The George Washington University  
Department of Statistics  
Introduction to Data Mining  
Stat 8289 (FALL 2012)

Instructor: Rida E. Moustafa  
E-mail: Shalash@gwu.edu

Office hours: By appointment  
Class time: Thursday 6:10-8:40, ROME 459

Course Description:  
Data mining is a multidisciplinary subject at the intersection of statistics, machine learning, visualization and computer science. This course is designed to introduce you to data mining techniques (automatic and semi-automatic) including predictive, descriptive and visualization modeling and their effective use in discovering interesting hidden patterns in large volume of data generated by businesses, science, web, and other sources. Focus is on the main process of data mining such as data preparation, classification, clustering, association analysis, and pattern evaluation. Some additional topics from the literature may be covered.

Learning Outcome: Due to the multidisciplinary aspects of data mining, the high market demand, and the complex nature of the development of data mining systems, data mining is in the forefront of information technology. The requirements of this profession including but not limited to: the ability to understand the data and the problem at hand; the ability to choose appropriate methods and techniques for analysis and use them effectively. The course objectives are:

- Synthesize the data mining fundamental concepts and techniques from multiple perspectives.
- Develop skills and apply data mining tools for solving practical problems.
- Advance relevant programming skills.
- Gain experience and develop research skills by reading the data mining literature.

Topics to be covered

1. Introduction to data mining (definition, motivation, applications, tasks)
2. Data Preparation and Data Exploration (data cleaning, integration, transformation, reduction, summarization and visualization)
4. Classification: Basic Concepts, Decision Trees, and Model Evaluation (evaluation steps, cross-validation, evaluation of numeric prediction);
5. Classification: Alternative Techniques: naive Bayes (NB) models and support vector machines (SVM);
6. Cluster Analysis: Basic Concepts, Algorithms and Visualization (partitioning methods, hierarchical clustering, density-based clustering, grid-based clustering, model-based clustering, outlier detection);
7. Advanced topics and applications.

Textbook(s):
Required: Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques,” Third Edition, 2012, SBN:9780123814791
Recommended:
Software: R/ Rattle /WEKA/ Matlab/ SAS. 

Brief description of the type of instruction and learning activities: To meet the course objectives, the course provides three types of learning experiences:

- Class lectures and readings. The lectures cover a broad range of data mining topics focusing on the concepts and techniques. The textbook provides a more detailed on these topics. The next important source of information is the lecture notes.
- Assignments: The assignments are to help students explore the ways of implementing data mining concepts and techniques on real data. Students are asked to use their preferred tool(s) and apply the skills learned in this course and programming skills in a programming language, e.g. Java, C, or C++ / statistical packages (SAS/R)/Matlab/ data mining tools (WEKA).
- Project and research exercise. Students are asked to develop a project based on the literature research in the data-mining subject and report the findings in a well-structured and well-written paper. Instructor will provide templates.

Grading: Grade is assigned as follows:

- Assignments and in Class exercise 20%
- Midterm 15%
- Final 25%
- Project 40%

Assignments: Practical exercises include: a programming project, in which students design, implement, test and validate data mining techniques; three programming assignments using data mining tools; a term paper on a topic selected in consultation with the instructor; and a weekly exercise/ in-class quiz. All graded work will usually be returned and discussed one week after due date. Late submissions will not be accepted. You are expected to work individually on each problem set.

Prerequisite: Background in statistical data analysis, multivariate statistics.
R/SAS/WEKA software will be used. You are expected to be familiar with the R/SAS software. R software is free and can be downloaded from CRAN website. GW labs provide access to SAS and have a site license for SAS. To obtain a copy for your PC contact the Advanced Technology Lab in the basement of Gelman library. See [http://citl.gwu.edu/pages/sas.html](http://citl.gwu.edu/pages/sas.html)
WEKA is an open source data mining tool and can be downloaded from: [http://www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/)

ACADEMIC INTEGRITY
I personally support the GW Code of Academic Integrity. It states: “Academic dishonesty is defined as cheating of any kind, including misrepresenting one’s own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information.” For the remainder of the code, see: [http://www.gwu.edu/~ntegrity/code.html](http://www.gwu.edu/~ntegrity/code.html)

SUPPORT FOR STUDENTS OUTSIDE THE CLASSROOM
Disability Support Services (DSS). Contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242. For additional information please refer to: [http://gwired.gwu.edu/dss/](http://gwired.gwu.edu/dss/)
The University Counseling Center (UCC) (202-994-5300) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. See [http://gwired.gwu.edu/counsel/CounselingServices/AcademicSupportServices](http://gwired.gwu.edu/counsel/CounselingServices/AcademicSupportServices)
SECURITY
In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.