7. Topic Modeling in Text Mining
8. Anomaly Detection
   - Statistical-based Methods
   - Density-based Methods
   - Clustering-based Methods