STATS 303

Statistical Machine Learning

2021-22 Spring Session 4

Dates / Synchronous meeting time:

Lecture:	MoTuWeTh 08:30 - 09:45	Recitation:	Th 19:00 - 20:00

Instructor's information

Dongmian Zou, Assistant Professor of Data Science dongmian.zou@dukekunshan.edu.cn

Office hour: Tu 11:15-12:15 Th 15:00-16:00, or by appointment

Dongmian Zou has a B.Sc. in Mathematics from the Chinese University of Hong Kong (2012) and a Ph.D. in Applied Mathematics from the University of Maryland, College Park (2017). Before joining Duke Kunshan, he served as a postdoctoral researcher at the University of Minnesota, Twin Cities from 2017 to 2020.

Dongmian is interested in mathematical aspects of data science. His primary research is in the intersection among applied harmonic analysis, machine learning and signal processing. Specifically, He is focusing on robust representations and structures in geometric and graph deep learning.

Dongmian loves cats, and he is better than any machine learning algorithms at distinguishing cats from dogs.

TAs

Eric Qu (recitation) zhonghang.qu@dukekunshan.edu.cn

Xue Chen (homework) xue.chen240@dukekunshan.edu.cn

What is this course about?

The impact of machine learning on modern society has been revolutionary in almost every discipline one could think about. Developing a theoretical understanding of the principles governing machine learning algorithms is of critical importance for their successful application.

Coming into this course, you have mastered a variety of machine learning tools in STATS 302 and have proudly achieved success in their application to real-world data. However, you may still have questions in mind that are yet to be answered. Can we interpret our models with statistical theory, and if yes, under what assumptions? How many data points do we need in order to be confident with a model learned from those data? How do we choose the most useful ones from a set of features? What do you gain from regularizing a linear regression? Why should you maximize the margin of a support vector machine? We will be answering these questions in this course.

This course aims to introduce the fundamental theory behind contemporary machine learning. In this course, you will develop understanding of the conceptual framework behind real applications of data science. You will learn how to analyze, understand and predict data from a statistical view. This course deals with not only the statistical methodologies, but also the assumptions, criteria, and evaluations behind those methods. You will also be revisiting some of the methods learned in STATS 302, but from a more theoretical point of view.

What background knowledge do I need before taking this course?

STATS 302 is the prerequisite for this course. You need to be familiar with concepts and techniques in Calculus, Linear Algebra and Probability (those are the prerequisites for STATS 302!). You should also have working knowledge of Python.

What will I learn in this course?

By the end of this course, you will be able to:

1. Understand and master the mathematical and statistical principles behind machine learning models and methods:
    a. Formulate statistical inference problems using both non-Bayesian and Bayesian approaches, and associate these problems with necessary optimization problems.
    b. Employ parametric methods for single- and multi-dimensional density estimation and application in classification and regression.
    c. Improve robust and interpretable methods using dimension reduction and feature selection.
    d. Adopt non-parametric methods including sampling and bootstrapping in case a parametric probability distribution cannot be obtained.
    e. Understand the capacity and limitations of machine learning methods. Apply the PAC learning framework to the analysis of statistical learning methods.
2. Evaluate and select machine learning methods for solving concrete problems in data science.
3. Apply and analyze statistical learning methods using standard libraries in Python.
4. Interpret results from statistical learning methods and communicate the results with both experts and non-experts in machine learning.

What will I do in this course?

- You will attend **lectures** in which I introduce the main concepts, principles and techniques.
- You will participate in **recitation** and we will explore concrete examples and state-of-the-art applications together. At the beginning of the course you will be assigned into groups (of 3~4). Each week, there is a worksheet for your group to work together on. During the recitation sessions, candidates from each group will present their answers and then the TA will lead the discussion.
- You will complete **homework** assignments. The problems in the homework can be mathematical, conceptual, or coding-related. You are encouraged to discuss among yourselves and consult with me for the homework, but you must write up the solution by yourself.
- You will work in your assigned group to explore a topic of interest that is not covered in class. I will suggest some references at the beginning of the whole session, but you can choose your topic beyond those (with my permission). During the last week of lectures, you will give a **presentation** on your discovery and what you have learned.
- You will take a **midterm** and a **final** written exam. They are both open-book and open-notes. However, you are NOT allowed to discuss among yourselves.

How can I prepare for the class sessions to be successful?

- Before the start of the whole session, you are encouraged to review what you learned in STATS302.
- Before each lecture / synchronous session, read or at least overview the assigned reading in the text. Think about the following question:
    - What is the current topic? What problem does it solve? Why is it important?
    - What mathematical/statistical techniques do I need to recall in order to understand this topic?
    - What application does this topic have? Does it relate in any way to what I learned in STATS302?
    - How would this topic relate to previous topics? What would be the difference?
- During each class, think actively and participate actively. Ask questions whenever you have any.
- After each class, review the notes and assigned reading. Make sure you understand both the intuition and the detailed techniques. Ask questions whenever you have any. Finish your homework and weekly journal in time.
- Before each exam, review the materials covered and the homework done. Ask questions whenever you have any.
- During the semester, meet regularly and talk with your teammates about the presentation. Work together with your groupmates for homework etc.
- Visit my office hour and use me as your source for learning. You are always welcome whether you have a question, need some help, or want to share an interesting idea with me.

What required texts, materials, and equipment will I need?

There is no required textbook because no single textbook covers the contents of this course. Lecture slides will be shared in a timely manner. The course materials are developed based on selected chapters from the following books.

- *Elements of statistical learning* by Hastie, Tibshirani and Friedman
    - Free e-textbook available at the official webpage: https://web.stanford.edu/~hastie/ElemStatLearn/
- *Pattern recognition and machine learning* by Bishop
    - Free e-textbook available at Microsoft webpage: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
- *Introduction to Machine Learning* by Alpaydin
    - E-textbook available from the Duke library: https://ieeexplore-ieee-org.proxy.lib.duke.edu/book/6267367
- *Understanding Machine Learning: From Theory to Algorithms* by Shalev-Shwartz and Ben-David
    - Free e-textbook available at the official webpage: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html

What optional texts or resources might be helpful?

The following books will be rarely referred to in the course, but they serve as a good complement.

- *Machine Learning: A Probabilistic Perspective*
    - E-textbook available: https://mitpress.mit.edu/books/machine-learning-1
- *Statistical Learning with Sparsity: The Lasso and Generalizations* by Hastie, Tibshirani and Wainwright

    - Free e-textbook available at the official webpage: https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS.pdf
- *An elementary introduction to statistical learning theory* by Kulkarni and Harman
    - E-textbook available from the Duke library: https://ebookcentral.proquest.com/lib/duke/detail.action?docID=697570
- *Foundations of Machine Learning* by Mohri, Rostamizadeh and Talwalkar (more advanced)
    - Free e-textbook available at the official webpage: https://cs.nyu.edu/~mohri/mlbook/

How will my grade be determined?

Activity	**Points**	**Comments**
Homework	20%	submit on Sakai; 6 in total, lowest score dropped
Presentation	10%	deliver during the last week
Midterm	30%	submit on Sakai; open-book
Final	40%	submit on Sakai; open-book

Please refer to the following scale for your grading.

A+= 98% - 100% **A** = 97% - 93%; **A-** = 90% - 92%; **B+** = 87% - 89%; **B** = 83% - 86%; **B-** = 80% - 82%; **C+** = 77% - 79%; **C** = 73% - 76%; **C-** = 70% - 72%; **D+** = 67% - 69%; **D** = 63% - 66%; **D-** = 60% - 62% **F** = 59% and below

What are the course policies?

Communications:

Course materials and announcements will be posted on Sakai regularly.

The best way to reach me is via email (I check and respond to emails regularly).

You are encouraged to communicate with your groupmates regularly.

Discussion Guidelines:

Civility is an essential ingredient for academic discourse. All communications for this course should be conducted constructively, civilly, and respectfully. Differences in beliefs, opinions, and approaches are to be expected. Please bring any communications you believe to be in violation of this class policy to the attention of your instructor. Active interaction with peers and your instructor is essential to success in this course, paying particular attention to the following:

- Be respectful of others and their opinions, valuing diversity in backgrounds, abilities, and experiences.
- Challenging the ideas held by others is an integral aspect of critical thinking and the academic process. Please word your responses carefully, and recognize that others are expected to challenge your ideas. A positive atmosphere of healthy debate is encouraged.
- Read your online discussion posts carefully before submitting them.

Homework:

Please deliver your assignment on or before the due date (late homework solutions will not be graded) on Sakai as a PDF file. You can either typeset your homework or scan / take a photo of a handwritten version. Points may be deducted for poorly presented solutions.

Academic Integrity:

As a student, you should abide by the academic honesty standard of the Duke Kunshan University. Its Community Standard states: “Duke Kunshan University is a community comprised of individuals from diverse cultures and backgrounds. We are dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Members of this community commit to reflecting upon and upholding these principles in all academic and non-academic endeavors, and to protecting and promoting a culture of integrity and trust.” For all graded work, students should pledge that they have neither given nor received any unacknowledged aid.

Academic Policy & Procedures:

You are responsible for knowing and adhering to academic policy and procedures as published in University Bulletin and Student Handbook. Please note, an incident of behavioral infraction or academic dishonesty (cheating on a test, plagiarizing, etc.) will result in immediate action from me, in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising). Please visit the Undergraduate Studies website for additional guidance related to academic policy and procedures. Academic integrity is everyone’s responsibility.

Academic Disruptive Behavior and Community Standard:

Please avoid all forms of disruptive behavior, including but not limited to: verbal or physical threats, repeated obscenities, unreasonable interference with class discussion, making/receiving personal phone calls, text messages or pages during class, excessive tardiness, leaving and entering class frequently without notice of illness or other extenuating circumstances, and persisting in disruptive personal conversations with other class members. Please turn off phones, pagers, etc. during class unless instructed otherwise. Laptop computers may be used for class activities allowed by the instructor during synchronous sessions. If you choose not to adhere to these standards, I will take action in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising).

Academic Accommodations:

If you need to request accommodation for a disability, you need a signed accommodation plan from Campus Health Services, and you need to provide a copy of that plan to me. Visit the Office of Student Affairs website for additional information and instruction related to accommodations.

What campus resources can help me during this course?

Academic Advising and Student Support

Please consult with me about appropriate course preparation and readiness strategies, as needed. Consult your academic advisors on course performance (i.e., poor grades) and academic decisions (e.g., course changes, incompletes, withdrawals) to ensure you stay on track with degree and graduation requirements. In addition to advisors, staff in the Academic Resource Center can provide recommendations on academic success strategies (e.g., tutoring, coaching, student learning preferences). All ARC services will continue to be provided online. Note, there is an ARC Sakai site for students and tutors. Please visit the Office of Undergraduate Advising website for additional information related to academic advising and student support services.

Writing and Language Studio

For additional help with academic writing—and more generally with language learning—you are welcome to make an appointment with the Writing and Language Studio (WLS). To accommodate students who are learning remotely as well as those who are on campus, writing and language coaching appointments are available in person and online. You can register for an account, make an appointment, and learn more about WLS services, policies, and events on the WLS website. You can also find writing and language learning resources on the Writing & Language Studio Sakai site.

IT Support

If you are experiencing technical difficulties, please contact IT:

- China-based faculty/staff/students 400-816-7100, (+86) 0512- 3665-7100
- US-based faculty/staff/students (+1) 919-660-1810
- International-based faculty/staff/students can use either telephone option (recommend using tools like Skype calling)
- Live Chat: https://oit.duke.edu/help
- Email: service-desk@dukekunshan.edu.cn

What is the expected course schedule?

Date	Class topic/unit name	Assignments due
Week 1	● Introduction ● Bayesian decision ● Parametric methods ○ Maximum likelihood ○ Parametric classification and regression	
Week 2	● Parametric methods ○ Model selection ○ Multivariate methods ● Nonparametric methods ○ Density estimation ○ Clustering ○ Mixture of Gaussian	• HW 1
Week 3	● Kernel methods ○ The kernel trick ○ Kernel PCA ○ RKHS ○ Gaussian Processes	• HW 2
Week 4	● Sampling methods ○ MCMC ● Information theory ○ Entropy ○ Maximal entropy	• HW 3
Week 5	● Learning theory ○ PAC learning ○ Uniform convergence	• HW 4 • Midterm exam
Week 6	● Learning theory ○ VC dimension ○ Fundamental theorem of statistical learning	• HW 5
Week 7	● More advanced topics (time permitting) ○ Complexity of learning ○ Rademacher complexities ○ PAC-Bayes ● Presentations	• HW 6
Final		• Final exam