Data Mining 2: Advanced Concepts and Algorithms

Instructor Information

Name: Natasha Balac, Ph.D.
Email: nbalac@eng.ucsd.edu

You may contact me by email or via Blackboard. **IMPORTANT:** We will try to answer your emails/questions/posting within 48 hours of receiving them.

Welcome

My teaching philosophy is to offer structured and clear explanation of the subject matter and ground in a variety of hands-on exercises and successful applications of the acquired new concepts, methods and technologies. This approach fosters deep understanding of the topics covered in class and accelerates up the preparedness of the students to apply the newly learned techniques onto their own projects. We encourage students to utilize the Blackboard for both interactions with the other students as well as with the TA and the professor.

Course Purpose and Prerequisites

As the amount of research and industry data being collected daily continues to grow, intelligent software tools are increasingly needed to process and filter the data, detect new patterns and similarities within it, and extract meaningful information from it. Data mining and predictive modeling offer a means of effective classification and analysis of large, complex, multi-dimensional data, leading to discovery of functional models, trends and patterns.

Prerequisites: Fundamentals of Data Mining and Data Preparation for Analytics required.

This course covers advanced data mining, data analysis, and pattern recognition concepts and algorithms. Building upon the skills learned in Fundamentals of Data Mining, this course includes models, machine learning algorithms, and advanced methods and applications.

Course Goal and Objectives

This course provides students with a foundation in advance data mining, pattern recognition, text mining and Big Data concepts and algorithms. It begins with an overview of the advanced data mining process and approaches. Practical exercises include various data analysis and machine learning techniques for model and knowledge creation though learning from examples.

Course Materials/Textbooks/Software

Ian H. Witten, Eive Frank, Mark A. Hall: **Data Mining: Practical Machine Learning Tools and Techniques**, Fourth Edition

Optional – free online:

David Olsen, Dursun Delen: http://lib.mdp.ac.id/ebook/Karya%20Umum/Advanced-Data-Mining-Techniques.pdf

Hurwitz, Nugent, Halper, Kaufman: Big Data For Dummies http://eecs.wsu.edu/~yinghui/mat/courses/fall%202015/resources/Big%20data%20for%20dummies.pdf

Software:

WEKA, a collection of Windows-based applications, is used in hands-on class assignments. Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data preprocessing, for modeling and data analysis and visualization.

Course Overview

This course has 10 sessions and the final assignment due at the end of the course. The topics are as follows:

Topics include:

- Advanced Data Mining techniques overview
- Data Mining with Big Data
- Artificial Neural Networks
- Probability Graph Models and Bayesian learning
- Support Vector Machines
- Ensemble learning: Bagging, Boosting, Stacking
- Random Forests
- Boosted Trees
- Data Mining Tools Review and Survey
- Text Mining

Practical experience:

- Hands-on data mining projects

Online Course Structure

Announcements	This is the first page you see upon entering your course. Your instructor will post weekly announcements and reminders here.
Introduction	Contains an introduction to the course and instructor biography.
Syllabus	Contains the course outline, learning objectives, weekly assignments and course details.
Lessons	If it's a fully online course, this section will have the instructor's weekly audio/image lectures. The lectures are self-paced and can be replayed like a video movie (start, pause, rewind, etc.).
Discussion Board	Questions pertaining to each lesson are posted weekly for you and your classmates to discuss and answer.
Assignments	Assignments, quizzes, Course Evaluation, and the Final Exam are available here.
Resources	Additional readings and handouts, web site links, and PowerPoint presentations are here.
Contacts	Instructor, student services and online learning support contact information is listed here.
Tools	Check your grades (My Grades), or access the Blackboard User Manual (User Manual) here.

Course Schedule

Week	Topic	Reading	Assignments	Points
1	Advanced Data Mining techniques overview	Witten,Frank,Hall 13.1; Olsen, Delen Ch. 1,2,10	Reading Assignment	
2	Data Mining with Big Data	Witten,Frank,Hall 13.2, Hurwitz, Nugent, Halper, Kaufman (optional)	Quiz #1	10
3	Artificial Neural Networks Part 1	ANN backprop reading under Resources Section	Topic review	
4	Artificial Neural Networks Part 2	ANN backprop reading under Resources Section	Assignments #1	20
5	Support Vector Machines	Witten,Frank,Hall 7.2	Assignments #2	10
6	Probability Graph Models and Bayesian learning	Witten,Frank,Hall 9.1, 9.2	Assignments #3	10

7	Ensemble learning: Bagging, Boosting, Stacking	Witten,Frank,Hall Chapter 12	Topic review	
8	Random Forests, Boosted Trees	Witten,Frank,Hall 12.2,12.3	Assignments #4	20
9	Text Mining	Witten,Frank,Hall 13.5; Text Mining reading under Recources	Assignments #5	20
10	Data Mining Tools Review and Survey	Tools Review documentation under Resources		
1-10	Class Participation	Class Participation	Class Participation	10
		TOTAL POINTS POSSIBLE		**100**

If you have 3rd Edition of the Book please refer to the table below:

Week	Topic	Reading	Assignments	Points
1	Advanced Data Mining techniques overview	Olsen, Delen Ch. 1,2,10	Reading Assignment	
2	Data Mining with Big Data	Witten,Frank,Hall 9.2; Hurwitz, Nugent, Halper, Kaufman	Quiz #1	10
3	Artificial Neural Networks Part 1	Witten,Frank,Hall 11.4; ANN backprop reading under Resources Section	Topic review	
4	Artificial Neural Networks Part 2	ANN backprop reading under Resources Section	Assignments #1	20
5	Support Vector Machines	Witten,Frank,Hall 6.4	Assignments #2	10
6	Probability Graph Models and Bayesian learning	Witten,Frank,Hall 6.7	Assignments #3	10
7	Ensemble learning: Bagging, Boosting, Stacking	Witten,Frank,Hall Chapter 8	Topic review	
8	Random Forests, Boosted Trees	Witten,Frank,Hall 8.3	Assignments #4	20

9	Text Mining	Witten,Frank,Hall 9.5, 17.3; Text Mining reading under Recourses	Assignments #5	20
10	Data Mining Tools Review and Survey	Tools Review documentation under Resources		
1-10	Class Participation	Class Participation	Class Participation	10
		TOTAL POINTS POSSIBLE		**100**

Requirements

In order to satisfy course requirements, class participants must participate in discussions, complete all course assignments on time (on or before the due date), and use graduate level writing/presentation for all written assignments.

IMPORTANT! Late assignments (anything posted or sent after the due date) will be graded -1 point for each day late unless due to a verifiable medical or family emergency. Assignments sent with the wrong naming convention or in the wrong format will be considered late until they are sent correctly. Late assignments will be accepted at the discretion of the instructor and cannot be accepted more than 1 week late. Grades are lowered for less-than-optimal (non graduate level) grammar, spelling, and presentation. Make sure all references are correctly cited and follow APA or MLA guidelines.

In general, the performance criteria for an A grade for assignments is listed below:
The assignment:

- Demonstrates a high level understanding of issues, including complexities.
- Is well focused and sequenced. Has a clear sense of purpose. Thoughts are clearly developed and easily understandable.
- Critically evaluates the topic beyond what is stated in readings, research, and discussions. Makes connections.
- Expresses views clearly. Provides specific examples, details, illustrations, anecdotes, etc. to support positions taken.
- Does more than repeat what the text says or what was said in class. Draws out additional important implications.
- Shows originality of thought.
- Uses proper citations for resources.
- Uses organizers: table of contents, topic headings, etc.
- Has no punctuation, grammar, spelling errors. Style, formatting, and appearance add to quality of final product.

Expect and plan for contingencies and technical problems (they WILL happen!). Written assignments MUST be sent as a PDF attachment! No exceptions.

Assignments/Quizzes/Discussion Board Participation

Discussion Board Participation—10 points

Regular presence in blackboard discussions. Substantial contributions are expected to gain full points. This may include taking a leadership role in weekly online discussions. See more information below.

Quiz #1—10 points

This is a simple 10 multiple-choice question quiz. You may only attempt the quiz one time. It is open book but you need to complete the quiz by noon on (put actual calendar date and time here) to have it count. Do NOT attempt this quiz at 11:45am. Try it a day or two before it is due to insure that you don't have any technical difficulties the day it's due. You will receive feedback immediately upon taking it. The quiz will be available from noon on Wednesday until it closes at noon on (put actual calendar date and time here)

.

Assignment #1—20 points

Handout explaining the details of assignment #1 focusing on Artificial Neural Networks can be found under the Assignments folder. The assignment is due by Sunday midnight of week #4.

Assignment #2—10 points

Handout explaining the details of assignment #2 focusing on Support Vector Machines can be found under the Assignments folder. The assignment is due by the end of week #5.

Assignment #3—10 points

Handout explaining the details of assignment #3 focusing on the topic of Probability Graph Models and Bayesian learning. The assignment is due by the end of week #6.

Assignment #4—20 points

Handout explaining the details of assignment #4 focusing on the topic of Ensemble Models. The assignment is due by the end of week #8.

Assignment #5—20 points

Handout explaining the details of assignment #5 focusing on the topic of Text Mining. The assignment is due by the end of week #9.

Grades

No late assignments or quizzes are accepted.

Grades are based on points and the letter grades are given as follows:

A+	97-100
A	94-96
A-	90-93
B+	87-89
B	84-86
B-	80-83
C+	77-79
C	74-76
C-	70-73
D+	67-69
D	65-66

F 0-64

You may check your grade anytime by clicking **Course Tools** and then **My Grades**. This will show you the points you have earned so far in this course.

Weighted Grades

Discussion Board Participation:	10%
Assignments/Quiz:	90%
TOTAL	100%

About Discussion Board Participation

A regular presence is expected in blackboard discussions and substantial contribution about class topics and discussion questions. What this means, essentially, is coming into blackboard regularly (at least 3 times a week) and posting your thoughts about the topic. Here are the attributes of effective discussion board participation:

- start a discussion on blackboard (add a thread)
- respond thoughtfully to a topic or another person's post
- provide links and resources related to the topic
- pose a thought-provoking question related to the topic
- provide pros and cons
- thoughtfully rebut another person's comments
- make your postings in a timely manner
- take a leadership role for weekly postings, be the one to start the discussion and encourage others

In grading, quality and quantity are considered. Regular contributions that add to the knowledge base of other students, links to additional resources, and providing substantive thought receive points. If you don't know a lot (yet!) about the topics, feel free to share some questions to others and/or search the Internet and share what you find with the class.

Note: Discussion Board participation is 10% of your grade.

About Assignments/Quizzes

Assignments are to be saved in Microsoft Word as a .doc file. Please name them as week1-lastname.doc and use the Assignment tool within Blackboard. When you click on the assignment name in the Assignments area, you will see the assignment directions. Under that is a link to complete the assignment. Click this and then you'll see a Browse button, click this, and locate your assignment (the Word document) on your computer and Open it. This will upload your assignment. Click Submit when you are done uploading it. You may only upload the assignment one time so be sure you have completed it and that you select the correct file. Assignments and quizzes are 90% of your grade.

The quizzes must be completed by 12pm, Pacific Standard Time, on the day shown above. It will no longer be available after that time. Plan ahead and give yourself plenty of time to complete it. These quizzes are based upon the lesson and the readings so do both before completing the quiz. You may only attempt the quiz once.

Student Resources

On any Blackboard screen, there are tabs across the top and one is called the Student Tab. There is information on how to get started as a student and who to contact if you encounter any problems. There are also videos and written instructions on how to do some of the most common things in Blackboard.

Another one of these tabs is called FAQ (Frequently Asked Questions). If you click on the Students Category (on the left), you can find step-by-step directions for everything from sending email to uploading your assignments to posting a reply on the discussion board.

Campus Emergencies

In the event of an emergency, information will be posted at UC San Diego Extension (http://extension.ucsd.edu/). Extension students must access the website to find out the status of the emergency situation. Email and or phone lines may not be accessible. Information will be updated online as the situation progresses and an ALL CLEAR will be posted once the situation is resolved.

Code of Conduct

All participants in a course at UC San Diego Extension are bound by the University of California Code of Conduct, found at http://www.ucop.edu/ucophome/coordrev/ucpolicies/aos/uc100.html.

Academic Honesty Policy

The University is an institution of learning, research, and scholarship predicated on the existence of an environment of honesty and integrity. As members of the academic community, faculty, students, and administrative officials share responsibility for maintaining this environment. It is essential that all members of the academic community subscribe to the ideal of academic honesty and integrity and accept individual responsibility for their work. Academic dishonesty is unacceptable and will not be tolerated at the University of California. Cheating, forgery, dishonest conduct, plagiarism, and collusion in dishonest activities erode the University's educational, research, and social roles.
If students who knowingly or intentionally conduct or help another student perform dishonest conduct, acts of cheating, or plagiarism will be subject to disciplinary action at the discretion of UC San Diego Extension.