MSIS-DL 435 Syllabus
Atef Bader, Ph.D.
Data Warehousing & Data Mining
Summer 2013

Contact Information

E-mail: a-bader@northwestern.edu
Office Hours: By appointment only

Course Description

Data mining involves statistical and artificial intelligence analysis applied to large-scale data sets in different fields, such as science, medicine, and business. This course covers the fundamental concepts of data mining for exploring, processing, and analyzing data. The course is organized into three sections. Section I introduces the techniques proven effective in classifying the data. Section II contains the concepts often used in rule association. Section III focuses on clustering when data is complex. Using real-world data sets from open source, the course uses Weka freeware tool for analysis and data mining. The course concludes with a final project and evaluates relative advantages of several contemporary algorithms.

Text

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

[ISBN-13: 978-0321321367]

Software

Weka

Prerequisite

MSIS-DL 317

Learning Goals

The goals of this course are to:

- Prepare data for analysis.
- Recognize popular data mining algorithms in classification, association, clustering, and regression.
- Identify the correct data mining techniques useful for the data.
- Discover and present patterns that are embedded in the data.

Evaluation

Homework (6): 420 points
Course Project: 280 points
Course Participation (Exercises and Discussion Board participation): 300 points
Total = 1,000 points

Discussion Board Etiquette

The purpose of discussion boards is to allow students to freely exchange ideas. It is important that we always remain respectful of one another's viewpoints and positions and, when necessary, agree to disagree, respectfully. While active and frequent participation is encouraged, cluttering a discussion board with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in

receiving less than full credit. Although frequency is not unimportant, content of the message is paramount. Please remember to cite all sources—when relevant—in order to avoid plagiarism.

Proctored Assessment
There is no proctored assessment requirement in this course.

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

Attendance
This course is asynchronous, meaning that we will not meet at a particular time each week. Even though we will not meet face-to-face in a physical classroom, participation in discussion boards is required and paramount to your success.

Late Work
Late assignments are not accepted without explicit permission from the instructor, and permission can be granted only in the case of an emergency and in advance of the assignment due date. Late work may be subject to a penalty in points.

Learning Groups
There will be no learning groups in this course.

Academic Integrity at Northwestern
Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., taking material from readings without citation or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit: <www.scs.northwestern.edu/student/issues/academic_integrity.cfm>.

Plagiarism is one form of academic dishonesty. Students can familiarize themselves with the definition and examples of plagiarism by visiting the site <www.northwestern.edu/uacc/plagiar.html>. A myriad of other sources can be found online as well.

Some assignments in this course may be required to be submitted through SafeAssign, a plagiarism detection and education tool. You can find an explanation of the tool at <http://wiki.safeassign.com/display/SAFE/How+Does+SafeAssign+Work>. In brief, SafeAssign compares the submitted assignment to millions of documents in very large databases. It then generates a report showing the extent to which text within a paper is very similar or identical to pre-existing sources. The user can then see how or whether the flagged text is cited appropriately, if at all. SafeAssign also returns a percentage score, indicating the percentage of the submitted paper that is similar or identical to pre-existing sources. High scores are not necessarily bad, nor do they necessarily indicate plagiarism, since

the score doesn't take into account how or whether material is cited. (If a paper consisted of just one long quote that was cited appropriately, the score would be 100%. This wouldn't be plagiarism, due to the appropriate citation. However, just submitting one long quote would probably be a pretty bad paper.) Low scores are not necessarily good, nor do they necessarily indicate a lack of plagiarism. (If a 50-page paper had all original material, except for one short quote that was not cited, the score might be around 1%. But not citing a quotation would still be plagiarism.)

SafeAssign includes an option in which the student can submit a paper and see the resultant report before submitting it to the instructor as a final copy. This ideally will help students better understand and avoid plagiarism.

Other Processes and Policies
Please refer to your SCS student handbook at <www.scs.northwestern.edu/grad/information/handbook.cfm> for additional course and program processes and policies.

Course Schedule

Important Note
Changes may occur to the syllabus at the instructor's discretion. When changes are made, students will be notified via an announcement in Blackboard.

Session 1

Learning Objectives
After this session, the student will be able to:

- Distinguish between a relational database and dimensional database.
- Design a data warehouse using two dimensions: time and product.
- Describe the reasons for using relational design on active data warehouse.
- Name three active data warehouse commercial products, and draw their architectures.
- Explain the three data mining topics—classification, association, and clustering.
- Differentiate among the three data mining topics—classification, association, and clustering.
- Name three commercial or academic mining tools in the market, and describe the usages of these tools.

Course Content

Reading—For this session please read:

Chapters 1, 2, 3 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Reading, MA: Addison-Wesley.

Web link:
Oracle Database Concepts

Multimedia:
Introduction to Data Mining
Six Data Mining Tasks

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

There are no assignments due this session.

Session 2

Learning Objectives

After this session, the student will be able to:

- Write the pseudo code for a decision tree algorithm.
- Calculate the Gini index and entropy index for a small data set of 15 cases.
- Describe the differences between decision tree and regression.
- Install Weka software.
- Describe the three parts of ARFF file that inputs into Weka: title session, data format session, and data session.
- Use Weka and the Weather.arff file to run J48 decision tree from Weka.
- Interpret the output from Weka.
- Make decisions from the result of the decision tree by using ROC curve and confusion matrix.
- Use a training set to build a model, and use the model to predict the result from a given data set.

Course Content

Reading—For this session please read:

Chapter 4 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

Web links:

Weka: Introduction

ROC Graphs: Notes and Practical Considerations for Researchers

Multimedia:

The Decision Tree Classification Algorithm

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

Assignment 1 is due by Sunday, July 7, 2013 at 11:55 p.m. (central time). For more information, click *Assignments* on the left navigation bar in Blackboard, and scroll to this assignment's item.

Session 3

Learning Objectives
After this session, the student will be able to:

- Differentiate between a decision tree and neural network algorithms.
- Use Weka to run a neural network analysis.
- Draw the neural network layers (input layer, hidden layer, and output layer).
- Interpret the weight and network connections of a neural network result.
- Compare the output of a J48 decision tree and a neural network result.
- Describe the algorithm of rule-based classifier.
- Explain the main hypothesis of naive Bayes algorithm and the impact on the analysis when the hypothesis is violated.
- Compare the results from naive Bayes and rule-based (IBK algorithm from Weka).

Course Content

Reading—For this session please read:

Chapter 5 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

Multimedia:

Naive Bayes Classification Algorithm

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

Assignment 2 is due by Sunday, July 14, 2013 at 11:55 p.m. (central time). For more information, click *Assignments* on the left navigation bar in Blackboard, and scroll to this assignment's item.

Session 4

Learning Objectives
After this session, the student will be able to:

- Calculate the support index, confidence index, and lift index.
- Set up parameters of Apriori (minsup, minconf) to run association analysis.
- Draw frequent itemset lattice, and explain the algorithm of Apriori using pseudo code.
- Select association rules from the Weka output.
- Name four measures for Apriori-based algorithm to measure the strength of association rules.
- Calculate statistics-based measures by a given contingency matrix.

Course Content

Reading—For this session please read:

Chapter 6 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

Assignment 3 is due by Sunday, July 21, 2013 at 11:55 p.m. (central time). For more information, click *Assignments* on the left navigation bar in Blackboard, and scroll to this assignment's item.

Session 5

Learning Objectives

After this session, the student will be able to:

- Apply association analysis formulation to non-asymmetric binary variables such as income and age.
- Discretize the continuous variables, and explain the impacts on the support index and confidence index.
- Using statistics-based methods such as t-test and z-test, determine whether an association rule is interesting.
- Use a Min-Apriori algorithm to determine the support index of a given text data.
- Name four example resources of sequence data, such as the genome sequence.
- Find all subsequences using support >= minsup

Course Content

Reading—For this session please read:

Chapter 7 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

Web link:

Mining Association Rules

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

Assignment 4 is due by Sunday, July 28, 2013 at 11:55 p.m. (central time). For more information, click *Assignments* on the left navigation bar in Blackboard, and scroll to this assignment's item.

Session 6

Learning Objectives
After this session, the student will be able to:

- Distinguish between clustering and classification.
- List five examples of what is not clustering.
- Calculate three distance measurements (e.g. Euclidean distance, Manhattan distance, or correlation matrix).
- Compare the results from a fuzzy and non-fuzzy analysis from Weka.
- Find a hierarchical clustering using a cobweb algorithm from Weka.
- Compare the analysis from cobweb analysis and EM analysis.

Course Content

Reading—For this session please read:

Chapter 8 of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

.
Assignment 5 is due by Sunday, August 4, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation bar in Blackboard, and scroll to this assignment's item.

Session 7

Learning Objectives
After this session, the student will be able to:

- Describe the steps of the chameleon analysis.
- Explain the differences between the shared nearest neighbor (SNN) approach and the chameleon method.
- List five characteristics of a spatial data set.
- Name two more spatial data set clustering methods (e.g., ROCK and Jarvis-Patrick clustering).
- List three main limitations of SNN clustering.

Course Content

Reading—For this session please read:

Chapter 9, of the textbook:

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Reading, MA: Addison-Wesley.

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

Assignment 6 is due by Sunday, August 11, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation bar in Blackboard, and scroll to this assignment's item.

Session 8

Learning Objectives

After this session, the student will be able to:

- Calculate Eigen value and Eigen vector given a square matrix.
- Describe the steps of varimax rotation, and explain the reasons of dimension reduction.
- Perform principal component analysis (PCA)
- Perform linear discriminant analysis (LDA).
- Explain the result from LDA.
- Name two more non-linear dimension reduction analysis types (e.g. multidimensional scaling [MDA] and stochastic proximity embedding [SPE]).
- Explain the MDA results.

Course Content

Reading—For this session please read:

Web link:
SPSS Tutorials

Discussion Board

Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment

There are no assignments due this session.

Session 9

Learning Objectives

After this session, the student will be able to:

- Describe the limitation of linear regression after reviewing linear regression.
- Explain why logistic regression is used.
- Describe the estimation by maximum likelihood.
- Explain the coefficients obtained by logistic regression.
- Evaluate the performance of the regression model.
- Use Weka Experimenter to compare different algorithms.
- Use Weka Knowledge Flow to program the analytical process into batch sequence.
- Use Weka Experimenter to combine at least three algorithms into one project.
- Read the Weka Experimenter outputs to find the best algorithm.
- Program at least one algorithm into Weka Knowledge Flow, and load the data.
- Explain the result from Weka Knowledge Flow after loading the data.

Course Content
Reading—For this session please read:

Handouts:
Weka: Experimental Tutorial
Weka: Knowledge Flow Tutorial

Discussion Board
Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment
There are no assignments due this session.

Session 10

Learning Objectives
No new learning objectives will be introduced in this session.

Course Content
No new learning content will be introduced this session.

Discussion Board
Each session, you are required to participate in all discussion board forums. Your participation in both posting and responding to other students' comments is graded. For this week's discussion topic(s), visit the discussion board in Blackboard.

Assignment
Final Project is due by Sunday, September 1, 2013 at 11:55 p.m. (central time). For more information, click *Assignments* on the left navigation bar in Blackboard, and scroll to this assignment's item.