Data Mining (95-791 Z4)

Syllabus

Mini 4, Spring 2018

This syllabus is adapted from Dr. Dubrawski’s 95-791 Data Mining Syllabus

Lecture Instructor:
Dr. Artur Dubrawski awd@cs.cmu.edu

Distance Learning Facilitator:
Karen (Lujie) Chen karenchen@cmu.edu

Teaching Assistant:
TBD

Prerequisites

95-796 “Statistics for IT Managers” or instructor’s permission based on the student’s knowledge of fundamentals of probability and statistics. Previous experience with data analysis will be considered a plus, although it is not absolutely necessary.

Course Motivation

Data mining – intelligent analysis of information stored in data sets – has gained a substantial interest among practitioners in a variety of fields and industries. Nowadays, almost every organization collects data, which can be analyzed in order to support making better decisions, improving policies, discovering computer network intrusion patterns, designing new drugs, detecting credit fraud, making accurate medical diagnoses, predicting imminent occurrences of important events, monitoring and evaluation of reliability to preempt failures of complex systems, etc.

About the Instructor

Artur Dubrawski is a scientist and a practitioner. He has been researching machine intelligence and its applications for twenty five years. In the past, he has been affiliated with an advanced data mining firm, Schenley Park Research, and served as Chief Technology Officer at Aethon, a local high-tech company making autonomous delivery robots. Currently Dr. Dubrawski is a faculty at the CMU Robotics Institute, where he directs the Auton Lab: a data mining and machine learning research group. Auton Lab’s work has yielded multiple deployments of analytic solutions and software in various government and industrial applications.
About the course facilitator

Karen Chen is a PhD student in the information system program of Heinz College, she is also associated with Auton Lab under supervision of Dr. Dubrawski. Her research interest is in big data analytics, machine learning and data mining application, in particular, the modeling of temporal dynamics of real time sensor data with application to health care and education. Some of her work involved analyzing physiological signals from continuously monitored patients as well as psychological signals of emotion states from facial expression analysis. Before her PhD career, she worked as a research staff with the Auton Lab for about 10 years, working on a variety of data mining and analytics projects in areas of public health, food safety, health insurance and fuel efficiency. She holds MISM degree and M.S. in statistics, both from Carnegie Mellon University and B. Eng degree in business and computer science from Shanghai Jiaotong University in China.

Course Objectives

This course will provide participants with an understanding of fundamental data mining methodologies and with the ability to formulate and solve problems with them. Particular attention will be paid to practical, efficient and statistically sound techniques, capable of providing not only the requested discoveries, but also estimates of their utility. The lectures will be complemented with hands-on experience with data mining software, primarily R, to allow development of basic execution skills.

The scope of the course will cover the following groups of topics.

Foundations. How to make data mining practical? (approximately 40% of class time)

Learning from data: why, what and how?

Fundamental tasks, issues and paradigms of learning models from data.

Real world data is noisy and uncertain. How much can we trust the results of our analyses?

Model selection

Reduction of dimensionality and data engineering

Measures of association between data attributes: information theoretic, correlational

Pragmatic methodologies for mining data (approximately 60% of class time)
Predictive analytics: classification and regression

Cost-sensitive model selection using ROC approach

Compression of data and models for improved reliability, understandability, and tractability of large sets of highly dimensional data

Association rule learning and decision list learning, decision trees

Introduction to density estimation, anomaly detection, and clustering

Overview of mining complex types of data

Illustrative examples of real-world applications

**Reading Material**

Unfortunately, the ideal textbook for this course does not exist. Instead, we will use a selection of readings excerpted from a variety of sources. These readings are intended to complement the material presented in class. Selected issues covered by the required readings will become topics of graded assignments and final examination.

All required material will be distributed electronically through course site, or pointers to the resources available on the internet for free download. Note that many of the readings are protected under copyright law. In order to use them in this course it was necessary to purchase official permissions from the copyright holders. Each enrolled student could have their HUB account charged with an equal share of the copyright fees. Although the exact amount of the individual share is not known at the moment of writing this document, it is estimated to not exceed $30.00. Please note that it is illegal to distribute copies of the copyrighted materials without obtaining permissions from their legal owners.

Interested students are welcome to go beyond the scope of the required readings. In particular, the following books are recommended - but not required - listed in no particular order:
3. Hastie, Tibshirani, Friedman: The Elements of Statistical Learning, Springer 2001

**Software and Hands-on Exercises**

We will primarily rely on R free software to demonstrate and operationalize concepts presented during lectures. Students are expected to download and install
the software, as well as learn basic usage skills on their own using tutorials available online. Appropriate resources will be recommended during the first lecture and/or recitations session.

Recitations will review concepts taught in lectures and connect them to homework problems through examples. Recitation sessions when software tools are introduced will provide hands-on-experience opportunity: the students will be asked to follow the presenter using their laptops and they will work on assigned exercises while in session.

**Assignments and Deadlines**

All assignments will be distributed electronically through the piazza. All reports (including homework) must be submitted electronically through email (TBD). There are two kinds of deadlines for each homework: The soft deadline and hard deadline, each with one week apart. You are encouraged to submit homework by soft deadline, in which case you will have 5% bonus point of your actual homework marks (for example, if you submit homework x before the soft deadline, and your homework mark is 90 out of 100, then your final mark for this homework is 90*1.05). You may choose to submit according to the hard deadline schedule without bonus nor penalty. Late homework will be accepted until 24 hours past the hard deadline, but it will be subject to an automatic 50% grade reduction.

**Grading**

Grades will be based upon the results of four homework assignments, one analytical project.

The analytical projects will be conducted in small groups of students. Each team will analyze specific real-world data. The project will be graded based on bi-weekly progress report (TBD), a report (TBD) and a recording of an oral presentation of the results (TBD).

The final grade for this course will is composed of following:
1. Homework (4 times 15%) 60%
2. Analytical project (in teams) 40%

**Academic Integrity**

Students are expected to strictly follow Carnegie Mellon University rules of academic integrity in this course. This means homework are to be the work of the individual student using only permitted material and without any cooperation of other students or third parties. It also means that usage of work by others is only permitted in the form of quotations and any such quotation must be distinctively marked to enable identification of the student's own work and own ideas. All external sources used must be properly cited, including author name(s), publication title, year of publication, and a complete reference needed for retrieval. Regarding the group projects, the work should be the work of only the group members. In all
their work students should not in any way rely on solutions to problems distributed in prior years or on the work of prior students or other current students. Violations will be penalized to the full extent mandated by the CMU policies. There will be no exceptions.

**Health and Wellness**

Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress. All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful. If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.