UWT TCSS 555 Data Mining Course Syllabus
Spring 2014

Instructors: Senjuti Basu Roy

Location: TLB 115

Class sessions:
MW 4:15-6:20
Instructor Office: CP 229
Office Hours: TBD (email: senjutib@uw.edu)

Class mailing list
Class moodle website

Course Overview:

Welcome. The data mining course presents methods and systems for mining varied data and discovering knowledge from data. After detailing a data mining system architecture and tasks, the course examines and compares specific methods in data mining, such as data preparation, classification, clustering, and text mining. Several applications are detailed, and tools to build new applications are provided. The task of knowledge discovery is then outlined as a higher-level goal of data mining. Familiarity with statistics, and database systems; in particular database design is expected.

The primary objectives of the course are:
• Understand algorithms and methods of data mining.
• Develop data mining programs and applications.
• Program using available data mining tools and general-purpose languages.
• Understand analysis, metrics, visualization and navigation of data mining results.
• Learn how to use a few commercial data mining tools (amongst a selection of: RStudio, Matlab, Jung library, Weka).

The outcomes of the course are:
Upon successful completion of the course, students are able to:
• Explain the basic principles of the primary data mining techniques.
• Explain the difference between data mining, data warehousing, machine learning, etc.
• Design mining models and manage databases to enable data mining technologies as part of larger systems.
• Describe issues facing latest trends in data mining.

Prerequisites:

(Under)Graduate Standing in CSS. TCSS 445 or equivalent database systems design experience at work or at internships. You should be proficient in database design and have an understanding of basic database system implementation techniques. In addition to that, basic understanding in probability and statistics is desirables.

We also recommend that all students have prior experience with at least one programming language such as C/C++/Java/C# and have demonstrated ability to work on coding algorithms and data structures. Students with no programming experience or interest should not take this course.
**Texts:** Data Mining: Concepts and Techniques (3rd edition), by Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kaufmann, 2011

Supplementary Text: Prof. Zaki’s text, Massive Data Mining by Jure Leskovec et. al. (available on moodle)

**Additional Readings:** The links to additional readings will be made available on moodle during the term. You may bring your readings to each class session so that you can refer to them during discussion. You will additionally do a number of readings of your own choosing for the project and the accompanying research paper that you will write. Your instructors may assign specific readings to individual students.

**Course Content:**

There will be several aspects to this course:

1. Instructor lectures

2. Individual assignments: Assignments should be done individually and submitted online on moodle by the due date.

3. Class Notes: Everyone takes class notes. You can form an ad-hoc group, meet in-person or collaborate online to discuss and then submit your class notes. Bring hard-copy to class each Monday. The notes should read like a chapter with your explanation of the concepts discussed in class that week.

4. In-Class discussion sessions or in-class labs (every Wednesday) on non-technical reading, data mining algorithms and implementation techniques.

5. Guest lectures may be invited depending on schedule and availability from industry describing latest research trends and implementation issues in commercial systems.

6. Class quiz: Each Wednesday the class will begin with a short quiz based on the reading assignments and class discussion from the previous week. This quiz is to ensure that you are taking the readings seriously and are absorbing the material presented in class.

6. Class project: A class project will be assigned and will be completed individually.

7. Midterm Exam

**Class Schedule: (may change - updated version always on moodle)**

<table>
<thead>
<tr>
<th>Week</th>
<th>topics covered</th>
<th>Readings (text; other)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 1</strong></td>
<td>Introduction and Math Foundations</td>
<td>Ch.1,2; Linear Algebra</td>
</tr>
<tr>
<td><strong>Week 2</strong></td>
<td>Data Preprocessing, Warehousing, OLAP</td>
<td>Ch.3,4</td>
</tr>
<tr>
<td>Week</td>
<td>Topic</td>
<td>Chapter/Sections</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Week 3</td>
<td>Frequent Pattern Mining</td>
<td>Ch. 6</td>
</tr>
<tr>
<td>Week 4</td>
<td>Classification: basic concepts</td>
<td>Ch. 8, 9</td>
</tr>
<tr>
<td>Week 5</td>
<td>Clustering: basic concepts</td>
<td>Ch. 10</td>
</tr>
<tr>
<td>Week 6</td>
<td>Recommendation Systems</td>
<td>Ch.13.3.5; supplementary reading</td>
</tr>
<tr>
<td>Week 7,8</td>
<td>Mining Social Network Graphs</td>
<td>Ch 10, Massive Data Mining, by Jure Leskovec.</td>
</tr>
<tr>
<td>Week 9</td>
<td>Outlier detection</td>
<td>Ch. 12</td>
</tr>
<tr>
<td>Week 10 &amp; 11</td>
<td>Evaluating Data Mining Models</td>
<td>Ch. 31 (Zaki text), IR evaluation techniques;</td>
</tr>
</tbody>
</table>

**Class Attendance and participation:**

Attendance and participation is expected. It will be counted towards your final grade (5%). We will not take attendance each day but since the class size is relatively small, absences will not be hard to miss. Early departures will not be counted as attending. Each student may miss up to one class with no penalty. This policy is not in effect if there is a flu epidemic (see last section).

**Class Etiquette:**

It is sometimes useful to distance ourselves from technology to obtain an environment of quality in-class discussion. Hence, kindly minimize the use of personal laptop computers in the classroom unless appropriate. Please turn off all cell-phones and PDAs unless taking notes. If placed on vibrate mode, please ensure that the ‘buzz’ does not disturb your neighbors during the sessions. Please bring a pen (black or blue) and a writing pad to class to take notes.

Come prepared to the class in the following manner: Read the assigned reading several days prior to the session on that topic.
Guest lectures will be invited. If you have suggestions for guest lecturers you can email the instructor about them. Maintain a sense of decorum and respect for the guest speaker. They are providing a valuable volunteer service by giving us the benefit of their expertise. We can make their visit meaningful for them only if we have read about the topic and come prepared to hear them. We can demonstrate our preparedness by asking meaningful questions that advance their research interests. Hence, be extra careful in designing your questions for these days. Give these papers extra attention.

All assignments are to be submitted to moodle on the date due, unless otherwise specified in class. If a handout is needed, make sure that your name is on each sheet that is handed in and that any hand-in with multiple pages is stapled. You should as a backup also upload your submissions to appropriate slots on Moodle to facilitate grade recording. Failure to submit /uploads to moodle prior to deadline will result in automatic denial of grade for that work. To keep the grading process on time, extensions are not available and should not be requested.

Additional material for writing support: You may choose current significant (science- or technology-related where applicable) news stories from the press to support your writing in the class. Provide full bibliographic information for each story in APA style.

Bibliographic style

Use APA Style for bibliographic references and citations in everything that you write. Make sure to use the special format for electronic references.

Grading Scheme*:

I reserve the right to make small adjustments to grade weights, or to add small assignments as the need arises.

<table>
<thead>
<tr>
<th>Item</th>
<th>Grade (% of final grade)</th>
<th>Due date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework Assignments (total 3)</td>
<td>3 x 7= 21%</td>
<td>Home works would be posted on Fridays. It would be due on Fridays after 2-weeks.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(HW-1 posting date, 12 th April)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(HW-2 posting date, 3rd May)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(HW-3 posting date, 24th May)</td>
</tr>
<tr>
<td>Reading, Class Notes, class participation (every week)</td>
<td>5+5%</td>
<td>Each week</td>
</tr>
<tr>
<td>Reading based Class Quiz (7 total)</td>
<td>7x4= 28%</td>
<td>Each week</td>
</tr>
<tr>
<td>Final Project</td>
<td>25%</td>
<td>June 10 before midnight (Posting Date, May 5)</td>
</tr>
<tr>
<td>Midterm</td>
<td>16%</td>
<td>May 1 (for 2 hours, followed by another 2 hours of regular class activities)</td>
</tr>
</tbody>
</table>

Additionally, there will be 2 extra-credit homework assignments. Each will be of worth 5%.
Grading will not be on a curve. The correspondence between percentage scores and letter grades is given as in the Grading matrix:

<table>
<thead>
<tr>
<th>Grade</th>
<th>GPA</th>
<th>Score</th>
<th>Grade</th>
<th>GPA</th>
<th>Score</th>
<th>Grade</th>
<th>GPA</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.0</td>
<td>98-100</td>
<td>B+</td>
<td>3.4</td>
<td>88-89</td>
<td>C+</td>
<td>2.4</td>
<td>77</td>
</tr>
<tr>
<td>3.9</td>
<td>95-97</td>
<td></td>
<td>3.3</td>
<td>87</td>
<td></td>
<td>2.3</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>A-</td>
<td>3.8</td>
<td>93-94</td>
<td></td>
<td>3.2</td>
<td>86</td>
<td></td>
<td>2.2</td>
<td>75</td>
</tr>
<tr>
<td>3.7</td>
<td>92</td>
<td></td>
<td>B</td>
<td>3.1</td>
<td>85</td>
<td>C</td>
<td>2.0-2.1</td>
<td>70-75</td>
</tr>
<tr>
<td>3.6</td>
<td>91</td>
<td></td>
<td></td>
<td>3.0</td>
<td>83-84</td>
<td>C-</td>
<td>1.8-1.9</td>
<td>68-69</td>
</tr>
<tr>
<td>3.5</td>
<td>90</td>
<td></td>
<td></td>
<td>2.9</td>
<td>82</td>
<td></td>
<td>1.5-1.7</td>
<td>65-67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B-</td>
<td>2.8</td>
<td>81</td>
<td></td>
<td>1.2-1.4</td>
<td>62-64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.7</td>
<td>80</td>
<td>D</td>
<td>0.9-1.1</td>
<td>60-62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.6</td>
<td>79</td>
<td>E</td>
<td>0.0-0.8</td>
<td>&lt;60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.5</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Grading Policy Details:

Attendance/Class Participation- 5 points: complete attendance (1 class absenteeism allowed if you email the instructor ahead of the class and you don't have an assigned duty in class), actively contribute to the sessions, enriching the class experience, help others in learning the concepts and contribute to projects and discussions. 3 points: All the above except for lesser attendance. 1 point: some participation, irregular attendance. 0 point: irregular attendance, no interaction and participation.

Quiz – Each quiz graded based on its questions. No extra credit awarded.

Assignments – Each assignment graded based on its questions and points therein. Complete work mostly qualifies for full grade. Partial credit is dependent on the amount of work completed. The assignments are supposed to enhance your understanding and help improve your hands on skills using data mining tools and languages such as R. These skills are central to success as a data miner in the industry and all assignments use real-world datasets. You should look out and discuss when you discover new things as you try to solve the problems.

General Student Post-conditions:

1. Experience with designing and implementing programs to handle data of significant size (about Million records), using appropriate data mining algorithms, data structures, code organization, and documentation.
2. Familiarity with characteristics of data mining methods wrt worst-case analysis and its influence on choices of data structures, algorithms, and program design.

Characteristics of a below average (C) outgoing student-

1. Able to implement and document a mediocre quality data mining algorithm when given guidance for how to design and implement it.
2. Knowledge of data mining algorithms: how they are defined, how they work.
3. Knowledge of when to employ appropriate statistical measures that have been discussed in the course.
Characteristics of an average (B) outgoing student

All of the above and the following:

1. Able to implement and document an above average quality data mining algorithm when given some guidance for how to design and implement it.
2. Understanding of when to employ what data mining algorithm in situations that are simple variations of those described in the course.
3. Has rudimentary ability to generate and evaluate different optimization alternatives based on the criteria described in the course.
4. Has rudimentary ability to reason about programs performance.

Characteristics of an above average (A) outgoing student

All of the above and the following:

1. Able to implement and document a high quality, high complexity (applicable in the real-world) program when given little guidance for how to design and implement it.
2. Understanding of when to employ appropriate data mining algorithms, even in novel situations.
3. Has a strong ability to generate and evaluate different design alternatives based on the criteria described in the course.

Has a strong ability to reason about the data mining solutions performance and ability to adapt to new data.

Services

Support for students with disabilities: DISABILITY SUPPORT SERVICES (Student Health and Wellness - SHAW): The University of Washington Tacoma is committed to making physical facilities and instructional programs accessible to students with disabilities. Disability Support Services (DSS) functions as the focal point for coordination of services for students with disabilities. In compliance with Title II of the Americans with Disabilities Act, any enrolled student at UW Tacoma who has an appropriately documented physical, emotional, or mental disability that “substantially limits one or more major life activities [including walking, seeing, hearing, speaking, breathing, learning and working],” is eligible for services from DSS. If you are wondering if you may be eligible for accommodations on our campus, please contact the DSS reception desk at 692-4828, or visit http://www.tacoma.washington.edu/studentaffairs/SHW/dss_about.cfm

CSS Mentors: The CSS mentors staff the Science Lab (SCIENCE 106) throughout the week. They can provide help with specific questions about specific classes. Please note, however, that they will not do your homework for you. Instead, they will help you when you get stuck (either in programming or in homework) and help you develop the reasoning skills you need to solve future problems. See the CSS Mentors website for information about when the mentors are in the labs and other information.

Center for Teaching, Learning & Technology: The Center for Teaching, Learning & Technology offers academic and technical support for students at all levels of expertise - review, upper division, graduate and TA. For your writing, reading, study skills and public speaking needs, please make an appointment online at www.tacoma.washington.edu/ctlt/ or visit KEY 202. For your Math needs, assistance is available on a drop-in basis, Monday Thursday, hours to be posted. For multimedia or video projects, please visit the Multimedia Lab located in MAT 251. For student software training, please register at www.tacoma.washington.edu/CTLT/training/student/index2.cfm

Safety Escorts: Safety escorts are available to accompany you to your vehicle Monday through Thursday from 5:00pm to 10:30pm, except holiday's, Breaks and Summer quarter. Dial #300 from a non-campus phone or #333 on a campus telephone and a Campus Safety Escort will walk you safely to your vehicle. In case of emergency, follow your professor’s instructions. When an alarm sounds, evacuate the building immediately. MATT, CP, WG, GWP, and BB buildings assemble in the Cragle Parking Lot south of the library. BHS, WCG, and DOU buildings assemble near the transit station next to the Pinkerton Building on Broadway across from Spaghetti Factory. Pinkerton occupants go to the convention center parking lot north of Pinkerton. For more information about emergency procedures and information, please go to: http://www.tacoma.washington.edu/safety/
Emergency Phoning: From campus phones, report emergencies by dialing 9-911 and state the T-number that is on a sticker on the phone; from non-campus phones dial 911. Building location numbers are posted on all buildings. For assistance with non-emergencies call Campus Safety at 2-4416 from a campus phone, and 253-692-4416 from a non-campus phone.

Inclement Weather: In the event of inclement weather, UWT's hotline at 253-383-INFO indicates whether classes have been cancelled. Please see the inclement weather page at: http://www.tacoma.washington.edu/policies_procedures/Inclement_Weather.pdf for more information.

Additional Policies

Collaboration: All assignments in this course are to be done as indicated earlier individually or in your own group(s). This does not mean that you cannot discuss anything about this course with others. What it does mean is that anything that you hand in must accurately represent your knowledge and work. It is expected that all students will read and follow guidelines as detailed in the UW Student Conduct Code.

Plagiarism: This class will heavily involve the use of the written works of others. Your own written work will involve discussing the ideas of others. When using the ideas of others, it is important to acknowledge whose ideas you are using, and to clearly distinguish the ideas of others from your own. To convey the impression, whether inadvertently or deliberately, that another's work is your own, is called plagiarism. Plagiarism is a serious offense in the university. Prof. Tenenberg has written a guideline on plagiarism and how to avoid it (i.e. by scrupulously citing your sources), and I expect that you will abide by it. Although this guideline is geared toward the use of other's computer programs, it applies equally well to other kinds of text, such as those that you will use in this class.