COURSE SYLLABUS
IST 565 Data Mining

Instructor: Bei Yu
Office: Hinds 320
Office Hours: By appointment
Phone: 315-443-3614
Email: byu@syr.edu

Prerequisite: IST 687. Exceptions maybe given to students who have acquired skills equivalent to what is taught in IST687

Audience: Undergraduate and Graduate Students

Description: Introduction to data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field. Hands-on experience with open-source software packages.

Additional Course Description:
This course will introduce popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be discussed and will be related to the issues in applying data mining to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The focus of this course is in understanding of data and how to formulate data mining tasks in order to solve problems using the data.

The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, students will develop skills that can be applied to business, science or other organizational problems.

The format of the class meetings will be a combined lecture and lab format, with lectures and class discussions to cover material and lab time to investigate small examples for the topic of the week. There will be weekly readings based on the textbook and on other materials, which will be posted on-line.

Credits: 3

Learning Objectives:
After taking this course, the students will be able to:

- Document, analyze, and translate data mining needs into technical designs and solutions.
- Apply data mining concepts, algorithms, and evaluation methods to real-world problems.
- Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.

Bibliography/ Texts / Supplies – Required:

- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) Introduction to Data Mining. (Free sample chapters available at authors' website http://www-users.cs.umn.edu/~kumar/dmbook/index.php)

Bibliography/ Texts / Supplies– Recommended

- Tom Mitchell (1997) Machine Learning. (http://www.cs.cmu.edu/~tom/mlbook.html)
- Brett Lantz (2015) Machine Learning with R (second edition).

Note to students: Given the diversified background of data science students, one textbook might not fit everyone. If you like rigorous algorithm presentation, I would recommend Mitchell's classic book on Machine Learning. If you like more lay-person explanation of machine learning, see if you like Lantz's book better.
The current required textbook is a balance between the two views. I will put a copy of this book to the Bird Library reserve room. You can check it out and read for up to two hours every time.

Tips for success in this class: Curiosity, critical thinking, math, and programming.

- Curiosity: Curious about the data, pay attention to the data details. Don't treat a data set as a blackbox. Don't treat an algorithm as a blackbox. Try see through them.
- Critical thinking: Data mining is essentially research. You will learn and practice methods to discover patterns, and also evaluate whether and why the discovered patterns are true and useful.
- Math: You will need some math knowledge, such as algebra and probability, to understand how the data mining algorithms work.
- Programming: Although GUI tools like Weka and Rapid Miner would allow users with no programming skills to play with data sets, data sets are rarely immediately ready for analysis in these tools. The results from off-the-shelf tools may need additional transformation to see patterns. Programming skills would help you pre- and post-processing the data. Programming would also help you gain more convenient control over algorithm tuning in your scripts.

Software

- Weka, R

Note to students: We will mainly use R but keep Weka as a backup tool for students who do not have enough R skills.

Requirements:

Your final grade is determined by your performance on the items in the table below. An overview of each item is provided in the remainder of this section.

Assessment Item	Weight %
Class exercises	15

Homework assignments	60
Project report	25
Total	100

- **Class exercises:** Students are required to participate in class discussions and exercises. These exercises are designed to encourage students to practice their newly learned knowledge, and thus the grading is based on participation only, not performance. All participations in the exercise forums will be tallied every week. If there is *x* number of exercises throughout the semester, and a student finishes *y* number of exercises in total, the student's grade is *y*/*x**15.
- **Homework assignments:** Assignments must be professionally prepared and submitted electronically to the LMS. All assignments should be submitted in Word files named as "*HW_Num_Lastname_Firstname.doc(x)*", e.g. "HW_1_Smith_John.doc". No PDF please.
- **Course project:** The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student.
  - Checkpoint 1: project idea proposal and presentation: Your idea proposal should include an overview of the data mining problem, the data set you will use and its availability, and your proposed data mining approach.
  - Checkpoint 2: project progress presentation: Show preliminary results and major challenges.
  - Checkpoint 3: Final project report: The final project report should describe the data mining problem, its significance and broader impact, the data mining approaches, results, and interpretation of the discovered patterns.

Grading:

For this class, an "A" would mean the student has the capability to independently solve a simple data mining task. Below is a common formula for number-to-letter grade conversion.

Grade	Points	Grade	Points	Grade	Points	Grade	Points
		B+	87-89	C+	77-79	F	0-69
A	93-100	B	83-86	C	73-76		
A-	90-92	B-	80-82	C-	70-72		

Grades of D and D- may not be assigned to graduate students.

Course Specific Policies on attendance, late work, make up work, examinations if outside normal class time, etc.:

- **Registration**: Students must register prior to the first class or may be restricted from registering. If you are registered but not present at the first class, you run the risk of being administratively deregistered from this course so that your seat can be given to a student on the wait list.

- **Late Policy for Assignments**: To ensure fast return, all assignments should be submitted on time. One-hour grace period is given to accommodate any incidents around deadline. Late policy will be enforced starting from the second hour. You are free to discuss the assignments with your classmates, but you must write up the report all by yourself. Plagiarism cases will be reported to the university.
- **Communications:** This course will use the 2U LMS system as the main communication platforms. Students are required to check their LMS accounts on a regular basis.

Academic Integrity Policy

Syracuse University's Academic Integrity Policy reflects the high value that we, as a university community, place on honesty in academic work. The policy defines our expectations for academic honesty and holds students accountable for the integrity of all work they submit. Students should understand that it is their responsibility to learn about course-specific expectations, as well as about university-wide academic integrity expectations. The policy governs appropriate citation and use of sources, the integrity of work submitted in exams and assignments, and the veracity of signatures on attendance sheets and other verification of participation in class activities. The policy also prohibits students from submitting the same work in more than one class without receiving written authorization in advance from both instructors. Under the policy, students found in violation are subject to grade sanctions determined by the course instructor and non-grade sanctions determined by the School or College where the course is offered as described in the Violation and Sanction Classification Rubric. SU students are required to read an online summary of the University's academic integrity expectations and provide an electronic signature agreeing to abide by them twice a year during pre-term check-in on MySlice. For more information about the policy, see http://academicintegrity.syr.edu. Respect Intellectual Property Rights and cite all sources in your work. Any valid citation style may be used. The following link may be used for further information regarding appropriate citation styles: *http://researchguides.library.syr.edu/citation*.

Disability-Related Accommodations

Syracuse University values diversity and inclusion; we are committed to a climate of mutual respect and full participation. If you believe that you need accommodations for a disability, please contact the Office of Disability Services (ODS), disabilityservices.syr.edu, located at 804 University Avenue, room 309, or call 315.443.4498 for an appointment to discuss your needs and the process for requesting accommodations. ODS is responsible for coordinating disability-related accommodations and will issue "Accommodation Authorization Letters" to students as appropriate. Since accommodations may require early planning and generally are not provided retroactively, please contact ODS as soon as possible. Our goal at the iSchool is to create learning environments that are useable, equitable, inclusive and welcoming. If there are aspects of the instruction or design of this course that result in barriers to your inclusion or accurate assessment or achievement, please meet with me to discuss additional strategies beyond official accommodations that may be helpful to your success.

Religious Observances Notification and Policy

SU's religious observances policy, found at supolicies.syr.edu/emp_ben/religious_observance.htm , recognizes the diversity of faiths represented in the campus community and protects the rights of students, faculty, and staff to observe religious holy days according to their tradition. Under the policy,

students should have an opportunity to make up any examination, study, or work requirements that may be missed due to a religious observance provided they notify their instructors no later than the end of the second week of classes through an online notification form in MySlice listed under **Student Services/Enrollment/My Religious Observances/Add a Notification**.

Student Academic Work Policy

Student work prepared for University courses in any media may be used for educational purposes, if the course syllabus makes clear that such use may occur. You grant permission to have your work used in this manner by registering for, and by continuing to be enrolled in, courses where such use of student work is announced in the course syllabus. I intend to use academic work that you complete this semester in subsequent semesters for educational purposes. Before using your work for that purpose, I will either get your written permission or render the work anonymous by removing all your personal identification.

Course evaluations:

There will be an end of course evaluation for you to complete this semester, described below. This evaluation will be conducted online and is entirely anonymous. You will receive a notification from the Syracuse University Office of Institutional Research & Assessment (OIRA) department in your email account with the evaluation website link and your passcode.

End of semester evaluation will be available for completion approximately week 14. This evaluation is slightly longer and it is used to gauge the instructor performance and make adjustments to the course to ensure it meets our student needs.

We faculty work hard to do the best possible job when preparing and delivering courses for our students. Please understand that not only does the school use the course evaluations to make decisions about the curriculum in order to improve where necessary, but they also use them to make decisions about faculty members. Please take the time and fill out this evaluation as your feedback and support of this assessment effort is very much appreciated.

Course Schedule:

Week	Topic	Textbook Readings	Submission items
0		Syllabus	Student self-intro video
1	Introduction to Data Mining	Ch.1	HW1
2	Data Exploration	Ch. 2-3	HW2
3	Association Rules	Ch. 6.1-6.3	HW3
4	Clustering	Ch. 8.1-8.3	HW4
5	Classification algorithm: decision tree	Ch. 4.1-4.3	HW5
6	Model Evaluation	Ch. 4.4-4.6	Project checkpoint 1
7	Classification algorithm: naïve Bayes	Ch. 5.3	HW6
8	Classification algorithm: kNN, SVMs, random forest	Ch. 5.2, 5.5	HW7, project checkpoint 2
9	Text mining		HW8
10	Review on classification applications		Student project presentation
	48 hours after the presentation		Final project report (checkpoint 3)