

Statistical Machine Learning
ABE 6933 section STAT, 3 credit hours
Class Periods: T, R 3-4 (9:35 AM - 11:30 AM)
Location: 129 Rogers Hall
Academic Term: Fall 2021

Instructor:

Dr. Nikolay Bliznyuk

Email: n[myLastName]@ufl.edu; replace [myLastName] by instructor's last name without adding extra characters

Phone: 352-392-1864 (only by prior appointment)

Office Hours: 239 Rogers Hall, times TBA

Teaching Assistant/Peer Mentor/Supervised Teaching Student:

Please contact through the Canvas website

- Hanyu Qian; email: qian[taFirstName]@ufl.edu; replace [taFirstName] appropriately.

Course Description

Methodology and application of tools of statistical (machine) learning targeted at graduate students in engineering, applied statistics/biostatistics and quantitative life sciences. Statistical approaches to machine learning are emphasized in order to expand on and complement existing courses in engineering. Application and the intuition behind statistical methods rather than formal derivations and full mathematical justification of the procedures are prioritized.

Course Pre-Requisites / Co-Requisites

Prerequisites: calculus-based probability and statistics (e.g., EGN6446 or PHC6092 or STA5328) and applied linear regression (e.g., STA6093 or STA6166). Knowledge of scientific/statistical computing (e.g., in R, Python or Matlab) and undergraduate mathematics (multivariate calculus and linear algebra) appropriate for a graduate student in data sciences will also be assumed. See Supplement B for details and materials for review. A short pretest will be given on the first day of classes to determine if a student has prerequisites met.

Course Objectives

- Learn the language of and the principles behind predictive modeling and model validation
- Learn and be able to use R and Python to implement and apply different classes of statistical learning methods and interpret results
- Establish command of methods through homework, exams and a final project
- Reinforce the use of R as a statistical computing language for data science – for statistical inference, prediction, scientific computing and data visualization

Materials and Supply Fees

None

Required Textbooks and Software

- *An introduction to statistical learning* (referred to as the ISLR book)
- James, G., Witten, D., Hastie, T., & Tibshirani, R.
- 2017, 1st ed, 7th printing or later (available free at <https://faculty.marshall.usc.edu/gareth-james/ISL/>)
- ISBN 1461471370

- *The Elements of Statistical Learning* (referred to as the ESL book)
- Hastie, T., Tibshirani, R., Friedman, J.
- 2009, 2nd ed, most recent printing (available free at <https://web.stanford.edu/~hastie/pub.htm>)
- ISBN 0387848576

R/Python languages and appropriate computing environments (e.g., R Studio Anaconda) are freely available

Recommended Materials

N/A

Course Schedule

Week	Topics
1	Course logistics; introduction to SML; nearest neighbors
2	Calculus-based probability: essential review of rvs, independence, moments, etc
3	Mathematical statistics: essential review of estimation and hypothesis testing
4	Maximum likelihood estimation (MLE) and its connections with Bayesian inference
5	Linear regression essentials
6	Linear regression in matrix form and its extensions
7	Classification: logistic regression; discriminant analysis and its flavors
8	Out-of-sample performance metrics: cross-validation and ROC curves
	Take-home midterm exam
9	Model selection and regularization; penalized regression (ridge, lasso)
10	Spline-based models, generalized additive models
11	Classification and decision trees (single)
12	Ensembles of trees: bagging, random forest, gradient boosting
13	Support Vector Machines
14	Unsupervised learning: dimension reduction and clustering
15	Advanced topics (high-dimensional inference, multiple testing, deep learning, etc)
	Project report OR take-home final exam due

Depending on the typical student background, the above schedule may be adjusted to reflect the interests of the class (e.g., less time on the probability and statistics review and more time on advanced topics).

Attendance Policy, Class Expectations, and Make-Up Policy

Attendance at all times is expected. Excused absences must be in compliance with university policies in the Graduate Catalog (<http://gradcatalog.ufl.edu/content.php?catoid=10&navoid=2020#attendance>) and require appropriate documentation. Make-ups are not allowed except for documented health, family emergency or work reasons.

Evaluation of Grades

Assignment	Total Points	Percentage of Final Grade
Homework Sets (8)	100 each	25%
Quizzes (8)	100 each	25%
Midterm Exam	100	25%
Final Exam or Project	100	25%
		100%

Quizzes (online, in Canvas) and homework assignments will be closely matched to the course topics and will span approximately 4 hours of lectures. All quizzes and all homework sets will be weighted equally by converting the score to the 100-point scale first. Extra-credit opportunities to gain “bonus points” will be provided periodically.

Grading Policy

Tentatively, the following grading scale will be adopted; grades may be curved to the advantage of students.

Percent	Grade	Grade Points
93.4 - 100	A	4.00
90.0 - 93.3	A-	3.67
86.7 - 89.9	B+	3.33
83.4 - 86.6	B	3.00
80.0 - 83.3	B-	2.67

76.7 - 79.9	C+	2.33
73.4 - 76.6	C	2.00
70.0 - 73.3	C-	1.67
66.7 - 69.9	D+	1.33
63.4 - 66.6	D	1.00
60.0 - 63.3	D-	0.67
0 - 59.9	E	0.00

More information on UF grading policy may be found at:
<http://gradcatalog.ufl.edu/content.php?catoid=10&navoid=2020#grades>

Students Requiring Accommodations

Students with disabilities who experience learning barriers and would like to request academic accommodations should connect with the disability Resource Center by visiting <https://disability.ufl.edu/students/get-started/>. It is important for students to share their accommodation letter with their instructor and discuss their access needs, as early as possible in the semester.

Course Evaluation

Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available at <https://gatorevals.ua.ufl.edu/students/>. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, in their Canvas course menu under GatorEvals, or via <https://ufl.bluera.com/ufl/>. Summaries of course evaluation results are available to students at <https://gatorevals.ua.ufl.edu/public-results/>.

In-Class Recording

Students are allowed to record video or audio of class lectures. However, the purposes for which these recordings may be used are strictly controlled. The only allowable purposes are (1) for personal educational use, (2) in connection with a complaint to the university, or (3) as evidence in, or in preparation for, a criminal or civil proceeding. All other purposes are prohibited. Specifically, students may not publish recorded lectures without the written consent of the instructor.

A “class lecture” is an educational presentation intended to inform or teach enrolled students about a particular subject, including any instructor-led discussions that form part of the presentation, and delivered by any instructor hired or appointed by the University, or by a guest instructor, as part of a University of Florida course. A class lecture does not include lab sessions, student presentations, clinical presentations such as patient history, academic exercises involving solely student participation, assessments (quizzes, tests, exams), field trips, private conversations between students in the class or between a student and the faculty or lecturer during a class session.

Publication without permission of the instructor is prohibited. To “publish” means to share, transmit, circulate, distribute, or provide access to a recording, regardless of format or medium, to another person (or persons), including but not limited to another student within the same class section. Additionally, a recording, or transcript of a recording, is considered published if it is posted on or uploaded to, in whole or in part, any media platform, including but not limited to social media, book, magazine, newspaper, leaflet, or third party note/tutoring services. A student who publishes a recording without written consent may be subject to a civil cause of action instituted by a person injured by the publication and/or discipline under UF Regulation 4.040 Student Honor Code and Student Conduct Code.

University Honesty Policy

UF students are bound by The Honor Pledge which states, “We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honor and integrity by abiding by the Honor Code. On all work submitted for credit by students at the University of Florida, the following pledge is either required or implied: “On my honor, I have neither given nor received unauthorized aid in doing this assignment.” The Honor

Code (<https://sccr.dso.ufl.edu/policies/student-honor-code-student-conduct-code/>) specifies a number of behaviors that are in violation of this code and the possible sanctions. Furthermore, you are obligated to report any condition that facilitates academic misconduct to appropriate personnel. If you have any questions or concerns, please consult with the instructor or TAs in this class.

Commitment to a Safe and Inclusive Learning Environment

The Herbert Wertheim College of Engineering values broad diversity within our community and is committed to individual and group empowerment, inclusion, and the elimination of discrimination. It is expected that every person in this class will treat one another with dignity and respect regardless of gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture.

If you feel like your performance in class is being impacted by discrimination or harassment of any kind, please contact your instructor or any of the following:

- Your academic advisor or Graduate Program Coordinator
- Robin Bielling, Director of Human Resources, 352-392-0903, rbielling@eng.ufl.edu
- Curtis Taylor, Associate Dean of Student Affairs, 352-392-2177, taylor@eng.ufl.edu
- Toshikazu Nishida, Associate Dean of Academic Affairs, 352-392-0943, nishida@eng.ufl.edu

Software Use

All faculty, staff, and students of the University are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against University policies and rules, disciplinary action will be taken as appropriate. We, the members of the University of Florida community, pledge to uphold ourselves and our peers to the highest standards of honesty and integrity.

Student Privacy

There are federal laws protecting your privacy with regards to grades earned in courses and on individual assignments. For more information, please see: <https://registrar.ufl.edu/ferpa.html>

Campus Resources:

Health and Wellness

U Matter, We Care:

Your well-being is important to the University of Florida. The U Matter, We Care initiative is committed to creating a culture of care on our campus by encouraging members of our community to look out for one another and to reach out for help if a member of our community is in need. If you or a friend is in distress, please contact umatter@ufl.edu so that the U Matter, We Care Team can reach out to the student in distress. A nighttime and weekend crisis counselor is available by phone at 352-392-1575. The U Matter, We Care Team can help connect students to the many other helping resources available including, but not limited to, Victim Advocates, Housing staff, and the Counseling and Wellness Center. Please remember that asking for help is a sign of strength. In case of emergency, call 9-1-1.

Counseling and Wellness Center: <http://www.counseling.ufl.edu/cwc>, and 392-1575; and the University Police Department: 392-1111 or 9-1-1 for emergencies.

Sexual Discrimination, Harassment, Assault, or Violence

If you or a friend has been subjected to sexual discrimination, sexual harassment, sexual assault, or violence contact the **Office of Title IX Compliance**, located at Yon Hall Room 427, 1908 Stadium Road, (352) 273-1094, title-ix@ufl.edu

Sexual Assault Recovery Services (SARS)

Student Health Care Center, 392-1161.

University Police Department at 392-1111 (or 9-1-1 for emergencies), or <http://www.police.ufl.edu/>.

COVID-19

- You are expected to wear approved face coverings at all times during class and within buildings even if you are vaccinated.
- If you are sick, stay home and self-quarantine. Please visit the UF Health Screen, Test & Protect website about next steps, retake the questionnaire and schedule your test for no sooner than 24 hours after your symptoms began. Please call your primary care provider if you are ill and need immediate care or the UF Student Health Care Center at 352-392-1161 (or email covid@shcc.ufl.edu) to be evaluated for testing and to receive further instructions about returning to campus.
- If you are withheld from campus by the Department of Health through Screen, Test & Protect, you are not permitted to use any on campus facilities. Students attempting to attend campus activities when withheld from campus will be referred to the Dean of Students Office.
- UF Health Screen, Test & Protect offers guidance when you are sick, have been exposed to someone who has tested positive or have tested positive yourself. Visit the [UF Health Screen, Test & Protect website](#) for more information.
- Please continue to follow healthy habits, including best practices like frequent hand washing. Following these practices is our responsibility as Gators.

Academic Resources

E-learning technical support, 352-392-4357 (select option 2) or e-mail to Learning-support@ufl.edu.
<https://lss.at.ufl.edu/help.shtml>.

Career Resource Center, Reitz Union, 392-1601. Career assistance and counseling. <https://www.crc.ufl.edu/>.

Library Support, <http://cms.uflib.ufl.edu/ask>. Various ways to receive assistance with respect to using the libraries or finding resources.

Teaching Center, Broward Hall, 392-2010 or 392-6420. General study skills and tutoring.
<https://teachingcenter.ufl.edu/>.

Writing Studio, 302 Tigert Hall, 846-1138. Help brainstorming, formatting, and writing papers.
<https://writing.ufl.edu/writing-studio/>.

Student Complaints Campus: <https://care.dso.ufl.edu>.

On-Line Students Complaints: <http://www.distance.ufl.edu/student-complaint-process>.

Supplements:

A: a brief FAQ; B: prerequisites and materials for review; C: project information.

SUPPLEMENT A. A BRIEF FAQ:

1. I am an ABE grad student. Does this course count for the “applied statistics” or “math” requirement? *Yes for both.*
2. Is the class math heavy? *This is not a pure math class but math notation and logical reasoning will be used extensively to communicate ideas precisely and succinctly.*
3. Why do I need a previous stats class as a prerequisite? *To ensure that you know the basics and that you are positive that you want to learn statistical machine learning. One really needs some practice with regression to answer this question positively.*
4. Why do I need to know some matrix/linear algebra? How much do I need? *In order to save your precious time. Notation, basic operations and main decompositions is probably all that you’ll need. We’ll review these early in the course in the context of linear regression.*
5. Do I need to know how to program to take this class? *No, but you need to be comfortable with basic scientific/statistical computing; e.g., reading and writing scripts, expressing ideas in pseudo-code, summarizing results graphically, etc.*
6. I noticed from the syllabus that R will be the language/environment of choice; I have not used R but I am quite comfortable with Matlab/C/Java/Python. *Not a problem; you’ll be able to pick up basic R very fast. See the previous item and think of R scripts as pseudo-code. Advanced programming in R (environments, nonstandard evaluation, mixing R with other languages) is nontrivial but won’t be used in this course (unless you specifically want to use it for your project).*
7. I can read and write basic R scripts but I am still not very effective using this for anything above small scale. *Not a problem; assignments will involve only a modest amount of R; for a final project, other languages/environments will be allowed.*
8. I am a doctoral stats/biostats student. The course has only modest prerequisites; will I be bored? *Not at all; more ambitious/nontrivial options for homework and projects will be presented.*
9. I am not a doctoral stats/biostats student. Will I have to compete with stats/biostats doctoral students to get an A in the class? *Not at all; you will be mainly competing with your own self before this course.*
10. Will you be using Canvas? *Some Canvas + I’ll setup a Dropbox folder for all materials; details TBA*
11. Does this class require a lot of work? *This is a 3-credit graduate course. As such, expect to spend, on average about 10 hours of work per week outside of class (i.e., in addition to the lectures) in order to achieve course objectives. Students with deficiencies in prerequisites may need to spend more time.*

SUPPLEMENT B. PREREQUISITES & MATERIALS FOR REVIEW:

The following background is necessary to fully benefit from this course: basic undergraduate quantitative training (multivariate calculus and basic matrix/linear algebra); exposure to calculus-based probability and statistics; experience reading and writing simple programs in a programming language (ideally, in R or Python); a course in applied statistics (recommended).

Due to multiple inquiries from prospective students, I will highlight and elaborate on the prerequisites necessary in order to prepare and fully benefit from my SML course. These are as follows:

1. Experience reading and writing simple computer programs in a scripting language (ideally, in R or Python). Some of these skills will come from an intro statistics course (e.g., loading and exporting data, using R as a scientific/graphical calculator, basic visualization) but it is also important to be familiar with (a little) more advanced topics such as basic data structures including vectors/arrays, matrices, lists and data frames (in R), for/while loops and how to write simple functions.
2. Basic undergraduate quantitative training (multivariate calculus and basic matrix/linear algebra). Although we’ll be mainly using these tools for notational purposes, there is no way that one can understand statistics (modeling

and estimation) without calculus, and linear regression (and its extensions) or the principal component analysis without linear algebra.

If you need a refresher on multivariate calculus and basic linear algebra, completing the first two short courses on Coursera (as an auditor, free so long as certificates are not needed) for "Mathematics for ML" specialization (link below) was found helpful by other students.

<https://www.coursera.org/specializations/mathematics-machine-learning>

3. Exposure to calculus-based probability and statistics; the keyword here is "calculus-based". There will be a brief review of these topics but it won't be sufficient or aim at teaching these topics from scratch. If you have not had such a course and your plan of studies allows for this, I would recommend taking my other course (ABE6933: Probability & Mathematical Statistics via Scientific Computing, offered every two years, particularly this Fall) or EGN 6446: Mathematical Foundations for Applied Data Science or STA5328 prior to taking the SML course. Essential knowledge is covered in the Coursera courses #1 and #2 (hypothesis testing) in the specialization <https://www.coursera.org/specializations/advanced-statistics-data-science>

4. A recent first graduate statistical methods class (such as STA6093 or STA6166), or equivalent knowledge. In view of this course being one of the three required options for the first machine learning course in the Master's in AI Systems program, this course is no longer a hard requirement, but rather a strong recommendation. You will likely find this course more useful once you have mastered applied regression first. Essential knowledge is covered in the Coursera courses #3 and #4 in the specialization <https://www.coursera.org/specializations/advanced-statistics-data-science>

SUPPLEMENT C. Project description

The project will emphasize creative application of the methods developed in the course. Ideally, the application would be to your line of research and data (your own or of your immediate collaborators - advisor or fellow students). If you do not have suitable data, please check out the sources at the end of this description for the publicly available datasets. Otherwise, a good project could be a replication and extension of the results of a paper of interest that uses the methods from our course. "Creative application" does not allow merely running someone else's code without making other contributions. **Plagiarism is totally inappropriate and prohibited (just do not do it); it will result in a failing grade for the course. Course staff will run all project reports through UF Ithenticate. All work should be done individually (unless explicitly permitted by the instructor - for more ambitious projects).** *Projects already completed for other classes/causes are not acceptable. Example 1 - unacceptable: in a previous semester, a student wrote a paper for a journal or did a project for a different class, and now wants to submit it without major changes or additional SML type of work as the SML class project. Example 2 - acceptable: in a previous semester, a student wrote a paper for a journal or did a project for a different class, but wants to do a major extension of the work using the techniques learned in the SML class. This would make a potentially very good project, but the student needs to be explicit about what is new and what not. Only the new work will constitute the course project in this case.* The project will be used to assess the knowledge and skills that students acquired in the course; for that reason, the work must be done individually and without assistance from the course staff.

Deliverables: a one-page proposal and a short technical report as described below.

Deadlines: (tentative and will be revised appropriately each year)

12-November-2021, by 5:00PM EST (TBA): submit your proposal by email to nbliznyuk@ufl.edu if you are planning to complete the optional project, so that we can meet on Thursday during the class time (individual slots TBA after your proposals have been received).

14-December-2021, by 8:00AM EST (TBA): final report (in pdf format, accompanying code and the actual data that you used, if using a publicly available source; put all in a folder named after you and create a zip or rar archive; test archive before submitting), submit using Dropbox file request (same link as above): tinyurl.com/nbliznyuk-submit-files

Expectations for the proposal (1 page):

The main goal behind the proposals is to ensure that the projects are neither too simple nor too ambitious (i.e., will require about 30 hours to complete - loosely, an equivalent of 3-4 weeks of homework effort, where writing will

play a significant role), there is no duplication among students and that you have the necessary relatively clean data to analyze. Please specifically discuss what you propose to do (e.g., “big picture” and specific methods), why you focus on this particular problem (significance, motivation and relevance to the course) and available data (specifically, what are primary response variable(s) and features, what are n and p , etc). Your project should be “shovel ready”, i.e., a bit of data preprocessing may be necessary but you should not be spending more than 20% (ideally, 10%) of your total time budget on cleaning and data manipulation. The typesetting format of the proposal should be the same as for the project (please read below).

Expectations for the report (8 pages):

Report should be organized as a short paper appropriate to your field; e.g., a short abstract (100 words), intro (including motivation), background and data, methods, analysis/results, conclusions/discussion. *Any software/languages/environments may be used for the project (i.e., not necessarily R). In most cases, the project should use several classes of methods (multiple linear regression or logistic regression as the baseline – possibly coupled with variable selection; at least one shrinkage method (if p is high) and/or a GAM (if p is low), and a tree-based ensemble method, typically, random forest (and definitely discuss variable importance summaries that may be extracted from it) and possibly boosting (because it is often the best method), ideally both) for classification or regression and examine out-of-sample performance of the methods using K -fold cross-validation.* If you are doing binary classification, please additionally include and discuss the ROC curves. If doing non-binary classification, please showcase a binary classification subproblem and include and discuss the ROC curves. The length is about 8 pages (not counting references, appendices or supplements) double-spaced, using 12 pt font: roughly 6 pages of text and 2 pages for your most essential tables and figures; **single-column only**. If necessary, the paper may have an Appendix with additional figures and tables. Data, code and other supplemental information should be made available as part of “Supplementary Materials” unless the data are confidential (please discuss “deliverables” in the proposal). Please check out the project evaluation rubric in a separate file.

Some sources of data for projects:

Google search “data for machine learning”, e.g.,

UCI ML Repository

<http://archive.ics.uci.edu/ml/>

Kaggle Competitions

<https://www.kaggle.com/datasets>

PLOS One journal website

read info about data availability in papers of interest and at

<http://journals.plos.org/plosone/s/data-availability>