DS 7335 Machine Learning II

Course Overview

Through a combination of lectures and case analyses, students will learn how to think more theoretically about machine learning. Machine learning centers on the development and application of algorithms that give computer systems the ability to automatically learn and improve from experience without being explicitly programmed. Key tasks include data preparation, algorithm selection, training, and evaluation of the algorithm's predictions when applied to new data.

Rather than focus on approaches that prioritize key business outcome, this class is designed to target theoretical understanding. This is intended as a bridge between the educational experience students have had during their master's program and the continued work that is required after they exit the program. Examples of applied practice will motivate the conversation and reinforce the necessity of abstract thinking in the real world.

Course Designer

From an English/film undergraduate degree to a career in photojournalism and a PhD in statistical machine learning (AI) from the University of Chicago, Dr. John Santerre's career path can hardly be defined as straightforward. Uniting these disparate paths is his compulsion to explore, combining visual and analytical approaches that lead to places otherwise unobservable. Dr. Santerre's AI research is built upon this compulsion, focusing on the application of advanced machine learning tools in early-stage companies and organizations. This focus has allowed him to work with a variety of teams at Argonne National Labs, United Therapeutics, Ecolab, and SAP. Dr. Santerre's collaboration partners are often just starting their data science journey and are looking for insight and lessons learned from a wide variety of perspectives. With more than ten years in industry and having been involved with more than thirty data science initiatives (from proof of concept through production), Dr. Santerre's experience has allowed him to see both trends in friction points and insight into quick wins that can keep a data science team moving. Crucial to this process is the exposure Dr. Santerre gains by working with Southern Methodist University. He has designed the required machine learning class for the machine learning specialization, teaches said course, and is currently advising students on their capstone projects, all related to using machine learning for health care.

Course Student Learning Outcomes

#	Goal	Assessment	Delivery
1	**Name**, **describe**, and **categorize** dominant machine learning approaches, including regression, supervised and unsupervised classification, and deep learning.	**Self-check** and **class discussion** that checks for understanding, recall, and contextualization	**Lecture** on these topics
2	**Identify** opportunities for transitioning between data set representations.		
3	**Name**, **describe**, and **contextualize** key events, figures,		

	and constructs in the history of machine learning from the 1960s to the present.		
4	**Evaluate** business insights (BI) use cases and trade-offs for different algorithms.	Hypothetical client pitch exercise, possibly embedded in or comprising one week's **homework assignment**	**Interview** or **lecture** segments on common BI use cases and applications for machine learning
5	**Advise on** the size, shape, and quality of data sets for use in algorithm training.	**Self-check** and **class discussion** that asks for more critical thinking	**Lecture** and other video segments (such as **screen captures**) walking through some of the elements that feature in identifying and evaluating specific cases
6	**Recommend** different models by balancing implementation considerations (feature engineering) with other factors.		
7	**Apply** common supervised and unsupervised machine learning algorithms to a variety of data sets.	**Homework** problems and **final project** focused on the understanding of algorithms	**Skills practice** and **external resources alongside** homework/project **Q&A** in live session
8	**Assess** (quantify) how well different models perform on specific problems.		
9	**Appreciate** the complex interrelations among types of machine learning algorithms and their mathematical representations (graphical, probabilistic).	**Self-check** and **class discussion** that gets at this degree of understanding and appreciation	**Course-level map**, **lecture** content, and live session **discussion** focused on conceptual course content

Course Instruction Using Synchronous and Asynchronous Sessions

The course uses a combination of synchronous class sessions and asynchronous material and activities to teach students the course material and guide them through the learning process. Synchronous class sessions occur once per week during the course of the term. These sessions consist of lectures, discussions, interviews, problem solving, in-class assignments, and quizzes based on the asynchronous material, including the course video lectures, assigned activities and work, and any readings assigned. It is expected that all asynchronous material will be completed (e.g., videos viewed, assigned readings read, and assigned work completed and turned in) prior to the synchronous session associated with that material.

Course Prerequisite

A student taking MSDS 7335 must be enrolled in the master of science in data science program at SMU and must have completed the DS 7331 (Data Mining) course with a grade of C– or better.

Course Textbook and Other Course Material

Hal Daumé III, *A Course in Machine Learning*, http://ciml.info/.

Rowel Atienza, *Advanced Deep Learning with Keras* (Birmingham, UK: Packt, 2018).

Technology Requirements

DS 7335 is a course taught online with both synchronous and asynchronous portions requiring the transfer of video. Students are expected to have access to a computer with reliable, high-speed internet access. Students are expected to have access to a computer with a web camera and capable of running the required software to access the learning management system, read online documents, watch course videos, and participate in the synchronous classes (including being on camera). Students are also expected to have access to a reliable phone connection to participate in the synchronous classes.

DS 7335 course utilizes Python to complete assignments.

All students enrolled in SMU have an SMU email account. Notifications from the learning management system and from the course instructor use your SMU email account. Students are encouraged to check this email account regularly.

Course Access

This course is accessible to registered students in the SMU MSDS program only. Course asynchronous material, course information, and course communications occur through the 2DS learning management system. Access to the 2DS learning management system is available at https://2ds.datascience.smu.edu/.

Students who experience technical issues with the learning management system or the Zoom classroom should contact technical support as described below.

Students will have access to only those courses and course sections in which they are currently enrolled or have been enrolled in previous terms. Access to other sections is at the discretion of the section instructor. Access to recordings of synchronous sections where the student did not participate or was not an enrolled student is prohibited to protect the privacy of the students who do attend and participate.

Communication and Technical Support

Direct communications with the instructor should be made in the manner indicated by the instructor. General questions and questions that are relevant to multiple students—that is, questions that are not specific to an individual and involve that individual's private information—should be posted on the course wall.

Technical support for the learning management system and the online classroom may be reached twenty-four hours a day, seven days a week via

- *Chat Support*: Click **Live Support** in the lower-right corner of the 2DS screen after logging in to the system to chat with a technical support representative. Chat support generally responds and engages in five minutes or less.

- *Phone*: Students should call 1-844-768-5637 (toll free) to speak with a technical support representative.
- *Email*: Email studentsupport@datascience.smu.edu to initiate a support request with a technical support representative.

For other questions or concerns, please contact the appropriate SMU department or send an email to datascience@smu.edu.

It is the student's responsibility to ensure that all communications are received or acted upon.

Course Procedures and Policies

This course has a number of policies and procedures that students should understand and follow if appropriate. The following sections present the general course policies and procedures that students must follow. Additional policies and procedures may be given by the instructor. Please discuss as early in the term as possible with the instructor any questions or concerns that you may have regarding the course procedures and policies as defined herein or any additions made by the instructor to the course procedures and policies.

Course Grading Policy

This course consists of a number of assignments and projects that are to be completed throughout the term. Every submitted assignment is graded on a scale of 0 to 100 and contributes to the cumulative percentage for the course. Individual percentage breakdowns for each type of assignment are below. Questions regarding the grading of any assignments should be directed to the course instructor as soon as possible and in accordance with any regrading policy instituted by the instructor. This course is not graded on a curve. The required cumulative percentage needed to earn each letter grade is given in table 1.

Table 1. Cumulative percentage required to reach each letter grade

Cumulative percentage	Earned grade
[100–93]	A
[93–90]	A–
[90–88]	B+
[88–83]	B
[83–80]	B–
[80–78]	C+

[78–73]	C
[73–70]	C–
[70–60]	D
< 60	F

The cumulative percentage for the course is determined by the course assignment components with their corresponding percentages defined in table 2.

Table 2. Grade components and weightings of the cumulative percentage

Percentage of cumulative percentage	Component
15 percent	Asynchronous Video Response Questions and Discussion
40 percent	Homework Assignments
30 percent	Final Project Presentation
15 percent	Class Participation

Asynchronous Video Response Questions and Discussion (15 percent): Throughout the videos are various concept-check questions to make sure the student understands the material before moving on to learn new material. These questions are often in the form of multiple-choice or matching questions. These questions may also be discussion questions in which the student will respond to a prompt and then be able to see all other students' responses after they submit their response. At that point, it is our hope that a discussion will ensue. The student has the option to keep the conversation going by responding to their peers' responses. The instructor will be checking for participation in these discussions and may even participate in the discussion themselves. Given the fact that most of the material in this course builds on the material presented before it, participation in the concept-check questions and the discussions must be completed during the week they are assigned.

Homework Assignments (40 percent): This is really where learning is accomplished. It is critical to practice using, applying, and interpreting the models presented in this course, and these assignments are a big part of that practice. Each week you will have a list of readings and assignments to complete before the live session. They will be given a completion grade based on the thoroughness of the student's responses. Python Scripts will be distributed during class.

Final Project Presentation (30 percent): The students will record a presentation of the analysis of their question(s) of interest. This will require students to record a presentation of their PowerPoint presentation (not to exceed 20 minutes) with screen capture software. The PowerPoint slides must be uploaded to the

2DS learning management system, and the presentation must be uploaded to Vimeo to share with the class.

Class Participation (15 percent): Class participation is of paramount importance. Please have your video camera on and ready to engage!

A course grade of Incomplete (I) will be given only in the case of extraordinary circumstances that prevent the student from finishing the semester. Students must have completed at least 50 percent of the course with a passing grade to be eligible for an Incomplete grade.

Course Synchronous Session Schedule

Week	Topic	Deliverable	Readings (complete before unit's videos)
1	Introduction to Three Matrix Structures		http://ciml.info/dl/v0_99/ciml-v0_99-ch12.pdf
2	Geometric Interpretation of ML		http://ciml.info/dl/v0_99/ciml-v0_99-ch03.pdf
3	What's in an Average?		http://ciml.info/dl/v0_99/ciml-v0_99-ch07.pdf
4	Probabilistic Approach to Machine Learning	**Assignment 1 due**	http://ciml.info/dl/v0_99/ciml-v0_99-ch09.pdf
5	Bayesian Approach		http://ciml.info/dl/v0_99/ciml-v0_99-ch16.pdf
6	Decision Making with Matrices		https://en.wikipedia.org/wiki/Analytic_hierarchy_process
7	Linear Algebra as a Transformation	**Assignment 2 due**	
8	Compressing and Expanding the Matrix		http://ciml.info/dl/v0_99/ciml-v0_99-ch11.pdf
9	Graph Representation		
10	Deep Learning Introduction	**Assignment 3 due**	http://ciml.info/dl/v0_99/ciml-v0_99-ch04.pdf
11	Neural Networks: Learning a Transformation Matrix		http://ciml.info/dl/v0_99/ciml-v0_99-ch10.pdf
12	When AI Breaks		
13	Bias in Machine Learning	**Assignment 4 due**	

14	Deep Learning Code Walk-Through		http://ciml.info/dl/v0_99/ciml-v0_99-ch05.pdf
15	Outside View!	**Final project due**	http://ciml.info/dl/v0_99/ciml-v0_99-ch08.pdf

*You can also find a list of optional enrichment material for each week at the end of this document.

Grade Grievance Policy

Students are responsible for saving all graded materials as evidence in case of a discrepancy with the assigned grades. Students are responsible for ensuring that all grades are correctly reflected on the grade store. Any identified discrepancies should be brought to the attention of the instructor as soon as the discrepancy is found.

Refer to the university catalog for the university policy and process for grade grievances.

Assignment and Collaboration Policy

Data science is an inherently collaborative subject, and learning often occurs best when subjects are taught both to and from peers. Collaboration is expected to occur both in learning the course material and in performing the course work. However, each student must hand in their own work performed by themselves unless explicitly allowed by written directions given by the instructor. Collaboration means helping one another learn the material. Collaboration does not mean copying answers from one another. A good process is to ask questions and have discussions in groups and to always write up answers alone.

Assignment submissions that contain substantially the same answers shall receive a grade of zero on the first instance and a course grade of F upon a second instance. In order to mitigate potential issues and questions of similarity, peers with whom a student collaborates should be clearly identified by that student in their submissions.

Scholarly Expectations

Work submitted at the graduate level is expected to demonstrate critical and creative thinking skills and be of significantly higher quality than work produced at the undergraduate level. To achieve this expectation, all students are responsible for giving and receiving peer feedback of their work. Students are also expected to resolve technical issues, be active problem-solvers, and embrace challenges as positive learning opportunities. Data science professionals must be able to teach themselves and teach others to fill in any gaps in their knowledge or to find a way of learning new material that is most conducive to their learning style. Data science professionals must also be able to work cooperatively and collaboratively with others—skills that students are expected to practice in this course. Students are expected to ask questions and ask for help when they need it and to offer help when others are in need.

Absent questions or requests for assistance, instructors must assume that students understand the material being covered and are able to complete the assignments. It is primarily through your questions that the instructor learns where the students are struggling to understand and on which topics more time needs to be spent for the students' benefit.

Timeliness

Because a fifteen-week term goes by quickly, assignments must be submitted by the designated due dates. Full credit cannot be earned by late or incomplete assignments. Assignments may lose up to 10 percent of their possible value each day late if submitted after the posted due date/time (e.g., assignments can lose all of their value at ten days past due). When a project incorporates peer review, it is imperative that all projects be available at the beginning of the review period and that reviews are completed by the end of the review period so that others may incorporate feedback into project revisions. You will have plenty of notification and time to complete course assignments. If you know you are going to be out of town, involved in a special event/project, or unable to access a computer, please plan ahead. Also ensure that you have a backup plan ready in the event you lose power, internet access, or your available technology.

Time Commitment

For this technical graduate-level course, it is expected that students will spend between three and four hours beyond course instruction for each hour spent in instruction. MSDS courses are designed to have approximately three hours of course instruction, or contact hours, per week of the course. Therefore, it is expected that students will spend between twelve and fifteen hours per week on this course.

Attendance Policy

Attendance and on-camera participation at the weekly synchronous sessions in this course are mandatory. Students with more than three unexcused absences will receive a final grade of F for this course. It is the student's responsibility to notify the instructor if a synchronous session will be missed for either an excused or an unexcused reason at least twenty-four hours, or as soon as reasonably possible, prior to the synchronous session.

Drop Policy

Refer to the university drop policy for a complete description of the drop and withdrawal policies for this course.

Campus Concealed Carry

Concealed handguns are prohibited on the Southern Methodist University campus. Pursuant to section 30.06, Penal Code (Trespass by License Holder with a Concealed Handgun), a person licensed under subchapter H, chapter 411, Government Code (Handgun License Law), may not enter SMU property with a concealed handgun. Report violations to the Southern Methodist University Police Department by dialing 9-1-1 or 214-768-3388 (nonemergency) or 214-768-3333 (emergency).

Americans with Disabilities Act

Disability Accommodations: Students needing academic accommodations for a disability must first be registered with Disability Accommodations & Success Strategies (DASS) to verify the disability and to establish eligibility for accommodations. Students may call 214-768-1470 or visit http://www.smu.edu/alec/dass to begin the process. Once registered, students should then schedule an appointment with the professor to make appropriate arrangements. (See University Policy No. 2.4.)

Religious Observance

Religiously observant students wishing to be absent on holidays that require missing class should notify their professors in writing at the beginning of the semester and should discuss with them, in advance,

acceptable ways of making up any work missed because of the absence. (See University Policy No. 1.9.) Failure to notify your professor prior to your absence will result in an unexcused absence and possibly a grade of zero for any assignments.

Excused Absences for University Extracurricular Activities

Students participating in an officially sanctioned, scheduled university extracurricular activity should be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to make arrangements with the instructor prior to any missed scheduled examination or other missed assignment for making up the work.

Academic Integrity

It is the philosophy of Southern Methodist University that academic dishonesty is a completely unacceptable mode of conduct and will not be tolerated in any form. All persons involved in academic dishonesty will be disciplined in accordance with university regulations and procedures. Discipline may include suspension or expulsion from the university.

Scholastic dishonesty includes but is not limited to cheating, plagiarism, collusion, the submission for credit of any work or materials that are attributable in whole or in part to another person, taking an examination for another person, any act designed to give unfair advantage to a student, or the attempt to commit such acts. Example of academic dishonesty: In this course, students who have taken the course before or students from past cohorts may have the answers to some homework problems. It is considered academically dishonest to share solutions with anyone who is currently taking the course before the instructor posts the solutions for those students. It is also academically dishonest to accept solutions before a student's instructor makes them available. This falls under the category of presenting someone else's work as your own and is not only a serious violation of the SMU Honor Code but also severely detrimental to the student's understanding of the material. In general, if it feels the slightest bit wrong, it probably is. The safest thing to do is to consult your instructor with any questions before action is taken.

Students caught being academically dishonest shall receive a grade of F for this course and will be referred to the SMU Honor Council for a hearing and possible sanctions including a three-year mark on the student's transcript or expulsion. On a more positive note, our overwhelmingly main goal is to facilitate and foster each student's educational experience to enable them to achieve their academic and professional goals as a data scientist. Furthermore, these measures are aimed at "protecting your degree" in order to ensure that those with a master of science in data science from SMU have the utmost respect throughout academic and industry fields. This is our passion, and it is an amazing experience when everyone is working together and working hard. Let's get to it!

University Honor Code

When you signed your letter of intent to enroll in the MSDS program, you initialed the following statement:

"I have read and agree to abide by the SMU Honor Code available online at https://www.smu.edu/StudentAffairs/StudentLife/StudentHandbook/HonorCode."

The honor code is taken seriously at all levels within the university. Students who are found to have violated the honor code will be disciplined, which often includes expulsion from the university.

Plagiarism

Plagiarism is the “practice of taking someone else’s work or ideas and passing them off as one’s own” (this definition is from Google Dictionary). An example of plagiarism is as follows:

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.

The following is NOT plagiarism:

“A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables” (https://www.easycalculation.com/statistics/learn-regression.php).

The difference is in the punctuation and the attribution. Note that one can self-plagiarize. If you are using something that you wrote (e.g., a blog or a previously published article), please reference yourself.

DO NOT PLAGIARIZE. If you have any question as to what is and what is not plagiarism, ask your instructor. As a general rule, always use your own words and cite your source.

The consequence for being caught plagiarizing is to earn at least a zero on the identified assignment and may include earning a course grade of F and a referral to the SMU Honor Council for your honor code violation.

Best Practices for Success in the Course

Attendance. Take responsibility for your commitment. Attendance means not only being there for synchronous sessions but also participating in asynchronous work.

Citizenship. You need to be actively engaged to succeed in this class. Talking on cell phones, texting, “Facebooking,” tweeting, or leisure web browsing are prohibited in class. I consider these a disruption (not to mention rude).

Integrity. A lot of the graded work occurs outside of class, so I expect honesty and integrity in what you submit for evaluation. Evidence of academic dishonesty will minimally result in zeros for all involved parties and perhaps university-level disciplinary action. Don’t risk your career.

Humility. Don’t get lost! Ask questions in class. If something isn’t clear to you, it probably isn’t clear to others either. Questions may arise because I haven’t made a connection clear or have inadvertently left out an important point. Your question gives me a chance to explain more clearly. Don’t be proud or shy.

Organization. Don’t procrastinate! This is a technology-driven course. Count on your computer failing or your wireless connection breaking the night before a due date. Start early, and give yourself a chance to succeed.

Deadlines. You will generally have a week to complete an assignment. Due dates and times will be clearly indicated. Late submissions will be penalized, but it is much better to turn in work late than not at all (or to turn in incomplete/sloppy work). Work turned in after solutions have been posted to the course website will receive no credit.

Getting help. If questions arise while doing assignments/exams, do your best to resolve these questions before the assignment is due, first by taking time to seek answers yourself, next by asking questions on the wall, and finally via email to your instructor or other students. I encourage you and expect you to seek help. For questions during exams, please email the live session instructor directly.

Collaboration. I encourage the formation of study groups and collaboration with your fellow students in tackling the assignments. Working together in groups on homework is permitted, even encouraged. However, every student should write up and complete their homework independently. Talking about problems with other people does help in learning, but just copying the solutions from one another doesn't help!

Looks do matter! All assignments must be NEATLY executed and organized. You risk a zero on any assignment submitted in a sloppy manner. See submission guidelines for more detail.

Have fun! Learning is meant to be a fun activity. Although it can be difficult, time consuming, frustrating, and sometimes disappointing, always seek to find the fun in what you are doing and learning. The gratification from learning complex concepts and applying them to solve hard problems is what we are all striving to achieve. Having fun while we are learning and teaching others just makes the learning easier and friendships better.

Everything below is enrichment material.

Week 1
https://www.ams.org/notices/200902/rtx090200212p.pdf
https://bloomberg.github.io/foml/#lecture-1-black-box-machine-learning
https://bloomberg.github.io/foml/#lecture-2-case-study-churn-prediction
https://bloomberg.github.io/foml/#lecture-3-introduction-to-statistical-learning-theory
https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/
*Hastie and Tibshirani wrote one of the most important books on machine learning from a statistician's perspective. The video is based in R, but the teaching is absolutely top notch.

Week 3
https://bloomberg.github.io/foml/#lecture-4-stochastic-gradient-descent
http://ciml.info/dl/v0_99/ciml-v0_99-ch14.pdf
http://lineardigressions.com/episodes/2018/12/16/convex-and-non-convex-optimization
http://lineardigressions.com/episodes/2017/12/3/maximal-margin-classifiers

Week 4
http://ciml.info/dl/v0_99/ciml-v0_99-ch09.pdf
https://bloomberg.github.io/foml/#lecture-16-maximum-likelihood-estimation
https://bloomberg.github.io/foml/#lecture-26-gaussian-mixture-models
https://bloomberg.github.io/foml/#lecture-27-em-algorithm-for-latent-variable-models
LR pdf: http://theanalysisofdata.com/notes/logreg.pdf

Week 5
https://bloomberg.github.io/foml/#lecture-18-bayesian-methods
http://lineardigressions.com/episodes/2017/2/19/empirical-bayes
Bayesian: http://videolectures.net/mackay_course_10/

Week 8
http://lineardigressions.com/episodes/2017/12/10/the-kernel-trick-and-support-vector-machines
http://videolectures.net/mlss2012_scholkopf_kernel/

Kernel SVM: https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec13.pdf
Dim. reduction: http://videolectures.net/mlss2012_lawrence_dimensionality_reduction/

Week 9
http://lineardigressions.com/episodes/2017/6/4/pagerank

Week 10
https://bloomberg.github.io/foml/#lecture-28-neural-networks
https://bloomberg.github.io/foml/#lecture-29-backpropagation-and-the-chain-rule
http://lineardigressions.com/episodes/2017/3/5/a-sprint-through-whats-new-in-neural-nets
RNN: http://videolectures.net/deeplearning2016_bengio_neural_networks/

Yuan Lacun: https://www.youtube.com/watch?v=8zdo6cnCW2w&fbclid=IwAR3rmC-hGVY21E_ZDdZaGVGzb4xPSUV8N7zHnQryDd9nRo5L6mB2knL3AME

Week 12
http://lineardigressions.com/episodes/2018/11/18/limitations-of-deep-nets-for-computer-vision
http://lineardigressions.com/episodes/2017/8/27/adversarial-examples-for-machine-learning
Interpretability: http://videolectures.net/DLRLsummerschool2018_kim_interpretability/

Week 14
http://lineardigressions.com/episodes/2018/12/30/re-release-word2vec
GNN: http://videolectures.net/deeplearning2017_goodfellow_generative_models/

People
Jeffery Dean: http://videolectures.net/deeplearning2016_dean_deep_learning/
Diaconis: http://videolectures.net/nips09_diaconis_bamc/
Anna Gilbert: http://videolectures.net/nips2011_gilbert_analysis/
Getoor: http://videolectures.net/ecml07_getoor_isr/
Smola: http://videolectures.net/nipsworkshops2012_smola_parameter_server/
Bianchi: http://videolectures.net/mlss07_bianchi_onlle/
Avirm blum: http://videolectures.net/sicgt07_blum_atosf/
Leonardo: http://videolectures.net/aaai08_berlin_srl/

Linear algebra notes: https://ml-cheatsheet.readthedocs.io/en/latest/linear_algebra.html
Comp. Stats. with python: http://people.duke.edu/~ccc14/sta-663-2016/index.html

Podcast: https://www.superdatascience.com/podcast