Multivariate Statistics

PSYX 522 – Autumn 2019

Course Location and Time

Skaggs Building 303
Thursdays 2:00 – 4:50pm

Instructor Information

Instructor: Daniel J. Denis, Ph.D.
Office: Skaggs Building 369
Phone: N/A
Email: daniel.denis@umontana.edu
Office hours: Wed., 1-3, Fri., 11-12.

Course Overview & Expectations

It is assumed that students entering this course have taken previous applied graduate statistics courses, and have a basic understanding of statistics and statistical inference from early concepts through to linear models such as ANOVA and multiple regression. Familiarity with calculus and matrix algebra is also encouraged, though not required. Prior exposure to statistical software (e.g., through independent study or completing assignments in a course, etc.) would also be a benefit, though also not required.

Credits

3.0

Learning Outcomes

1. To provide you with the opportunity to obtain knowledge of various classical and modern multivariate statistical methods, supervised and unsupervised machine learning algorithms, and a **foundation** for further learning. As well, to demonstrate your understanding of techniques by obtaining software output and explaining such output.

2. To provide you with the ability to critically evaluate various multivariate analyses found in modern social and natural science literature, as well comprehend modern machine learning algorithms that you may come across.

3. To acquire an understanding of what is “under the hood” of modern “learning” techniques in AI and machine learning so you may situate these in proper context and gain a sense of maturity in understanding what these things actually mean at a deeper technical level. **What does it mean for a machine to learn?** By the end of the course, you will understand what this means.

Course Description

This course will survey topics which include requisite mathematics for multivariate statistics, regularization techniques in regression (e.g., ridge, lasso), discriminant analysis and support vector machines, principal components analysis, exploratory factor analysis and blind source separation, cluster analysis, neural networks, decision trees, among others, and the opportunity to apply many of these techniques using software. The key to understanding and using statistics is to be able to rely on your knowledge of **fundamental concepts** so that you may learn a variety of statistical procedures that you may need (or read) in your career. **The wealth of quantitative methods in current existence could easily take many lifetimes to master in entirety.** This course will help instill the fundamentals, so that upon completing it you are in a good position to learn any techniques you wish, now or into the future. Lacking such a foundation would make learning and interpreting (on a moderately deep level) new techniques virtually impossible, and you'd likely never see the wider forest.

Course Depth vs. Breadth

This course is necessarily a "breadth" course, as it is impossible to cover all of multivariate statistics in depth in the amount of time allotted for this course. For each of the multivariate or machine learning procedures that exist, there are many BOOKS written on these individual topics, and countless peer-reviewed journal articles. It is unreasonable to think that this course alone will make you an "expert" on any of the various multivariate techniques. Each research scenario, job, and data analysis is different (design issues are usually extremely difficult to figure out), and "cookbook" approaches to statistical analysis, even if somewhat helpful and having their place as a learning tool, can be dangerous if they are not used with caution. This course will introduce you to the underlying technical details of these procedures, so that you have some background on the "anatomy" of multivariate analysis before attempting to apply it to problems in research. Most, if not all, of any statistical methods you will likely encounter in your career are based on the same fundamentals studied in this course. **Seek to understand the fundamentals.** If you master the fundamentals and grasp the "big picture," your ability to learn new things in the future will be unstoppable. If you lack that foundation and resort to seeing all quantitative methods as distinct with no unifying foundations, you will forever be lost in the trees! Aim to see the forest and you will have a solid foundation for future learning.

Required Texts & Sources

Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. [We will be using this text as the "authority" document in the class. This is a highly advanced and very technical text spanning 733pp. Our goal in using it will be to employ it primarily as the authoritative reference (note: the software books below will not be used as authorities). We will not be covering near the complete book (e.g., we won't be dissecting proofs, etc.). As the Instructor, I will extract from Izenman what we require for the class to accomplish our goals and give you a solid, precise, and correct foundation in the subject of multivariate statistics and related areas (e.g., machine learning).

Chapter Class Notes with material drawn from sources such as Izenman and others. Chapter class notes will also contain some software applications using SPSS and R (and possibly Python in places).

Fox, J. (2010). *Appendices to Applied Regression Analysis, Generalized Linear Models, and Related Methods*. On-line document. Covers essential linear algebra and calculus (and a review of probability and statistics) that we require for understanding multivariate statistics and machine learning. I will select topics from this document for lecture and discussion in class (which will dominate most of September). It is a large document, and we will extract what is most important for understanding later class material.

Applications in Software

The following two books are useful for applying the statistical models and algorithms we discuss in class. It is recommended that you use them as **Python references** now and into the future as you gain more knowledge of multivariate statistics and machine learning applications. What these books are NOT are "stand-alone" rigorous textbooks, and should not be read or treated as such. These are NOT "source books." They are primarily software manuals. Though technically Géron can be read in order like a book, the early chapters especially are simply a spattering of ideas to get you started, with very little explanation of what is actually going on (other than installing the software). After installing Python, you are encouraged to try out some code and ideas in Chapters 1-3 just to get comfortable using Python, but don't worry too much about understanding the output for now. Chapter 4 and onward is a bit more systematic and those chapters can be read to help elucidate ideas we are discussing in class. Géron and VanderPlas may also be used for student seminar material at the end of the course, and are definitely books you'll want to keep as references if you work in data analysis or data science for the rest of your career.

Though these software books can help elucidate some concepts discussed in class, the Izenman text as well as class handouts are the "authority" documents of the class. Software books are excellent if you already know what they are talking about, hence use them primarily for code ideas now or into the future. They can also be used in places to help elucidate class concepts, but very seldom do software books do justice to the underpinning mathematical ideas. Rather, they give an "approximation" to the concept, which is fine, so long as we realize it as such. **The Izenman text is our most precise document for this course and for understanding concepts and the mathematics underlying them.** Beyond this is philosophical inquiry into the methods, which we will not get into (though to some extent we will have to face some of the philosophical issues, and will discuss in class – we cannot avoid a philosophical discussion of factor analysis as it is used in psychology, for example). **Why the rigor? Why the math FIRST?** So that you see what is "under the hood" of labels we give to statistical procedures and concepts (e.g., Artificial Intelligence, Machine Learning, etc.) so that you may learn what these things semantically refer to in reality. Multivariate analysis and associated machine learning tools are based primarily on calculus, matrix algebra, and fundamental statistics. **You cannot understand the topics of this course without an understanding of its mathematical foundations. It is impossible to do so. The very "core" of many of the procedures in this course is based on essential math theory.**

Class assignments will direct you to particular pages of the Python sources (as well as possibly other sources) to apply what we are learning in class using software. What's the difference between these two books? Géron attempts to include theory as well as software routines, whereas VanderPlas sticks to software code alone. Typical of software books, they sometimes cover topics in 1/2 page that theory books cover in a chapter. That's fine, so long as you first understand the theory (i.e., so long as you've consulted books like Izenman or similar, etc., first). Many topics in Géron however are impossible to understand (or teach) without deeper theory (even some of that theory isn't covered in Izenman because he avoids those topics altogether likely because he didn't feel he could give them adequate treatment in only a few pages). In this course, we cover primarily ideas and concepts first featured in Izenman (or class notes). You should never apply a statistical method or algorithm in software before first studying it and clearly understanding it. Applications in software and "hacking" follow.

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow. O'Reilly.

VanderPlas, J. (2017). Python Data Science Handbook. O'Reilly.

Software documents will also appear on Moodle providing further options for applying these models to software. I will announce when new postings appear. These can be used as well in completing the two major assignments for the course, as well as for seminars (which is a major component of the course).

***** Software Issues in Python** – if you experience software issues, feel free to e-mail me with a description of it before you spend too much time trying to resolve it. I may be able to help you much more quickly (or I will know where to look to figure out the issue). E.g., "I'm trying to fit this code, but for some reason it isn't working and gives me this error."

Spyder Python is relatively easy to use *when everything is working fine.* When there is a problem (usually a computer problem), it can be extremely frustrating, and it may take some problem-solving on your part and mine to figure out the issue. There are also quirks in our software books that I've had to dig up on, likely because of operating system differences. Any designated areas of software application that I refer you to in these books I will do my best to make sure those quirks are resolved and will send you updates via e-mail on any changes that need to be made to existing code. Though any IDE in Python can theoretically be used, in this class I will be discussing and referring to Spyder in assignments and e-mail communications (and handouts posted to Moodle).

Optional Texts & Resources

Johnson, R. A. & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. New Jersey: Prentice Hall.

Rencher, A. C. & Christensen, W. F. (2012). *Methods of Multivariate Analysis*. New York: Wiley.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research*. Sage publications: London.

Hays, W. L. (1994). *Statistics*, 5th ed. Wadsworth Publishing Company, Belmont CA.

Kirk, R. E. (2008). *Statistics: An introduction*. Thomson/Wadsworth: Belmont, CA.

Field, A. (2009). *Discovering statistics using SPSS*. Sage Publications: California.

Upton, G., & Cook, I. (2006). *Oxford Dictionary of Statistics*. Oxford University Press. New York.

Morgan, G.A., Leech, N. L., Gloeckner, G. W. & Barrett, K. C. (2011). IBM SPSS for Introductory Statistics: Use and Interpretation, 4th ed. Routledge: New York.

Leech, N. L., Barrett, K. C. & Morgan, G. A. (2011). IBM SPSS for Intermediate Statistics: Use and Interpretation, 4th ed. Routledge: New York.

Office Hours

Office hours are held weekly. You are also strongly encouraged to e-mail questions to the instructor as they arise. Writing your question out in an e-mail, as clearly as you can (even if very long) is an **excellent** way to clarify what you do not understand, and often, you achieve a deeper understanding of the topic itself. Please be as detailed and specific as you can in your e-mail so I know how to frame my response to best suit your needs.

Evaluation

Your final grade will be based on the following:

1. **Assignments (25%)** – graded on a binary scale (0, 1) – if you did the work, you get the credit (if you didn't or made a weak attempt, no credit – feedback will be provided on assignments, and you are also welcome to speak to the instructor at any time in person or e-mail for further feedback). End of class problems (or issues to think about) will also regularly be given. You will be expected to come to class next meeting prepared to discuss in rountable format.
2. **Student Seminar (25%)** *** Major Course Component
3. **Final Exam (50%)** 1/2 Theory / 1/2 Application

Student Seminar

Each seminar will be approximately 30 minutes (you speak for 30, discussion for 5). Seminars are an EXCELENT way to learn new topics, which you'll need to do for the rest of your career if you continue on in research or quantitative sciences. Seminars will be evaluated using the following criteria:

- Topic Knowledge & Expertise (30%)
- Level of Difficulty, Complexity and Depth (30%)
- Presence and Clarity of Exposition (20%)
- Organization, Delivery, and Thought Process (20%)

Letter Grade Distribution

Percentage	Grade	Percentage	Grade	Percentage	Grade
100	A	79	B +	59	D +
99	A	78	B +	58	D +
98	A	77	B +	57	D +

Percentage	Grade	Percentage	Grade	Percentage	Grade
97	A	76	B	56	D
96	A	75	B	55	D
95	A	74	B	54	D
94	A	73	B	53	D
93	A	72	B -	52	D -
92	A	71	B -	51	D -
91	A	70	B -	50	D -
90	A	69	C +	< 50	F
89	A -	68	C +		
88	A -	67	C +		
87	A -	66	C		
86	A -	65	C		
85	A -	64	C		
84	A -	63	C		
83	A -	62	C -		
82	A -	61	C -		
81	A -	60	C -		
80	A -				

Course Guidelines & Policies

Disability Modifications

The University of Montana assures equal access to instruction through collaboration between students with disabilities, instructors, and Disability Services for Students. If you think you may have a disability adversely affecting your academic performance, and you have not already registered with Disability Services, please contact Disability Services in Lommasson Center 154 or call 406-243-2243. I will work with you and Disability Services to provide an appropriate modification.

Academic Misconduct

You are expected to adhere to the university's Student Conduct Code with regard to academic integrity. Academic misconduct in this course will not be tolerated and will result in an academic penalty. **If you are suspected of cheating on a test or exam, you will receive zero on that test or exam and be asked to leave the class permanently**. In short, even if you do not know the answer to a question, you're much better off guessing than risking the chance of getting caught cheating.

Incompletes

Departmental and university policies regarding incompletes do not allow one to change "incomplete" grades after 1 year has passed since the "I" was granted.

Tentative Course Schedule (Subject to Change)

Date	**Topic**	**Primary Readings/Other**	**Secondary**
29 Aug.	Introductions, Syllabus, Course Preview & Big Picture	Article Readings	Ensure Access to Moodle Confirm You are on Mailing List
5 Sept.	Statistics, Calculus and Matrix Algebra for Multivariate	Fox Appendix	Install Python & Test Functionality
12 Sept.	Statistics, Calculus and Matrix Algebra for Multivariate	Fox Appendix	-
19 Sept.	Statistics, Calculus and Matrix Algebra for Multivariate	Fox Appendix / Izenman, Chap. 3	-
26 Sept.	Introduction to Spyder Python for Multivariate Statistics & ML	Izenman, Chap. 5	Géron, Chapter 4 (Training Models) & A#1 Assigned
03 Oct.	Regularized Regression (Ridge, Lasso)	Izenman, Chap. 5	Géron, Chapter 4 (Training Models)
10 Oct.	(Discriminant Analysis) and Support Vector Machines	Izenman, Chaps 8, 11	Géron, Chapter 5 (SVM) VanderPlas, Chapter 5, pp. 405-420
17 Oct.	Support Vector Machines	Izenman, Chap. 11	A#1 DUE
24 Oct.	Dimensionality Reduction (PCA)	Izenman, Chap. 7	Géron, Chapter 8 VanderPlas, Chapter 5 (pp. 433-445)
31 Oct.	Dimensionality Reduction (Latent Var., EFA)	Izenman, Chap. 15	A#2 Assigned
07 Nov.	Dimensionality Reduction (Cluster Analysis)	Izenman, Chap. 12 and Handout	VanderPlas, Chapter 5 (pp. 462-476)
14 Nov.	Intro to Artificial Neural Networks	Izenman, Chap. 10	Géron, Chapter 10 (pp. 257-278)
21 Nov.	Student Seminars (25%)	1. Tensorflow and Training Neural Networks 2. Decision Trees 3. Random Forests 4. Bagging (Bootstraping), Boosting (e.g., AdaBoost) 5. Python Environments (e.g., Shell, IDE, Jupyter, etc.) 6. Visualization in Python (e.g., Matplotlib) 7. Intermediate Python (e.g., advanced functionality, etc.) 8. SVM Regression 9. Softmax Regression A#2 DUE	
28 Nov.	**THANKSGIVING – NO CLASS**		
05 Dec.	Student Seminars (25%)		
TAKE-HOME	**FINAL EXAM (50%)**	All material covered in class is testable.	