DATA 319: Model-based and Data-based Methods for Data Analytics

Fall 2023

Course Logistics

Prefix/Number:	DATA 319.1 (WSU Everett); DATA 319.1 (WSU Pullman); DATA 319.2 (WSU Vancouver)
Credit Hours:	3
Prerequisites:	• DATA 219, CPT S 215, CPT S 223, or CPT S 233 *and* • MATH 220 or MATH/DATA 225 *and* • STAT 360
Classroom:	Room 340 (WSU Everett); SLOA 7 (WSU Pullman); VECS 125 (WSU Vancouver)
Class Time:	M/W/F 10:10-11:00 am
Final Project:	December 13 (Wednesday) 10:30 am-12:30 pm in the classroom

Instructor Information

Instructor:	Gani Nurmukhametov
Office:	WSU Everett Room 404
Phone:	(425) 405-1659
Email:	gani.nurmukhametov@wsu.edu
Office Hours:	typically[1] M/W 12:00-1:00 pm and Tu/Th 10:00-11:00 am or by appointment in Zoom

Introductory Note

First of all, welcome to DATA 319! I am looking forward to meeting you and helping out with your journey at Washington State University.

Secondly, sorry this is such a long document. Despite that, I hope you'll make time to read it through at least once during the first week of class as it contains a useful summary of the material we will be covering and will serve as a repository of important information and links.

Finally, definitely feel free to reach out if you have any questions or concerns about the course or material, I'm always happy to chat about data, math, or academic life and would welcome opportunities to offer my perspective or simply serve as a sounding board.

Course Description

This course provides an introduction to modern modeling methods for data analysis, focused on applications to high dimensional data. As datasets with hundreds or thousands of variables have become more common and easier to manipulate, an understanding of the relevant theoretical and practical concerns is an increasingly important part of data analytics. This focus on the nuances of high-dimensional data will allow us to explore the properties of many real-world data sets and gain hands-on experience with analytical methods implemented in the Python programming language. Throughout the course we will also build on student's foundational knowledge of mathematics and statistics to introduce multivariate concepts. The skills learned in this course are relevant to careers in industries specializing in large multivariate data sets, including (but not limited to!) data analytics, operations research, biological sciences and engineering.

Course Materials

We will use two main textbooks for this course:

1. Applied Multivariate Statistical Analysis by Härdle and Simar (ISBN-13: 978-3540030799)
2. Mining of Massive Datasets by Leskovec, Rajaraman, and Ullman (ISBN-13: 978-1108476348)

[1] I may occasionally change the time or cancel office hours on a given day. Please follow my announcements on Canvas.

You may also find these textbooks useful:

- Applied Multivariate Statistical Analysis by Johnson and Wichern
- Numerical Algorithms by Solomon

Software

Each course module incorporates computational examples working in the Python programming language. Prior Python experience at the level of DATA 219 or 302 will be assumed. These sections will reinforce the main material by allowing you to perform and implement the analyses discussed in the study materials and videos. Some homework problems and all of the projects will also require the use of computational methods and previous programming experience and knowledge of statistical software tools from prerequisite courses will be assumed. Throughout the course we will help students build familiarity and experience with standard analytics libraries including pandas, numpy, matplotlib, scikit-learn, keras, and tensforflow. Please get familiar with the following resources on Python:

- An Introduction to Statistical Learning with Applications in Python by James, Witten, Hastie, Tibshirani, and Taylor
- Data Science from Scratch: First Principles with Python by Grus
- Hands-On Machine Learning with Scikit-Learn and TensorFlow by Geron
- Python Data Science Handbook by VanderPlas

Expectations for Student Effort

You are expected to spend a minimum of 9 hours per week for a three-credit course, of which 3 hours are spent in instructor-lead activities (lectures and discussions) and 6 hours are spent in outside activities. These outside activities include, but not limited to: reading, studying, problem solving, writing, homework, and other preparations for the course. Achievement of course goals may require more than the minimum time commitment. For the most accurate and up to date information go to Academic Regulations.

Attendance and Participation

Due to the nature of the course (with students joining synchronously from several WSU campuses) and the wide range of topics that we will cover, daily attendance will be essential for your success. Although it is not officially a part of the course grade, missing class could adversely affect your grade by impacting your understanding of the material. That said, I understand that you may occasionally have difficulties keeping up with the pace of the course. Please communicate with me in advance if possible, so I can help point you to useful resources.

Class Communication and Instructor Interaction

We will use the Canvas forums for course discussions. This is a great place to ask questions from your peers, as well as to get feedback on your ideas. Announcements and other official communications will be posted on Canvas as well as sent to your official WSU email accounts. You should check these messages regularly to stay informed about upcoming due dates and updates to the syllabus.

I am accessible by email at gani.nurmukhametov@wsu.edu. I teach multiple courses, so please as a courtesy to me include "DATA 319(Sync)" in the subject line for any messages concerning this course. Most likely I will read your email as soon as I receive it, but it still may take me a little while to respond. I will commit to responding within 48 hours (usually I will get back to you much faster) but this does mean that queries sent immediately before a deadline may not receive substantive responses in time to be directly helpful, so please plan ahead ☺

I strongly encourage you to utilize the resource that is freely available to you, that is my office hours! If you cannot attend my office hours during their regular time, do not hesitate to send me an email and we can schedule a short appointment in Zoom that fits both you and me.

Learning Outcomes and Assessment

Students who successfully complete the course will be able to:

- Construct scripts for processing, analyzing, and visualizing high-dimensional data in Python
  - **Relevant Assessments:** All Module Assignments, Midterm Project, Final Project
- Conduct exploratory multivariate data analysis, including constructing relevant visualizations
  - **Relevant Assessments:** Module 1 Assignment, Midterm Exam, Midterm Project
- Apply and analyze the MapReduce framework and implement simple tasks using PySpark on Colab
  - **Relevant Assessments:** Module 1 Assignment and Module 2 Assignment
- Compute matrix factorizations and interpret the resulting outputs in the context of the original data
  - **Relevant Assessments:** Module 1 Assignment, Midterm Exam
- Use frequent itemset modeling and association rules to characterize relevant data
  - **Relevant Assessments:** Module 2 Assignment, Midterm Exam
- Apply the A Priori algorithm and its variants to frequent itemset problems
  - **Relevant Assessments:** Module 2 Assignment, Midterm Exam
- Understand and implement high-dimensional data embeddings in $\mathbb{R}^n$
  - **Relevant Assessments:** Module 2 Assignment, Midterm Project
- Select and apply relevant metrics for (dis)similarity on high-dimensional data
  - **Relevant Assessments:** Module 2 Assignment, Final Project
- Select and perform appropriate methods for dimension reduction, including PCA, MDS, Factor Analysis, and Discriminant Analysis
  - **Relevant Assessments:** Modules 2 and 4 Assignments, Midterm Project
- Define the characteristics and properties of the multivariate normal distribution
  - **Relevant Assessments:** Module 3 Assignment, Midterm Exam
- Characterize the concept of interdependence among multidimensional data through correlation
  - **Relevant Assessments:** Module 3 Assignment, Midterm Exam
- Implement appropriate multivariate null hypothesis testing
  - **Relevant Assessments:** Module 3 Assignment, Midterm Exam
- Perform, justify, and interpret the use of unsupervised analytic methods including clustering techniques
  - **Relevant Assessments:** Module 5 Assignment, Final Project
- Perform, justify, and interpret the use of supervised analytic methods
  - **Relevant Assessments:** Module 6 Assignment, Final Project
- Identify ethical concerns relevant to designing and implementing data models and analyze potential impacts using appropriate frameworks
  - **Relevant Assessments:** Weekly Reading Discussions, Module 7 Assignment, Final Project

Assignments and Assessments

There will be five main types of graded assignments in this course.

- **Reading Discussions:** In addition to participation in Canvas discussion posts, each week you will be responsible for completing one or more short readings on a relevant data analytics topic and responding to corresponding discussion questions. There will be short quizzes corresponding to each module covering the basic concepts from the lecture material that will also count in the participation grade.
- **Written Assignments:** For each module (roughly every two or three weeks), a problem set will be assigned. These will usually be a mixture of direct questions about the lecture material and opportunities for you to apply the methods we discuss to real data. Individual responses to the assignment will be due at midnight two weeks after they are posted. No late work will be accepted but at the end of the semester your lowest score will be dropped. Written assignments must be submitted as .pdf files.
- **Midterm Project:** Beginning in Week 6 you will work individually to use the tools discussed in class to implement a recommendation system and explore the concepts of dimension reduction.
- **Midterm Exam:** There will be a single exam during Week 8, covering the material that we will have encountered to that point. Both theoretical concepts and tasks in Python will be covered on the exam. The format of the exam to be announced later during the course.
- **Final Project:** Beginning in Week 12 you will work in groups of three or four to complete a project employing the multivariate methods discussed throughout the class on a large dataset. In addition to identifying a particular question (or questions) of interest to address, each group will identify several of the topics from the second half of the class that are relevant to the chosen dataset and perform an exploratory analysis using those techniques. The final submission will include a report describing the analytical work that was performed and the Python code used to process the data and resulting outputs. Each group will record a presentation of their findings.

Grading Policy

The "weights" of course assignments towards the overall course grade, and the grade scheme I will use to determine your letter grade are as follows:

Assignment	Percentage
Reading Discussions	10%
Written Assignments	40%
Midterm Project	10%
Midterm Exam	15%
Final Project	25%

Grade	Percentage	Grade	Percentage
A	95-100	C+	77-79.99
A-	90-94.99	C	73-76.99
B+	87-89.99	C-	70-72.99
B	83-86.99	D	60-60.99
B-	80-82.99	F	0-59.99

Late Work Policy

Late submissions are not normally allowed. However, earlier submissions are allowed at any time before due date. Extensions *may be* allowed by contacting me well in advance. If asking for an extension, do so on real emergencies—not as a habit. After "answer keys" have been posted to the class, that assignment or exam cannot be made up, so please reach out to me early if needed and don't get too far behind.

It is better to turn in a partial homework than no homework. The phrase "little is little but nothing is nothing" fully applies here.

For an excused missed midterm exam, the accommodation is at the sole discretion of the instructor and may include a reweighting of the remaining components making up the student's grade or taking a makeup exam. Again, please reach out to me well in advance.

Collaboration Policy

You are encouraged (and sometimes required) to work with other students for the assignments in the class. However, the work that you submit should be your own and in particular should be written in your own words and communicate your own understanding of the solution. If you do collaborate, please list the names of the other students you worked with on your submission. You may be asked to explain your work in person to obtain full credit. Obtaining solutions for course problems from external sources will be considered a violation of the academic integrity policy with consequences described below.

Online Discussion Policy

The essence of education is exposure to diverse viewpoints. In your discussion posts you'll meet students with vastly different opinions and backgrounds. You're encouraged to disagree with the substance of others' ideas and opinions but do so with an active sense of respect for one another, and without losing focus on the topic at hand. Personal attacks, inflammatory statements, flaming, trolling, and disruption of the discussion do not have a place in academic discourse. Postings must comply with University policy on use of computing resources, including those regarding harassment and discrimination, as well as conform to the WSU Community Standards.

I will aim to promote high-quality academic discussions by removing any posts I view as disruptive of the educational process and alerting students whose posts have been removed that they have violated course expectations. Students who continue to misuse the discussion boards after a warning may be subject to removal of access rights, course failure, and referral to the Office of Community Standards. Visit WSU Netiquette Guidelines.

Weekly Topics

The following outline describes the preliminary plan for our class. An updated version will be posted on the course Canvas page and updated throughout the semester.

Week 1:

- Course outline and introduction
- Linear algebra review
- **Textbook:** HS Chapter 2, MMDS Chapter 1

Week 2:

- Multivariate visualizations
- Random vectors and matrices
- **Textbook:** HS Chapter 1, MMDS Chapter 2

Week 3:

- Matrix factorizations
- Conditioning and stability
- **Textbook:** Solomon Section 2, MMDS Chapters 2 and 6

Weeks 4-5:

- Data embeddings (distance metrics)
- Curse of dimensionality
- Introduction to dimension reduction
- **Textbook:** HS Chapter 10, MMDS Chapters 3, 6, and 11

Weeks 6-7:

- Multivariate normal distribution
- Multivariate inference
- **Textbook:** HS Chapters 4 and 5, MMDS Chapter 9

Week 8:

- Inference versus data mining
- Machine learning formulations
- Midterm Exam
- **Textbook:** MMDS Chapter 9

Week 9:

- Dimension reduction
- **Textbook:** HS Chapters 10-13, MMDS Chapter 11

Weeks 10-11:

- Supervised learning
- **Textbook:** HS Chapter 20, MMDS Chapters 12 and 13

Weeks 12-13:

- Unsupervised learning
- **Textbook:** HS Chapter 13, MMDS Chapter 7

Week 14:

- Computational advertising
- Ethics introduction
- **Textbook:** MMDS Chapter 8

Week 15:

- Ethics of predictive analytics
- **Textbook:** Instructor materials

Final Exam Period: Final Project presentations

WSU Academic Calendar

Please refer to the WSU academic calendar to be aware of university holidays and improrant deadlines throughout the semester.

University Syllabus

Students are responsible for reading and understanding all university-wide policies and resources pertaining to all courses (for instance: accommodations, care resources, policies on discrimination or harassment), which can be found in the University Syllabus.

Lauren's Promise: WSU's Commitment to Address Discrimination and Harassment

On October 22, 2018, Lauren McCluskey, 21 years old, was murdered by a man she briefly dated on the University of Utah campus, where she was a student. Lauren was raised in Pullman, Washington. Together with her parents, who are professors at WSU, this university community stands firmly behind Lauren's Promise: **WSU will listen and facilitate support and reporting options if someone is threatening you**.

WSU prohibits discrimination and sexual misbehavior. Discrimination includes discriminatory harassment, sexual harassment, and sexual misbehavior. Sexual misbehavior includes stalking, violence between intimate partners, and all types of sexual violence. If you are in immediate danger, call 911.

If you have experienced or have witnessed discriminatory behavior, you can contact the WSU Office of Civil Rights Compliance & Investigation (CRCI) and/or the WSU Title IX Coordinator at 509-335-8288. These offices can give you confidential resources and explain your choices to report the behavior. (Go to crci.wsu.edu for more information).

See Policy Prohibiting Discrimination, Discriminatory Harassment, Sexual Harassment, and Sex and Gender Based Violence (Executive Policy 15) and WSU Standards of Conduct for Students (Chapter 504-26 WAC).

Reasonable Accommodations

Students with disabilities or chronic medical or psychological conditions can request reasonable accommodations. If you need reasonable accommodations to fully participate in your courses, please go to your campus' Access Center/Services website (see links below). Follow the procedures to request accommodations. You may also contact your campus office to schedule an appointment with an Access Advisor.

The Access Center/Services will notify your instructors of your requested accommodations, but you made need to communicate with your instructors about how some of your accommodations will work (by email, Zoom, or in person).

Contact an Access Advisor on your campus:

- Central Support (for Pullman, Global, Everett, Bremerton, and Puyallup campuses).
  Phone: 509-335-3417 or email: access.center@wsu.edu.
- Spokane Campus. Phone: 509-358-7816 or email: spokane.access@wsu.edu.
- Tri-Cities Campus. Phone: 509-372-7352 or email: tricities.accessservices@wsu.edu.
- Vancouver Campus. Phone: 360-546-9238 or email: van.access.center@wsu.edu.

Arrangements for Religious Reasons

Washington State University tries to accommodate students for religious reasons. Please reach out to me within the first two weeks of the semester to schedule examinations or other required course activities during the absence. You should include the specific dates of the religious activity. If I approve the absence, then your grade is not affected. However, you are still responsible for any course work required during the absence.

If you disagree with my response/decision, check Academic Regulation 104 - Academic Complaint Procedures. If you think your request was treated unfairly, contact the Office of Compliance and Civil Rights.

Emergencies on Campus

To receive emergency alerts on your phone or by email, click on the link to the page of your campus below. These alerts may include information about active shooter situations and severe weather.

In case of an active shooter, follow these ideas: "Run, Hide, Fight" .

In any emergency, remain ALERT by observing and paying attention to WSU emergency alerts. ASSESS your specific situation, and ACT to ensure your own safety and the safety of others if you are able.

- Bremerton Campus
- Everett Campus
- Pullman Campus
- Spokane Campus
- Tri-Cities Campus
- Vancouver Campus

Student Support Resources

WSU wants you to succeed. When problems happen, it is important to get help early. The Student Care Network has links to resources for each campus. For more resources for physical and emotional health, financial, legal, academic, and other support, visit Campus Resources and Support. Each WSU location also has a Student Care Team. The team includes professionals who can recommend resources and services to help you succeed.

Other Relevant University Policies and Statements

Academic Integrity

Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course.

You are responsible for reading WSU's Academic Integrity Policy, which is based on Washington State law. Students who violate the Academic Integrity Policy will fail the assignment, will not have the option to withdraw from the course pending an appeal, and will be reported to the Center for Community Standards. Multiple violations of the policy will cause you to fail the course.

Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students (identified in Washington Administrative Code (WAC) 504-26-010(2)). You need to read and understand all of the definitions of cheating. If you have any questions about what is and is not allowed in this course, please reach out to me before proceeding.

If you wish to appeal my decision relating to academic integrity, please use the form at the Center for Community Standards website. You must submit this request within 21 calendar days of the decision.

Incomplete Grade Policy (Academic Rule 90h)

Incompletes are granted only with my permission and are subject to the following guidelines:

1. You must request an incomplete in writing or by e-mail to me before the end of the semester. This request must be signed and dated by you (or identified by your WSU e-mail address) and must explain the reasons behind the request for the incomplete.
2. In order to be considered for an incomplete grade, these two consditions should be met:
   - You must complete a minimum of 75 percent of the assigned course work.
   - You must have a mathematical possibility of scoring a 60 percent or above for the entire course.
3. If extraordinary circumstances (e.g., family emergency, serious illness) are involved, I retain the the discretion to grant an incomplete even if the minimum conditions outlined in item 2 above are not met.

If an incomplete grade is granted, the standard WSU policy applies (i.e., ALL work must be completed within one full year from the end of the enrollment semester at issue, unless a shorter time is specified by the instructor. Otherwise, an automatic grade of "F," or failing, will be entered on your transcript).

Library Support

All students enrolled in Washington State University online courses can use the WSU Libraries online databases and receive reference and research assistance from their home campus. Students can also borrow books and other circulating material as well as access full-text journal articles.

General Library Support Links:

- Global Campus
- Pullman Campus
- Spokane Campus
- Tri-Cities Campus
- Vancouver Campus
- College of Nursing

Online Tutoring

As a WSU student enrolled in an undergraduate course, you have FREE unlimited access to Online Tutoring. This is not a course requirement, but a resource for you to utilize as needed.

With three ways to access a tutor you can choose the one that best fits your needs:

- Submit a paper: Writing Lab tutors will respond to papers in ANY academic subject. Just submit your paper, ask specific questions on the submission form, and a tutor will respond within 24-48 hours.
- Live tutoring: eChat rooms allow students to meet with tutors in one-on-one tutoring sessions via a fully interactive, virtual online environment.
- Leave a question: Students can leave specific questions for a tutor in any of our subjects by taking advantage of our eQuestions option. Our tutors will respond to your question within 24-48 hours.

More details and the list of available tutoring subjects can be found at www.eTutoringOnline.org

Copyright

Any course-related materials, presentations, lectures, etc. are the instructor's intellectual property and may be protected by copyright. The use of University electronic resources for commercial purposes, including advertising to other students to buy notes, is a violation of WSU's computer abuses and theft policy (WAC 504-26-218). Selling class notes through commercial note taking services without written advance permission from the faculty, could be viewed as be as copyright infringement and/or academic integrity violation(WAC 504-26-010 (3)(a,b,c,i)).