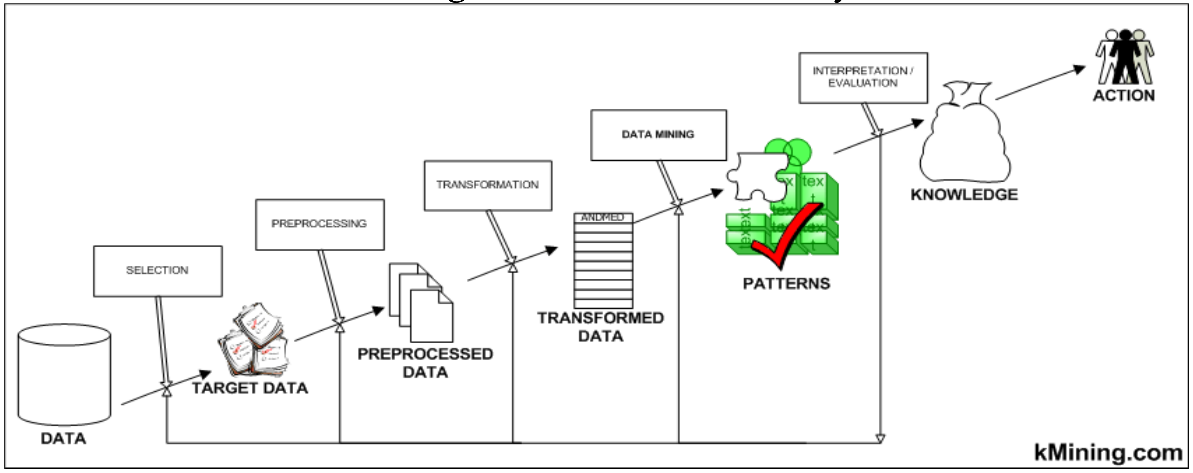
95-791 Data Mining

H. John Heinz III College
Carnegie Mellon University



Instructor: Associate Professor. Murli Viswanathan
Phone: 81109926
E-mail: mkrishna@cmu.edu

Prerequisites: 95-796 “Statistics for IT Managers”

Meeting Times:
Lecture: See class schedule

Class Web Site: Shared folder and **CANVAS**

Required Textbooks (e-book version is fine):

- Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition (https://www.cs.waikato.ac.nz/ml/weka/book.html) by Ian H. Witten, Eibe Frank, Mark A. Hall.

Other Useful Reference books

- Machine Learning with R - Second Edition Paperback – July 31, 2015, by Brett Lantz, Packt Publishing
- An Introduction to Statistical Learning: with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
- Mitchell: Machine Learning, McGraw-Hill, 1997
- Han and Kamber: Data Mining: Concepts and Techniques. Morgan Kaufmann 2000
- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron

Recommended Readings:

- **See Canvas**

Software:

- **WEKA (https://www.cs.waikato.ac.nz/ml/weka/index.html)**
- **TABLEAU (see Canvas for instructions)**

Note that assignments can be done in WEKA or Python

Course Rationale

Public and private enterprises are drowning in data. Expert projections suggest a 4,300% increase in annual data production that will create 35 zettabytes by 2020. In spite of this explosion in data collection we are poor in knowledge. *Data mining refers to the process of analyzing data from commercial and scientific databases to discover previously unknown and hidden knowledge which is predictive.* This knowledge could represent commercial trends or scientific discoveries. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud.

Did you know that Uber uses a data mining program (code-named Greyball) that tries to predict which new Uber drivers are secret cops, regulators or investigators who could make trouble for the company, deployed in "Boston, Paris and Las Vegas, and in countries like Australia, China and South Korea" where the company is fighting with the authorities.

“Greyball used a variety of techniques to identify Uber's potential adversaries, from geofencing the region around public buildings, blocking riders whose credit cards were issued by police credit-unions, and surveilling the app's users by hijacking their own phones, trying to find users who repeatedly opened and closed the app (as a means of identifying whether there were Uber cars around). The company also sent undercover agents to electronics discount stores to record the phone numbers assigned to the cheapest burner phones, on the theory that cash-strapped police departments would look to save money wherever possible.
All of this was combined with data-mining using social networks, through which Uber counterintelligence employees would look up suspicious new users on social media to see if they were linked to law enforcement. People tagged as adversaries of the companies were denied rides, and drivers who accidentally picked them up were ordered to cease the ride and kick them out.”[1]

[1] Extracted from https://boingboing.net/2017/03/04/sounds-legit.html

Course Objectives

This course is designed to give students a solid grounding in the methodologies, technologies and algorithms employed in the data mining field. The emphasis is on understanding the application of a wide range of modern machine learning techniques to specific data analysis scenarios rather than on mastering the theoretical underpinnings of the techniques. The course covers methods that are aimed at prediction, forecasting, classification, and clustering. It also introduces several cutting edge and interactive visualization and data mining tools and learning well-known data mining process methodologies.

More specifically, the goals of this course are:

- Understanding the nature of real-world data and the issues faced in developing intelligent learning systems;
- Supervised and Unsupervised learning (classification/regression);
- Algorithms for data mining and tools including TABLEAU, WEKA, and R.
- To introduce data mining techniques and the CRISM-DM standard
- To introduce exploratory analytics and visualization
- To understand the limitations of various techniques
- To decide when to use which technique
- To work with data mining tools and learn how to analyze data

Individual Project

Students will research and develop a solution for a Data Challenge following the CRISM-DM methodology.

Grading

Students are expected to attend classes and participate. There will be several tutorials during labs, assignments, and a data mining project. The final exam will be closed book (open notes) and based on the theory covered in the lectures. The project will be based on formulating a real-world data mining problem and applying data mining tools.

Lab Tutorial Participation	10%
Group Project	20%
Assignments	30%
Final Exam	40%
	100%

The letter grades will be assigned based on the following percentage points earned from the four components of the class (as defined above):

97% – 100%: A+
93% – 97%: A
89% – 93%: A –
85% – 89%: B+
81% – 85%: B
77% – 81%: B –
73% – 77%: C+
69% – 73%: C
65% – 69%: C –
Below 65%: R

Academic Integrity

The Heinz School takes very seriously its mission to produce graduates who are committed to ethical behavior in all phases of their professional lives. In this regard, the school views any cheating and plagiarism as serious offences. You are required to review the material on academic integrity presented in the master's program handbook and to monitor your own actions carefully to prevent even the appearance of violations of academic integrity guidelines. Any violations of academic integrity in this class will have the following consequences:

a) at the minimum, no credit for assignment in question; and
b) in more serious offences, failing the class.

Schedule (subject to minor revision)

	TOPICS	READINGS	ASSIGNMENT
Lecture 1	• Introduction to Data Mining • SEMMA/CRISP DM • Data Preparation & visualization for Knowledge Discovery	Chapter 1,10	TABLEAU tutorial
		http://www.crisp-dm.org/Process/index.htm	Project information Assignment 1: Data Wrangling & Visualization
Lecture 2 *Assignment 1 due*	• Inputs: Concepts, instances, attributes • Output Knowledge Representation • Classification - Basic methods	Chapter 2 Chapter 3,7	Assignment 2: Data Mining - classification Learning to use WEKA Tutorials
Lecture 3	• Classification: Decision Trees and Rules • Evaluation	Chapter 4,6	
Lecture 4 *Assignment 2 due*	• Association Rules and Market Basket Analysis	Chapter 4,6	Assignment 3: Association Learning
Lecture 5 *Assignment 3 due*	• Unsupervised Learning • Clustering Techniques	Chapter 6	Assignment 4: Clustering Data
Lecture 6 *Assignment 4 due*	• Evaluation and Credibility • Advanced Techniques	Chapter 5,9	
Exam Week *Project due*	• Final Exam	Chapters 1,2,3,4,5,6,9,10	

NOTES

- Chapters are from the required textbook “Data mining: practical machine learning tools and techniques.—3rd Ed.” by Ian H. Witten, Frank Eibe, Mark A. Hall.
- Class notes and all other material will be posted on **CANVAS** prior to the lecture.
- Important announcements will be sent by email through CANVAS.