SYLLABUS

DSCI 4520 – Introduction to Data Mining

Fall 2014

CLASS (DAY/TIME): T 6:30-9:20, BLB 055
INSTRUCTOR: **Dr. Nick Evangelopoulos**
OFFICE HRS: TW 1:00-2:00pm, T 5:00-6:00pm, BLB 365D
CONTACT INFO: OFFICE PHONE: 940-565-3056
E-MAIL (preferred): **Nick.Evangelopoulos@unt.edu**

Textbooks in printed and PDF file format

- Kattamuri Sarma, *Predictive Modeling with SAS Enterprise Miner*, Second Edition, SAS Press 2013, ISBN: 978-1-60764-767-6 (required, printed text)
- *Getting Started with SAS Enterprise Miner 13.1*, SAS Publishing 2013, (required, PDF)
- *Getting Started with SAS Text Miner 12.1*, SAS Publishing 2012, (required, PDF)
- *Getting Started with SAS Enterprise Miner 5.3*, SAS Pub. 2008, (recommended, PDF)

Software

IBM SPSS Statistics 22, IBM SPSS Modeler 15, SAS Enterprise Miner 13.1, SAS Text Miner 13.1. All these are available at the CoB lab, physically and via VMWare.

Web Site

http://www.cob.unt.edu/itds/courses/DSCI4520/DSCI4520.htm
http://www.cob.unt.edu/itds/faculty/evangelopoulos/evangelopoulos.htm

Purpose of the Course

This course deals with the problem of extracting information from large databases and designing data-based decision support systems. The extracted knowledge is subsequently used to support human decision-making in the areas of summarization, prediction, and the explanation of observed phenomena (e.g. patterns, trends, and customer behavior).

Techniques such as visualization, statistical analysis, decision trees, and neural networks can be used to discover relationships and patterns that shed light on business problems. This course will examine methods for transforming massive amounts of data into new and useful information, uncovering factors that affect purchasing patterns, and identifying potential profitable investments and opportunities.

Learning Objectives

1. Understand the problems and opportunities when dealing with extremely large databases.
2. Review data visualization software used for interpreting complex patterns in multidimensional data. Learn to identify what information is useful and what is not.
3. Provide an understanding of predictive models and algorithms, as well as exploratory algorithms.

4. Examine all phases of decision making, including discovery and data query, data analysis and confirmation, presentation, and implementation of results.

Class Attendance
Regular class attendance and informed participation are expected.

Academic Integrity
This course adheres to the UNT policy on academic integrity. The policy can be found at http://vpaa.unt.edu/academic-integrity.htm. If you engage in academic dishonesty related to this class, you will receive a failing grade on the test or assignment, or a failing grade in the course. In addition, the case may be referred to the Dean of Students for appropriate disciplinary action.

Students with Disabilities
The College of Business complies with the Americans with Disabilities Act in making reasonable accommodations for qualified students with disability. If you have an established disability as defined in the "Act" and would like to request accommodation, please see your instructor as soon as possible.

Deadlines
Dates of drop deadlines, final exams, etc., are published in the university catalog and the schedule of classes. Please be sure you keep informed about these dates.

SETE
The Student Evaluation of Teaching Effectiveness (SETE) is a requirement for all organized classes at UNT. This short Web-based survey will be made available to you at the end of the semester/session, providing you a chance to comment on how this class is taught. I am very interested in the feedback I get from students, as I work to continually improve my teaching. I consider SETE to be an important part of your participation in this class.

Cell Phones
As a courtesy to your instructor and to your fellow classmates, you are asked to set your cell phone to vibrate, or switch it off. In case of a personal emergency, if you must use your cell phone, you are asked to step out of the classroom.

Incomplete Grade (I)
The grade of "I" is not given except for rare and very unusual emergencies, as per University guidelines. An “I” grade cannot be used to substitute your poor performance in class. If you think you will not be able to complete the class satisfactorily, please drop the course.

Campus Closures
Should UNT close campus, it is your responsibility to keep checking your official UNT e-mail account (EagleConnect) to learn if your instructor plans to modify class activities, and how. This may include changing assignment due dates, rescheduling quizzes and exams, etc.

Point Allocation

	DSCI 4520	DSCI 5240
Homework exercises (8 exercises)	25%	22.5%
In-class quizzes (8 quizzes, 3 dropped)	5%	4.5%
Mid-term Exam (in-class)	25%	22.5%
Final Exam (take-home)	20%	18.0%
Project (4 individual parts and 1 group part)	25%	22.5%
Graduate Presentation		10.0%
TOTAL	100%	100.0%

Bonus for attending/giving a graduate presentation	0.5%	0.5%
Bonus for graduate presentation – Option D		up to 20%

Letter Grades: 90% or more = A 80% or more = B
70% or more = C 60% or more = D Below 60% = F

Homework Exercises

There will be 8 homework exercises that you will have to turn in. Exercises will be using IBM SPSS Statistics, IBM SPSS Modeler, SAS Enterprise Miner, and SAS Text Miner. The homework exercises ask you to perform certain types of analysis, capture screen shots, and answer questions. Related handouts and PowerPoint slides with data description, step-by-step instructions, and assignment details, will be available on Blackboard. HW7 closely follows the text in *Getting Started with SAS Text Miner*, referred to as "GSTM text" below. **Homework is turned in electronically using Blackboard**, in the form of a report document. If you turn in your HW report late, 50% of HW credit is awarded.

HW1. Multiple Regression for TargetD using IBM SPSS Statistics. MYRAW data.
HW2. Logistic Regression for TargetB using IBM SPSS Statistics. Small sample effects. MYRAW data.
HW3. Overview of SEMMA process in SAS Enterprise Miner. Decision Tree and Logistic Regression. Model comparison. HMEQ data.
HW4. Scoring, Reporting in SAS EM. HMEQ data.
HW5. Clustering in SAS EM. PROSPECTS data.
HW6. Association Analysis in SAS EM. ASSOCS data.
HW7. Text Analytics in SAS Text Miner. Text cleanup, synonyms, stop list, topic extraction, and predictive modeling using text data. VAEREXT data. Based on the *GSTM* text.
HW8. Introduction to IBM SPSS Modeler. Decision Tree and Logistic Regression. HMEQ data.

Term Project

This course has a term project. You will be asked to analyze data related to the KDD-cup 98, an International competition for professional data miners. The data set will be available on Blackboard. Handouts describing what you have to do will be distributed in class. During the first 4 parts you will work individually and submit your work as a Word document that includes screen shots from Enterprise Miner and answers to various questions as described on the handouts. You will turn in your reports by uploading them on **Blackboard**. Grading and late

penalty policies for PR1-PR4 are the same as with HW1-HW8. During the last part of the project you will form groups. Each group will include **at least one undergraduate** student and **at least one graduate student**. The **maximum group size will be 6**. Groups will be self-managed. If the group is not satisfied with some member's contribution they may choose to dismiss that person from the group. In such a case, alternative individual assignment will be given to the dismissed group member. The group will turn in a single PR5 report, listing all group member names, in **printed hard copy** format. A summary of the project parts follows below.

	Topic	**Work type**
PR1.	Open the data, produce statistics and graphs	Individual
PR2.	Decision Trees/Regression	Individual
PR3.	Regression/Decision Trees	Individual
PR4.	Neural Networks	Individual
PR5.	Final Written Report. Comparison and evaluation of 3 models (Decision tree, logistic regression, neural net).	**Group**

DSCI 4520 TIME SCHEDULE – Fall 2014

The schedule below is a tentative outline for the semester. It is meant to be a guide and several items are subject to change. Certain topics may be stressed more or less than indicated.

Date	Topics	Assignment due
Aug. 26	Intro to Data Mining, Ch1 Multiple Linear Regression	
Sept. 2	Logistic Regression, Ch 6 Stepwise Procedure	HW1 (regression in SPSS, MYRAW)
Sept. 9	SEMMA, CRISP-DM, Model comparison, Ch7	HW2 (Log Reg in SPSS, MYRAW)
Sept. 16	Scoring, Reporting	HW3 (SEMMA, HMEQ)
Sept. 23	Decision Trees, Sarma Ch 4	HW4 (scoring, HMEQ)
Sept. 30	Decision Trees, Sarma Ch 4	PR1 (data explor., DONOR_RAW)
Oct. 7	Neural Networks, Sarma Ch 5 Clustering Analysis	PR2 (trees/reg, DONOR_RAW)
Oct. 14	Association Analysis Review for Exam 1	HW5 (clustering, PROSPECT) PR3 (reg/trees, DONOR_RAW)
Oct. 21	NO CLASS (Exam preparation)	PR4 (neural nets, DONOR_RAW)
Oct. 28	***** Exam 1 (in-class) *****	Grad Pres. signup deadline (type D only)
Nov. 4	Text Mining, Sarma Ch. 9	HW6 (market basket, ASSOCS)
Nov. 11	Text Mining, Sarma Ch. 9 Buffer lecture (use as needed)	Grad Pres. signup deadline (types A-C)
Nov. 18	Grad student presentations Buffer lecture (use as needed)	Grad Presentations are due by 5PM HW7 (text mining, VAEREXT) 0.5 pt. extra credit for attendance (UG st.)
Nov. 25	Grad student presentations Buffer lecture (use as needed)	HW8 (SPSS Modeler, HMEQ) 0.5 pt. extra credit for attendance (UG st.)
Dec 2	Course review Take-home final handed out	PR5 (project report) 0.5 pt. extra credit for attendance
Dec 9	***** Final Exam (take-home, due 12noon on Blackboard) *****	