STA380.18 Learning Structures & Time Series Syllabus
#04730 Spring, 2019

PROFESSOR:	Tom Sager
OFFICE	CBA 3.434B
OFFICE HOURS:	By appointment
TELEPHONE:	512-471-5232
E-MAIL:	TomSager@mail.utexas.edu
CANVAS WEB SITE:	https://courses.utexas.edu
TA:	Mr. Alek Dimitriev (dimitriev.aleksandar@gmail.com)
TEXTBOOK:	Not required, but recommendations are given – see discussion below.

PREREQUISITES:

1. Introductory statistics – a basic calculus-level course in statistical methods. Familiarity with the concepts and properties of mean, variance, covariance, correlation, confidence interval, hypothesis test, linear regression will be assumed in this course.
2. Linear algebra – familiarity with linear combinations, systems of linear equations and their solutions, matrix operations (addition, multiplication, inverse, determinant) will be assumed.
3. Mathematical maturity – comfort with mathematical reasoning. This is more important than any specific statistical or mathematical subject-matter background.

Familiarity with calculus will be assumed. Calculus will not be used extensively, but has a beneficial effect on mathematical maturity. SAS and JMP statistical computer software will be used extensively throughout the course. It is not expected that you will have used SAS or JMP extensively before. The computational aspects of the course are self-contained.

THE NATURE OF THIS COURSE:

Statistical machine learning (SML) is the application of computer software to learn from data. There are two main types: In *supervised learning*, there is an observed outcome that we try to predict. For example, we observe whether or not individuals in a group buy a product, and we use data from those individuals to try to predict whether other people will buy the product. The observed buying behavior provides the “supervision” that shapes a prediction rule. In *unsupervised learning*, there is no observed outcome but there are data for the individuals in the group. We analyze the group data to discover relevant structure of the group. For example, how many distinct consumer types are there? How can we identify who belongs in each type? What are the hidden motivations of consumers? How can we display the analysis visually to gain more insight?

This course is one of three related courses in your curriculum that are directly focused on statistical machine learning. In the summer, you took a course that surveyed supervised learning methods (*Introduction to Predictive Modeling*). In the fall, you went more deeply into supervised learning (*Advanced Predictive Modeling*). Also in the fall, you had some exposure to unsupervised learning methods in your other analytics courses, like *Marketing Analytics*. In the spring, this course, *Learning Structures and Time Series*, will dig deeply into the methods of unsupervised machine learning, plus time series modeling and forecasting. Time series is a supervised statistical learning methodology – but there is not enough room for it in *Advanced Predictive Modeling*. So this course is about unsupervised statistical machine learning for all kinds of data and about supervised learning for time-sequenced data.

The topics covered in unsupervised learning include data reduction, principal components, cluster analysis, factor analysis, multidimensional scaling and other data analytic techniques for understanding the structure of data when there is no outcome variable (no dependent variable). These methods all focus on making the data easier to understand by making the data simpler. Data may be complex, without being big, on account of complex internal relationships. But size complicates the task of learning from data. Data may be big because the number of rows of data is large. For example, there may be millions of clicks on internet ads. This is called a "big-*n*" problem. Data may be big because the number of variables is large. For example, there may be hundreds of measurements on each customer – everything from demographics (e.g., age, sex, residence) to past purchases in scores of different product categories, to click-throughs on hundreds of ads and time spent on each. This is called a "big-*p*" problem. And the problem could be both big-*n* and big-*p*.

The complexity of learning from data rises exponentially with the size of the data. This is known as the "curse of dimensionality." The contribution of unsupervised machine learning to managing the curse is to find simple structure (if it is really there) in the data so that big-*n* or big-*p* problems become, in essence, smaller-*n* or smaller-*p* problems when looked at in the right way.

The topics covered in supervised time-sequenced learning include modeling and forecasting of time series, autoregression, moving averages, ARIMA and other techniques. A typical problem for supervised time-sequenced learning is to forecast future sales, given past sales and other related past data. This part of the course will make extensive use of regression modeling that you will have studied in supervised statistical learning in your other courses. You will learn how to identify the right model for time-series data and forecast future values. You will also learn how to quantify the quality of your forecast by estimating its likely error.

In the balance between theory and practice, this course emphasizes practice more than theory. The goal is to make you a knowledgeable practitioner of the arts that you will be taught. You will have plenty of practice with real data sets. You will use SAS and JMP to analyze data. (Recruiters often request SAS; other courses in your MSBA curriculum provide experience with the R statistical software programming environment.) You will be taught to use SAS from the beginning without any knowledge or previous experience with the software being assumed. However, your introduction to SAS will be speedy, as it is assumed that you are computationally adept. But the models that underlie the statistical methods are mathematical. You must understand enough about the mathematical model to realize what the mathematical model permits you to do – and what it forbids you to do.

<u>TEXTBOOK</u>:

I have not found a textbook that satisfies all of my desiderata for this course:

- Focus on the main unsupervised SML methods;
- Coverage of time series models and forecasting;
- Theoretical-practical balance tipped toward the practical;
- Use of SAS.

Many books cover one or more of the above adequately. But no book even attempts to do all within two covers. Rather than require three or four books, only parts of each of which would be used in the course, I teach from notes that I will provide to you, and I recommend supplementary reading in selected books, which you may use as your motivation spurs and bank balance permits. Parts of the following textbooks are relevant for some portions or aspects of this course:

- *Multivariate Data Analysis* (7th edition, 2009), by Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolph E. Anderson – non-mathematical (algebra – and not much of that), no proofs, includes both supervised and unsupervised methods from the statistics perspective, presented clearly and explained well, applications-oriented. New version is *very* expensive. Earlier editions are probably good enough. eBay lists newest edition in international version (same text, different paper and format) for about $30 new.
- *Applied Multivariate Data Analysis* (2nd edition, 2010), by Brian S. Everitt and Graham Dunn – includes both supervised and unsupervised methods from the statistics perspective, applications-

oriented, many interesting datasets with plots and diagrams, math level is matrix algebra, some derivations. New $50-$100; used $15+ from online sources.

- *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning* (2nd editon, 2013) by Alan J. Izenman – surveys a vast number of supervised and unsupervised methods from a machine learning perspective, integrated with interesting datasets, sophisticated math used abundantly. About $70-$100+, new, from online sources.
- *Introduction to Time Series Analysis and Forecasting* (2008) by Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci – solid treatment of time series methods, especially regression and regression-related methods, applications-oriented but with some derivations, math level is matrix algebra. About $100 new from online sources.
- *The Little SAS Book: A Primer* (5th edition, 2012) by Lora Delwiche and Susan Slaughter – well-organized, well-written introduction to the way SAS works and to writing programs in SAS (still, it has 350 pages!), PROC REG is highest-level PROC. You will be a SAS expert if you learn this book well. About $50-$60 from online sources. Earlier (and cheaper) editions are probably good enough.

COURSE CONTENT:

I expect to cover 4-5 unsupervised learning techniques, selected from the following list:

- Information sufficiency
- Principal components
- Cluster analysis
- Factor analysis
- Multi-dimensional scaling
- Canonical correlation (**)
- Correspondence analysis

(**) Canonical correlation has aspects of both supervised and unsupervised learning.
I expect to spend about two weeks on each unsupervised technique and about six weeks on time series.

GRADING:

Four components of your work will be evaluated numerically:

Homework	200 points maximum
Quizzes	100 points maximum
Midterm Exam	300 points maximum
Final Exam	400 points maximum
COURSE SCORE (Total)	**1000 points maximum**

Your grade will be based entirely on your total points. At the end of the course, I will rank-order the COURSE SCORES from highest to lowest. I will then divide the ranked list into letter grade categories, based upon the level of mastery that I evaluate the points to represent. There is no predetermined COURSE SCORE that will guarantee an A. There is no predetermined grade distribution. The exact number of A's, B's, etc. in this course will depend upon the efficacy of your efforts.

HOMEWORK: Homework will be assigned at least once for every major topic covered. For the purpose of computing the homework portion of your COURSE SCORE, your scores on the homework sets collectively will be pro-rated to a 200-point scale according to the formula (*total points earned on all homeworks*) ÷ (*maximum possible points on all homeworks*) * 200. I require that your homework submissions be entirely your own individual efforts. In particular, no discussion or electronic exchanges of any type with other students are permitted on homework assignments. Please direct your questions about the homework to the TA or to the instructor.

LATE HOMEWORK: Homework is due on or before the date and time announced. Generally, homework will be collected at the beginning of class. Late homework will not be accepted except under two circumstances: (1) I have approved late submission in advance; (2) legitimate reasons beyond your control prevented both a timely submission and also timely prior approval. Such reasons include documented medical emergencies, but do not include last-minute software or hardware issues that has been exacerbated by your procrastination. You will have sufficient time to do homework that you should be able to complete it well in advance. Occasionally, a true emergency may make it infeasible to obtain my approval in advance for a late submission. In such cases, late submission may be accepted, at my discretion, if you provide acceptable written explanation and/or documentation as soon as feasible to explain the true emergency that caused the late submission. If you anticipate the possibility of an unavoidable delay, your safest course is to contact me in advance and explain your circumstances. If you cannot contact me in advance, then contact me as soon as possible. A late homework that is not accepted will be recorded as zero points. A penalty may be applied to a late homework that is accepted.

QUIZZES: Most classes will begin with a short 10-point quiz, administered online through Canvas. Your Total Quiz Score will be the sum of all of your quiz marks, except for your lowest three marks, which will be dropped automatically. For the purpose of computing the quiz portion of your COURSE SCORE, your Total Quiz Score will be pro-rated to a 100-point scale according to the formula *Total Quiz Score* ÷ (10 * *number of quizzes* – 30) * 100. If you miss a quiz by being absent or by being late to class, your mark on the quiz will be zero. In rare cases, additional drops or make-up quizzes may be permitted, at my discretion, upon presentation of acceptable written explanation for the absence or lateness. However, you must first use all three of your automatic drops before excuses for justifying additional drops or make-up quizzes will be considered.

EXAMS: The midterm exam may be scheduled for an evening in March to be arranged. Tentatively, I anticipate the midterm exam during an evening of the week prior to spring break. The final exam will be given at the date, time, and place published by the Business Analytics Program Office. The final exam will be comprehensive. For both the midterm and final exams, you may use a simple hand calculator and a limited amount of reference material (to be announced), but you may not use a computer.

SAS and JMP: I use statistical software to illustrate implementation of statistical theory. In this course I will use SAS (Statistical Analysis System) and JMP. For SAS, I will focus on the *free* web-based SAS OnDemand. JMP is a separate software product published by SAS Institute that is especially good for visualization and that integrates with SAS. For JMP, I will use JMP Pro 14 on a PC running the Windows 7 OS. At UT, you may buy software licenses for SAS and/or JMP to run on your PC for modest fees. I recommend buying the licenses for their convenience. SAS support for Macintosh native OS machines has been phased out. However, SAS may run on a Macintosh that is running Windows emulation software. You should feel reassured that the SAS language is essentially “the same” for all modes of use that you may encounter. Moreover, there is little difference for the user among recent versions of SAS. Please see the information about SAS that I have posted on Canvas for details.

COMPUTERS IN CLASS: I use a laptop computer extensively in class as a means to display data and analyses, and to show how to accomplish statistical tasks in SAS or JMP. Prior to each class, I will post on Canvas all of the files that will be used in that class. You may find it helpful to download these files and print them out and/or bring your laptop to class so that you can follow the class demonstrations and take notes. Having the files in front of you electronically or as printouts as we discuss them will maximize your learning.

CLASSROOM COURTESY:

- Turn off cell phones, pagers, and other noisy electronic devices before entering class.
- Mute the volume control on your laptop.
- Avoid surfing the internet or answering email in class.
- Avoid arriving late to class.
- Avoid unscheduled breaks.
- Respect the questions and opinions of other students as you would have them respect yours.

MISCELLANEOUS:

- Unless otherwise announced, you are responsible for material covered in class and on handouts, emails, or Canvas postings.
- It is unfair to allow a student to raise his/her score by submitting extra work unless all students are allowed the same opportunity. Therefore, extra work for extra credit will not be permitted.

ACADEMIC DISHONESTY: All students are expected to observe the UT Honor Code fully. Your responsibilities regarding the Honor System are described at http://deanofstudents.utexas.edu/sjs/spot_honorcode.php and associated links, which are incorporated herein by reference. I urge you to become familiar with this. If the application of the Honor System to this class and its assignments is unclear in any way, it is your responsibility to ask me for clarification.

STUDENTS WITH DISABILITIES: Upon request, the University of Texas at Austin provides appropriate academic accommodations for qualified students with disabilities. Services for Students with Disabilities (SSD) is housed in the Office of the Dean of Students, located on the fourth floor of the Student Services Building. Information on how to register, downloadable forms, including guidelines for documentation, accommodation request letters, and releases of information are available online at http://deanofstudents.utexas.edu/ssd/index.php. Please do not hesitate to contact SSD at (512) 471-6259, VP: (512) 232-2937 or via e-mail if you have any questions.

TENTATIVE CLASS SCHEDULE:

Assuming a semester of 30 class meetings of 90 minutes (1.5 hours) each, I anticipate approximately the following coverage of topics, approximately in the given order:

Hours	**Topic**
1.5	Introduction to unsupervised statistical learning
3.0	Review of basic statistics; big-*n* reductions
1.5	Introduction to SAS and JMP
6.0	Principal components analysis
6.0	Factor analysis
6.0	Cluster analysis
6.0	Multidimensional scaling
15.0	Time series analysis

Please interpret this loosely. The schedule is only a general guide. I expect to vary from it. Exams will take place outside of class time.