**AMS 380: Data Mining**

**Department of Applied Mathematics & Statistics**

**Stony Brook University**

**1. Course Description:**

This course will teach the basic ingredients of classical and contemporary statistical data mining methods including dimension reduction, variable selection, pattern recognition, and predictive modeling using traditional general linear models and generalized linear models, and modern statistical learning methods such as classification and regression tree, random forest, neural networks, etc. We will also teach how to run these procedures with the statistical programming language R.

**2. Text book:**

**An Introduction to Statistical Learning** **with Applications in R**. (2017). Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. **Springer.** (<http://faculty.marshall.usc.edu/gareth-james/ISL/book.html>)

**Data Mining with R: Learning with Case Studies**, Second Edition (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series), 2017, **Luis Torgo**;ISBN-10: 1482234890; ISBN-13: 978-1482234893

**3. Class Time, Instructor & TA Office Hours:**

**Lectures: Tuesday/Thursday 9:45am-11:05am**

**Professor Wei Zhu (**[**wei.zhu@stonybrook.edu**](mailto:wei.zhu@stonybrook.edu)**);**

**Office Hours: Tuesday/Thursday 3:30pm-4:30pm**

**Professor Zhu Office Hour Zoom link:** [https://stonybrook.zoom.us/my/profweizhu?pwd=RjVIVXg3YUhudzZZQ3pheHUydTJBUT09](https://www.google.com/url?q=https://stonybrook.zoom.us/my/profweizhu?pwd%3DRjVIVXg3YUhudzZZQ3pheHUydTJBUT09&sa=D&source=calendar&ust=1612638200522000&usg=AOvVaw38bNXGCV9Ed7C5zZlqGnCH)

**Mr. Jiecheng Song (Jason) (**[**jiecheng.song@stonybrook.edu**](mailto:jiecheng.song@stonybrook.edu)**);**

**Office Hours: Wednesday 10:00am-12:00noon**

**\* Please see Blackboard for Zoom links to lectures & TA office hours**

**4. Grading:**

**\*Final course score = maximum (cumulative score, final exam at 100%)**

**Cumulative Score:**

1. **Team Project: 20%**
2. **Quizzes: 20% (\* the first 3 HWs will count as 3 Quizzes)**
3. **Midterm (Thursday, March 25, during class): 30%**
4. **Final Exam (Tuesday, May 11, 8:00-10:45am): 30%**

**Note:** This course is offered every semester plus during the summer sessions. The quizzes, midterm and final exam are open book exams. The students will perform various analyses using methods taught in class on data sets provided or simulated during the exam. The required programming language is R that will be taught throughout the course. Some questions in these tests are more theoretical in nature and will require derivations. Each student must work on his/her exam independently – for all the exams and quizzes. \* Include a week-by-week breakdown of the topics \*

|  |  |
| --- | --- |
| Week 1. | Statistics and the R programming language |
| Week 2. | Linear model with R, introduction |
| Week 3. | General linear models with R, classic variable selection methods |
| Week 4. | Regularized regression |
| Week 5. | Generalized linear models (e.g. Logistic regression model) with R |
| Week 6. | Review, midterm exam |
| Week 7. | Cluster analysis |
| Week 8. | Principal component analysis |
| Week 9. | Classification and Regression Tree (CART) |
| Week 10. | Statistical resampling methods, the Jackknife and the Bootstrap |
| Week 11. | Random Forest |
| Week 12. | Neural Networks & its relation to logistic regression |
| Week 13. | Case Studies |
| Week 14. | Review for the final exam |

**5. Required Syllabus Statements**

The University Senate Undergraduate and Graduate Councils have authorized that the following required statements appear in all teaching syllabi (graduate and undergraduate courses) on the Stony Brook Campus.

**Student Accessibility Support Center Statement**

If you have a physical, psychological, medical or learning disability that may impact your course work, please contact Student Accessibility Support Center, ECC (Educational Communications Center) Building, Room 128, (631) 632-6748. They will determine with you what accommodations, if any, are necessary and appropriate. All information and documentation is confidential.

**Academic Integrity Statement**

Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. For more comprehensive information on academic integrity, including categories of academic dishonesty please refer to the academic judiciary website at <http://www.stonybrook.edu/commcms/academic_integrity/index.html>

**Critical Incident Management**

Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of University Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Further information about most academic matters can be found in the Undergraduate Bulletin, the Undergraduate Class Schedule, and the Faculty-Employee Handbook.

**Student Absences Statement**

Students are expected to attend every class, report for examinations and submit major graded coursework as scheduled. If a student is unable to attend lecture(s), report for any exams or complete major graded coursework as scheduled due to extenuating circumstances, the student must contact the instructor as soon as possible.  Students may be requested to provide documentation to support their absence and/or may be referred to the Student Support Team for assistance. Students will be provided reasonable accommodations for missed exams, assignments or projects due to significant illness, tragedy or other personal emergencies. In the instance of missed lectures or labs, the student is responsible for ***making up the lectures (including: review posted slides, review recorded lectures, and make up the missing homework)***.  Please note, all students must follow Stony Brook, local, state and Centers for Disease Control and Prevention (CDC) guidelines to reduce the risk of transmission of COVID. For questions or more information click [here](https://www.stonybrook.edu/commcms/comingback/students.php).