ITMD 522 SYLLABUS

ITMD 522 Data Mining and Machine Learning

Hours: 3 credit hours / 45 contact hours

Instructor: Yong Zheng

Textbook, title, author, and year:
(all are optional but are strongly recommended)

Machine Learning: *The Art and Science of Algorithms that Make Sense of Data*, Peter Flach. 2012

Python Data Science Handbook: Essential Tools for Working with Data 1st Edition, Jake VanderPlas. 2016

Specific course information

a. **Catalog description:** Data mining is a useful tool to uncover patterns and underlying relationships in large data by using data analytics and knowledge discovery techniques. Machine learning algorithms additionally learn from the data and make predictions or decisions by different optimization methods. This course is a graduate level survey of concepts, principles and techniques related to data mining and machine learning. Students will be familiar with data preprocessing skills and the popular data mining and machine learning techniques, including the supervised learning (regressions and classification) and unsupervised learning(clustering and association rules analysis), as well as semi-supervised learning and ensemble learning. Students will also learn the related applications, including text mining/NLP, Web mining, information retrieval and recommender systems. Students will learn R and Python programming for data mining and machine learning and be able to handle real-world data or applications.
b. **Prerequisites/Corequisites:** ITMD 514 or ITMS 514

Specific goals for the course

a. **Program Educational Outcome:** Deliver optimal technical and policy technology solutions for the problems of business, industry, government, non-profit organizations, and individuals in each student's particular area of focus.
b. **Course Outcomes:** At the completion of the course, each student will have the capability to deal with real-world data mining problems by using standard knowledge discovery in databases (KDD) process. More specifically, students are able to perform data selection, data preprocessing and data mining on real-world data sets. Students will have demonstrated their knowledge and skills in classical data mining techniques, including classification, clustering, association rule analysis, etc. Students are able to use related tools (R and Python) to apply these data mining techniques. They are expected to be familiar with popular applications, including Web mining, information retrieval and recommender systems.
c. **Course Student Outcomes:**
Upon successful completion of the course the student should be able to do the following:
    - Describe what data mining is and why it is so useful
        - Discuss real-world data mining techniques and applications
    - Recall and distinguish between related terms: artificial intelligence, data mining, machine learning, etc
    - Explain the standard Knowledge Discovery in Databases (KDD) process
    - Perform popular data preprocessing: data selection, data cleaning, data transformation, etc
    - Explain the difference between classification and clustering
    - Perform and evaluate different classification algorithms over real-world data sets
    - Perform and evaluate different clustering algorithms over real-world data sets
    - Explain how association rule analysis works and how it is used to assist business intelligence
    - Perform and evaluate associate rule mining to discover useful patterns over real-world data sets
    - Describe the principles of and build related applications, including Web Mining applications, information retrievals, recommender systems, etc
    - Explain how regression analysis works and how useful it is
    - Use data mining tools (e.g., R and Python) to perform data mining tasks
    - Describe basic evaluation protocols, challenges (e.g., overfitting, imbalance issues) and corresponding solutions in different data mining tasks
    - Explain the pros and cons of each commonly used technique for specific data mining tasks. For example, among different classification techniques, what are their pros and cons.
        - Determine which techniques should be used in various situations for specific problems.
    - Identify problems and determine which of multiple appropriate techniques are optimal to solve the problems.

Topics to be covered

a. Intro: Data Science & KDD: Preprocessing
b. Classification by KNN & Naïve Bayes
c. Application: Information Retrieval and Text Classifications
d. Classification by Decision Trees and SVM
e. Ensemble Methods a Multi-Label Classifications

f. Clustering Techniques
g. Association Rules and Web Mining
h. Outlier Detection, Feature Selection and Reductions
i. R for Data Science
j. R for Data Science with Practice
k. Recommender Systems and Collaborative Filtering
l. Traditional Recommender Systems
m. Python for Data Science
n. Python for Data Science with Practice
o. Advanced Topics in Machine Learning
p. Exams or Project Presentations
q. Applications: Recommender Systems
r. Exams or Project Presentations