JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Introduction to Data Science
553.436, 553.636
Spring, 2020 (4 credits, EQ)

Instructor

Professor Tamas Budavari, budavari@jhu.edu
https://engineering.jhu.edu/ams/faculty/tamas-budavari/

Office: Whitehead Hall 212C, 410-516-7914
Office hours: Wednesdays 11:10am-12noon, and by appointment

Teaching Assistants and Office Hours

TA	Email	Office Hour
Aranyak Acharyya	aachary6@jhu.edu	Th 4:00 PM - 5:00 PM
Dapeng Yao	dyao10@jhu.edu	F 12:30 PM – 1:30 PM
Jasmine Hu	xhu39@jhu.edu	Tu 10:00 AM – 11:00 AM
Jiahao Shi	jshi35@jhu.edu	M 10:00 AM – 11:00 AM
Jiaying Yi	jyi24@jhu.edu	M 9:00 AM – 10:00AM
Jingyi Gao	jgao38@jhu.edu	W 3:00 PM – 4:00 PM
Qihao Pan	qpan12@jhu.edu	F 2:30 PM – 3:30 PM
Tingwen Guo	tguo12@jhu.edu	M 2:00 PM – 3:00 PM
Wei Jin	wjin@jhu.edu	M 3:15 PM – 4:15 PM

Meetings

Lectures: Monday, Wednesday 12:00PM-1:15PM, Remsen Hall 101

Sections:

Section	Time	Location	Contact
EN.553.436 (01)	F 09:00 AM - 09:50 AM	Maryland 202	Tingwen Guo
EN.553.436 (02)	F 12:00 PM - 12:50 PM	Maryland 217	Jasmine Hu
EN.553.436 (03)	F 10:00 AM - 10:50 AM	Maryland 202	Aranyak Acharyya
EN.553.636 (01)	F 10:00 AM - 10:50 AM	Maryland 217	Dapeng Yao
EN.553.636 (02)	F 11:00 AM - 11:50 AM	Maryland 217	Wei Jin
EN.553.636 (03)	F 01:30 PM - 02:20 PM	Maryland 114	Qihao Pan
EN.553.636 (04)	F 03:00 PM - 03:50 PM	Maryland 104	Jiahao Shi

Office Hour	Monday	Tuesday	Wednesday	Thursday	Friday
Aranyak Acharyya				4:00-5:00 PM	
Dapeng Yao					12:30-1:30 PM
Jasmine Hu		10:00-11:00 AM			
Jiahao Shi	10:00-11:00 AM				
Jiaying Yi	9:00-10:00 AM				
Jingyi Gao			3:00-4:00 PM		
Qihao Pan					2:30-3:30 PM
Tingwen Guo	2:00-3:00 PM				
Wei Jin	3:15-4:15 PM				

Textbook
Recommended books:

- Hastie, T., R. Tibshirani, and J. Friedman (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer
- Zeljko Ivezic, Andrew J. Connolly, Jacob T VanderPlas and Alexander Gray (2014) Statistics, Data Mining and Machine Learning in Astronomy, Princeton University Press
- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flanner. Numerical recipes in C: The art of scientific computing. Cambridge: University Press.

Online Resources
Please log in to Blackboard for all materials related to this course.

Course Information

- Data mining is a relatively new term used in the academic and business world, often associated with the development and quantitative analysis of very large databases. Its definition covers a wide spectrum of analytic and information technology topics, such as machine learning, artificial intelligence, statistical modeling, and efficient database development. This course will review these broad topics, and cover specific analytic and modeling techniques such as advanced data visualization, decision trees, neural networks, nearest neighbor, clustering, logistic regression, and association rules. Although some of the mathematics underlying these techniques will be discussed, our focus will be on the application of the techniques to real data and the interpretation of results. Because use of the computer is extremely important when mining large amounts of data, we will make substantial use of data mining software tools to learn the techniques and analyze datasets.
- **Recommended Course Background**
  EN.553.413 Applied Statistics and Data Analysis
- **Prerequisites**
  Calculus 1 (AS.110.107 or equivalent)
  Linear Algebra
- **Required, Elective or Selective Elective**

Course Goals

Specific Outcomes for this course are that

- Students will learn the foundations of data mining
- Students will learn to apply analysis techniques

This course will address the following Criterion 3 Student Outcomes

- An ability to apply knowledge of mathematics, science and engineering (to solve problems related to materials science and engineering) (Criteria 3(a))
- An ability to design and conduct experiments, as well as to analyze and interpret data (using statistical, computational or mathematical methods) (Criteria 3(b))
- An ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability – the design process (Criteria 3(c))
- An ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability – recognition of constraints within design (Criteria 3(c))
- An ability to function on multidisciplinary teams (Criteria 3(d))
- An ability to identify, formulate and solve engineering problems (Criteria 3(e))
- An understanding of professional and ethical responsibility (Criteria 3(f))
- An ability to communicate effectively (writing) (Criteria 3(g))
- An ability to communicate effectively (oral presentation) (Criteria 3(g))
- The broad education necessary to understand the impact of engineering solutions in a global, economic, environmental and societal context (Criteria 3(h))
- A recognition of the need for and an ability to engage in life-long learning (Criteria 3(i))
- A knowledge of contemporary issues (Criteria 3(j))
- An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice (Criteria 3(k))

Course Topics

- descriptive statistics
- probability, probability density function, moments
- sampling from the distributions, density estimation
- linear methods for regression and classifications
- bayesian inference, numerical methods
- model selection and averaging
- support vector machines
- neural networks
- cluster analysis, k-means
- principal component analysis
- spectral methods

- robust statistics
- databases

Course Expectations & Grading

Grades will be based on 2-4 homework assignments (30%), 2 exams (50%), and a final project (20%) presented in a poster session during the week of finals

Key Dates

Wednesday, Feb 26 – Exam 1 in class (preliminary)
Wednesday, Apr 15 – Exam 2 in class (preliminary)
Tuesday, May 12 – Project presentations 9am – 12noon

Assignments & Readings
Will be posted on the Blackboard site for this course.

Ethics
The strength of the university depends on academic and personal integrity. In this course, you must be honest and truthful. Ethical violations include cheating on exams, plagiarism, reuse of assignments, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition.
Report any violations you witness to the instructor.
You can find more information about university misconduct policies on the web at these sites:

- For undergraduates: http://e-catalog.jhu.edu/undergrad-students/student-life-policies/
- For graduate students: http://e-catalog.jhu.edu/grad-students/graduate-specific-policies/

Students with Disabilities
Any student with a disability who may need accommodations in this class must obtain an accommodation letter from Student Disability Services, 385 Garland, (410) 516-4720, studentdisabilityservices@jhu.edu .