BOR 6335 Data Mining

Course Description

This course provides an overview of data mining and fundamentals of using RapidMiner and OpenOffice open access software packages to develop data mining models. Using the CRISP-DM methodology, the principles and practice of data mining are illustrated through the data sets and exercises in the textbook. This course follows a typical path of a data mining project starting with learning how to read data, transforming data into useable formats, developing the data mining model, and interpreting the results of the model. The course provides basic model validation methods to ensure the validity and strength of the data mining model. Ethical considerations will be explored to examine moral issues and laws concerning data mining techniques and methodologies.

Course Bibliography and Required Readings

North, M. (2012). *Data Mining for the Masses*
ISBN: 0615684378
ISBN: 13: 978-0615684376

Available as free eBook at:
http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/DataMiningForTheMasses.pdf

Rapid Miner 5 Software
Open Source software downloadable at http://rapid-i.com/content/view/26/84/

- Go to Download and you will need to register to download Rapid Miner 5. This software will be necessary to complete this course.

OpenOffice Software
Open Source software downloadable at http://www.openoffice.org/

Data Sets for Lessons
Downloadable datasets for each Lesson:
http://sites.google.com/site/dataminingforthemasses/

Prerequisites

A working knowledge of maneuvering through basic computer software programs, working with EXCEL or other types of spreadsheet software (this course will use Open Office), and the ability to understand basic mathematics.

Technical Skills Required for This Course

AS with all online courses, students must be able to operate a computer and have the necessary technical skills to navigate around a web page. Additional technical skills are not a prerequisite for this course, however your computer must meet certain minimum requirements to operate Blackboard.

Time Spent on this Course

Students can expect to spend a minimum of 6 hours per week to complete all readings and assignments. The lessons themselves take as long as it requires the student to read the materials and watch or listen to media presentations.

Course Objectives/Learning Outcomes

Objective 1: Introduce basic data mining principles and technics. Gain knowledge about the methods of Data Mining to include proactive analysis, predictive modeling and identifying new trends.

Objective 2: Become conversant with terminology of Data Mining and how these terms can be applied to gathering, understanding, and analyzing data sets.

Objective 3: Learn how to produce quantitative analysis reports relative to Data Mining.

Objective 4: Be able to explain the concepts of basic quantitative methods and their purpose and importance to data mining to identify proper application of reports for agencies.

Objective 5: Be able to identify and analyze basic data mining algorithms, methods, and tools.

Objective 6: Learn developing areas of data mining such as ethical issues of data mining techniques, text mining, and web mining.

Objective 7: Be able to apply critical thinking, problem-solving and decision-making skills to the entire process of data mining and the results of such.

Objective 8: Create and environment conducive to online learning of difficult data mining methods and techniques using various open access Internet resources, readings and data management tools to complete complicated data mining procedures, analyses and the use of available technology to complete these goals.

One consistent skill which you will need in any future career is that of effective writing and the ability to clearly communicate your thoughts. Therefore, you will be assigned weekly projects that evaluate your ability to write clearly. Your instructor will grade your assignments on technical skills, such as clear organization, spelling and grammar usage, as well as a subjective assessment of whether or not you are able to think critically and analyze assigned subject matter.

Grading Policies

This course uses weekly data mining exercises to measure the student's comprehension of the presented materials and the use of the skills and techniques presented in each lesson. It is imperative that students read ALL of the provided materials and practice on the methods and procedures assigned each week. Staying current with course is important. The subject matter in and of itself is particularly intense so staying on top of the subject matter and utilizing outside sources to better understand the information is commanding.

Assignment	**Percent of Grade**	**Due**
Weekly Assignments/Exercises Modules 2-7	80%	Sundays - Modules 2 - 7
Participation in the Discussion Board – Module 1 and Module 8 will have Discussion Boards	20%	Sunday – Module 1 Wednesday – Module 8

Angelo State University employs a letter grade system. Grades in this course are determined on a percentage scale:

A = 90 – 100 %
B = 80 – 89 %
C = 70 – 79 %
F = 59 % and below.

The professor will post discussion questions for Module 1 (Week 1) and Module 8 (Week 8) and you will be required to make at least two responses toward each of the discussions as a minimum. These postings should be made thoughtfully and you should be able to provide evidence for your conclusions through the reading materials or other source documents available to you. A source document for your postings on the discussion is not Wikipedia.

The weekly assignments will be on learning data mining procedures as provided in the chapters of the assigned book. These assignments will provide a working knowledge of the subject matter in each chapter as well as the ability to maneuver through the Rapid Minder software to manipulate the data as required per applied lesson. Due dates are listed in the individual lessons.

This course is being taught as a basic data mining course consistent with the suspected ability of students not majoring in computer science or mathematical statistics. The understanding of computer learning methods has much inconsistency among students so those with more experience may deem the material less challenging at times but will be more-so to others. Working together to share knowledge will make the experience far more gratifying for all. Do not hesitate to ask others for their opinions, assistance or otherwise as this material is most likely new to all. Additionally, since this is a graduate course, you are expected to think critically of the weekly subject matter and be able to develop the rationale for your opinions as to what is working or not working and be able to provide some evidence for your conclusion(s).

Writing Guidelines

There are no research papers or written assignments in this course. Each student in responsible to complete weekly data mining exercises and upload the completed project via Blackboard.

Uploading Assignments

A video that describes how to upload assignments in Blackboard can be viewed by clicking this link:
Uploading Blackboard Assignments - video

A printable version of these instructions can be viewed by clicking this link:
Uploading Blackboard Assignments - PDF

Warning

Any PLAGIARISM will not be tolerated and can result in the failure of a course and dismissal from the University.

Rubrics

Discussion forums and writing assignments will be graded using a standardized rubric. It is recommended that you be familiar with these grading criteria and keep them in mind as you complete the writing assignments. There are two rubrics. Click the link to download the PDF document:
Discussion Rubric
Writing Assignment Rubric

Date and Time of Final Exam

There is no final exam in this course.

Course Organization

Module 1:

Introduction to Data Mining and CRISP-DM (Ch 1)

The ultimate aim of this lesson is to introduce the basic concepts and practices of Data Mining. This chapter will discuss the tools necessary for data mining such as software. Instructions how to download the open access programs that will be used in this course will be included as well as an explanation of how and why they will be used. The steps of Cross-Industry Standard Process for Data Mining (CRISP-DM) will be introduced that led to the formalization and standardization the data mining process.

Organizational Understanding and Data Understanding (Ch 2)

This lesson will discuss the contexts and perspectives of data mining to help the student gain a real-world understanding of how data mining can solve organizational problems. Potential uses of data mining and how their applications are important will be brought out. The purpose, intent and limitations will be explained in this chapter. The fundamentals of data bases, data collection and data organization are important as these impact the reliability and quality of all data mining activities. Organizational and Operational data are the two types of data that can be mined and descriptions of each will be introduced. Lastly, data and human privacy and security are paramount when embarking on data mining techniques. The importance of such will be emphasized in Chapter 2.

Module 2:

Data Preparation (Ch 3)

Before any data mining models can constructed, three steps of CRISP methodology must be achieved. Organizational Understanding and Data Understanding and the first two and Data Preparation is the third that will be discussed here. You must install both Open Office and RapidMinder to complete this lesson. Refer to Chapter 1 for details on downloading these

programs. Basic data preparation will begin in Open Office and then on to data preparation tools in RapidMiner. Data collation and organization is important to begin data mining activities. Data scrubbing to increase quality of data and learn ways to handle missing data, reduction of data, handling inconsistencies in data and reducing attributes will be introduced through hands-on exercises.

NOTE: You must be very familiar with this Module and Chapter 3 in your book as this is the basis of the rest of the course.

Module 3:

Correlation (Section 2 - Ch 4)

This lesson stresses the importance of knowing how correlation affects the attributes in a data set. Correlation is a statistical measure of how strong these relationships are. This is the basic construction of a data mining model and easier to build, understand, run and interpret thus an serving as a less difficult starting point. Correlation is primarily a Classification part of data mining and is rarely used for prediction other than examining general trends and how the movement of one attribute affects another. Correlations help one learn how strongly interactions are when, and if, they occur.

Association Rules (Ch 5)

The goal of this lesson is the show how the association rule in data mining can identify linkages in data that can have a practical application. CRISP's cyclical nature is used to understand that data mining sometimes involves some back and forth "digging" in order to move to the next step. Support and Confidence percentages are calculated to know the importance of these metrics and determining their strength in the data set.

Module 4:

Means Clustering (Ch 6)

Means clustering is data mining model that falls primarily on the side of Classification in a Venn diagram. It is not necessarily a predictor but simply takes known indicators from a data set and groups them together based on similarity to group averages. An example of this might be the likelihood of young persons ages 18-24 being predisposed to delinquency. One cannot predict a particular person will engage in delinquent behavior but means clustering is an effective way to group observations based on what is typical for the group. RapidMiner will give one a powerful means to find natural groups in a data set.

Discriminate Anaylsis (Ch 7)

This chapter includes Discriminate Analysis and begins the cross-over between Classification and Prediction. With a similar process as k-clustering, and with the right attributes, one can generate predictions for a data set. Discriminate Analysis can be used when the classification for one observation is known and another not known. This method is useful in predicting potentially successful career paths for students based on personality traits, preferences, and aptitudes.

Module 5:

Linear Regression (Ch 8)

This lesson will teach the student how to recognize the necessary format for data in order to generate linear regression numeric predictions when training and scoring data sets. Each attribute in the data set is evaluated statistically for its ability to predict the target attribute. This allows researchers to remove attributes from the model that are not strong predictors. Linear Regression is important to data mining models in that once predictions are calculated, the results can be summarized in order to determine if there are differences in the predictions in subsets of the scoring data.

Logistic Regression (Ch 9)

Logistic regression is a way to predict whether something will happen or not and confident we are of the predictions. Different that linear regression, logistic regression uses nominal data to categorize observations in a data set into their probable outcomes. Logistic regression will be used in this model to quickly and safely predict the outcome of a particular phenomenon in the data set and to determine how accurate this prediction is.

Module 6:

Decision Trees (Ch 10)

This lesson will demonstrate how Decision Trees are excellent predictive models when the target attribute is categorical in nature and the data set has mixed types of attributes. This method handles missing or inconsistent values while still generating useable results. Decision trees will tell us what is predicted, how confident we are in our prediction, and how we arrived at the prediction. The how represented by nodes and leaves labeled by branch arrows thus the name ‘decision tree’.

Neural Networks (Ch 11)

Neural networks use artificial neurons to compare attributes to one another and look for strong connections. This method has been said to be compared to artificial intelligence or to attempt to mimic the human brain. In this, data mining models using neuron networks are not as limited regarding value ranges as some other methodologies. Neural networks find interesting observations that may not be obvious but still give data miners the opportunity to solve problems.

Module 7:

Text Mining (Ch 12)

This method of data mining allows the observer a powerful way to analyze data in an unstructured format such as paragraphs of text. The text is broken down into 'tokens' allowing further manipulation such as phrases, word groupings, case sensitivity among other things. Trends can be revealed using this method. The 'tokens' can be modeled just as other data sets and then analyzed through RapidMiner. Text documents such as company complaints or positive comments can be mined to reveal specific results from multiple documents.

Module 8:

Evaluation and Deployment (Ch 13)

This lesson will discuss some cautions that should be taken before putting any real-world data mining results into practice. Suppose you find that your decision making methods while developing your data mining model have turned out less than ideal results. The entire model may not need to be scrapped. The process of Cross Validation can be performed to check for the likelihood of false positive predictive models in RapidMiner. Students will learn how Cross-Validation, a performance operator that can be used to check a data set's ability to perform, is used to examine the validity and strength of the data mining model.

Data Mining Ethics (Ch 14)

Ethical considerations concerning data mining will be examined in this lesson. As with all statistical research, there are people behind the data who one may be examining shopping habits, health issues, and even the possibility of being a juvenile delinquent. Data mining ethics should always be held at a level above and beyond the desired results of a research model. This lesson will look at moral issues, laws, and truths concerning the proper behavior that should govern the data miner in research endeavors.

Communication

Participation

In this class ***everyone***, brings something to the table. Your ideas and thoughts do count, not only to me, but the entire class. Feel free to ask questions either via e-mail or the discussion board. **Check the discussion board regularly.** Many student questions are applicable to the class as a whole, as are the responses. You may be surprised how many of your classmates have the same questions and concerns as you. I may simply post your particular question on the discussion board and allow your classmates to provide the answer through their own posts.

To some, this may be their first online class and naturally, it could seem somewhat intimidating. As a class, we are together to help each other with this learning process and share our collective knowledge on how best to communicate; how to resolve technical issues that may arise (if we have the expertise), and to assist each other to find answers to our questions.

We will learn and work as a team.

Courtesy and Respect

Courtesy and Respect are essential ingredients to this course. We respect each other's opinions and respect their point of view at all times while in our class sessions. The use of profanity & harassment of any form is strictly prohibited (Zero Tolerance), as are those remarks concerning one's ethnicity, life style, race (ethnicity), religion, etc., violations of these rules will result in immediate dismissal from the course.

Attendance

This is an online course and attendance is not taken. However, failure to participate in the discussion board, to communicate or respond to e-mails from the professor, is an indication something is wrong. Therefore, we have made both a significant component of the course grade as an enticement to keep you engaged in the learning process. Failure to participate or communicate on the part of a student will result in an appropriate reduction of your grade and possibly in your failure of this course.

Late Work

Late work will result in a deduction of 10 points per day. No late work will be accepted after the third day an assignment or discussion is late.

Incompletes

The University policy on grades of "Incomplete" is that the deficiency in performance must be addressed satisfactorily by the end of the next long (16 week) semester or the grade automatically becomes a "F". Grades of "Incomplete" will only be awarded to students who have demonstrated sufficient progress to earn the opportunity to complete the course outside of the normal course duration. The award of an "Incomplete" will only be made in rare circumstances, with the concurrence of the student and the professor on what specific tasks remain and when they are due for the grade to be changed to a higher grade. The determination of the need to award an "Incomplete" is entirely up to the professor's personal judgment.

Important Dates

Students may add this course up to the last Friday of the first week of class.

Students may drop this course up to the 6th day of the class or the last drop date as specified by the University Administration.

Office Hours and/or Hours of Outside-of Class Contact

This is an online course, thus there are no set office hours. Refer to the Instructor Information section in the Blackboard course for details regarding the instructor's contact information.

University Policies

Academic Integrity

Angelo State University expects its students to maintain complete honesty and integrity in their academic pursuits. Students are responsible for understanding and complying with the university Academic Honor Code and the ASU Student Handbook.

Accommodations for Disability

The Student Life Office is the designated campus department charged with the responsibility of reviewing and authorizing requests for reasonable accommodations based on a disability, and it is the student's responsibility to initiate such a request by contacting the Student Life Office at (325) 942-2191 or (325) 942-2126 (TDD/FAX) or by e-mail at Student.Life@angelo.edu to begin the process. The Student Life Office will establish the particular documentation requirements necessary for the various types of disabilities.

Student absence for religious holidays

A student who intends to observe a religious holy day should make that intention known in

writing to the instructor prior to the absence. A student who is absent from classes for the observance of a religious holy day shall be allowed to take an examination or complete an assignment scheduled for that day within a reasonable time after the absence.