ME 597: Data Analytics for Scientists and Engineers

Instructor

- Ilias Bilionis, *Associate Professor of Mechanical Engineering, Purdue University*

Audience

Advanced undergraduate students, graduate students, and professionals interested in applications of data analytics to engineering problems.

Course Description

This course provides an introduction to data analytics for individuals with no prior knowledge of data science or machine learning. The course starts with an extensive review of probability theory as the language of uncertainty, discusses Monte Carlo sampling for uncertainty propagation, covers the basics of supervised (Bayesian generalized linear regression, logistic regression, Gaussian processes, deep neural networks, convolutional neural networks), unsupervised learning (k-means clustering, principal component analysis, Gaussian mixtures) and state space models (Kalman filters). The course also reviews the state-of-the-art in physics-informed deep learning and ends with a discussion of automated Bayesian inference using probabilistic programming (Markov chain Monte Carlo, sequential Monte Carlo, and variational inference). Throughout the course, the instructor follows a probabilistic perspective that highlights the first principles behind the presented methods with the ultimate goal of teaching the student how to create and fit their own models.

Course Learning Outcomes

After completing this course, you will be able to:

- Represent uncertainty in parameters in engineering or scientific models using probability theory

- Propagate uncertainty through physical models to quantify the induced uncertainty in quantities of interest

- Solve basic supervised learning tasks, such as: regression, classification, and filtering

- Solve basic unsupervised learning tasks, such as: clustering, dimensionality reduction, and density estimation

- Create new models that encode physical information and other causal assumptions

- Calibrate arbitrary models using data

- Apply various Python coding skills

- Load and visualize data sets in Jupyter notebooks

- Visualize uncertainty in Jupyter notebooks

- Recognize basic Python software (e.g., Pandas, numpy, scipy, scikit-learn) and advanced Python software (e.g., pymc3, pytorch, pyrho, Tensorflow) commonly used in data analytics

Prerequisites
- Working knowledge of multivariate calculus and basic linear algebra
- Basic Python knowledge
- Knowledge of probability and numerical methods for engineering would be helpful, but not required
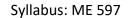
Required Tools

Jupyter Notebook

Jupyter notebooks are interactive documents that can simultaneously contain text, mathematics, images, and executable code. The executable code can be in many programming languages (e.g., R, Matlab), but we are only going to use Python in this course. The course uses Jupyter notebooks for the following content: Reading Activities, Hands-on Activities, and Homework Assignments. The rationale behind this choice is that it allows the student to focus on the mathematical methods rather than the programming and it ensures the reproducibility of the course content. Of course, understanding the code in Jupyter notebooks does require knowledge of Python, albeit it does not require knowing how to structure and call Python code from the command line. Jupyter notebooks can be run either on the students' personal computers (instructions vary with operating system and can be found here) or in several cloud computing resources. The recommended method for this class is to use Google Colab which is available free of charge and requires only a standard Google account. The activity links included in the course will take you automatically to a copy of the latest version of the corresponding Jupyter notebook which you can then save and edit on your Google Drive. If you do not want to create a Google account, then it is also possible to run the notebooks at Purdue's Jupyter Hub using your Purdue Career Account. This method requires more expertise and you may encounter several issues. When this occurs, please post a question on Piazza and the TA will help you out.

Access to the Jupyter Notebooks Repository

As stated earlier, the recommended method for using the Jupyter notebooks of this class is to use Google Colab. The links to all the activities will take you directly to a Google Colab copy of the Jupyter notebook. If you want to use any alternative method (e.g., your personal computers Purdue's Jupyter Hub, or anything else), you will need access to the Jupyter Notebook repository for the class. Select this link for a Git version control repository.

(If you have no idea what Git is, this is a good tutorial.) All the activities are inside the "activities" folder and they are named using the following convention: "activity_type_<lecture_number>.<activity_type_count>.ipynb." For example, "hands-on-06.3.ipynb" is the third hands-on activity of Lecture 6.
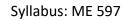
Grading

This course will be graded based on the following criteria:

Assessment Type	Description	% of Final Grade
Homework	There will be ten (10) homework assignments. The homework assignments will be both theoretical (e.g., prove this, derive that) and computational (e.g., use this data to fit that model, create and fit a model for this situation). The assignments will be in the form of a Jupyter notebook with empty space reserved for your writing or coding. If you wish, you can do the writing by hand (instead of the latex required by Jupyter notebooks), scan it and submit a single PDF. Submissions should be made through Gradescope.	100%

Grading Scale

Your course grade will be based on the following grading scale: A+ 95 – 100; A 88 – 95; A- 85 – 88; B+ 80 – 85; B 73 – 80; B- 70 – 73; C+ 67 – 70; C 62 – 67; C- 60 – 62; D+ 57 – 60; D 52 – 57; D- 50 – 52; F 0 – 50. The instructor reserves the right to change with grading scale in a way that favors students.

Course Schedule

Section	Lectures	Assignments
Introduction	1. Introduction to Predictive Modeling (8/25)	NONE
Review of Probability Theory	2. Basics of Probability Theory (8/27) 3. Discrete Random Variables (9/1) 4. Continuous Random Variables (9/3) 5. Collections of Random Variables (9/8) 6. Random Vectors (9/10)	• **Homework 1** ○ Available in the course: 8/25 ○ Due Date: **9/1 at 11:59 PM ET** *(9/2; 3:59 UTC)* • **Homework 2** ○ Available in the course: 9/1 ○ Due Date: **9/15 at 11:59 PM ET** *(9/16; 3:59 UTC)*
Uncertainty Propagation	7. Basic Sampling (9/15) 8. The Monte Carlo Method for Estimating Expectations (9/17) 9. Monte Carlo Estimates of Various Statistics (9/22) 10. Quantify Uncertainty in Monte Carlo Estimates (9/24)	• **Homework 3** ○ Available in the course: 9/15 ○ Due Date: **9/29 at 11:59 PM ET** *(9/30; 3:59 UTC)*
Principles of Bayesian Inference	11. Selecting Prior Information (9/29) 12. Analytical Examples of Bayesian Inference (10/1)	• **Homework 4** ○ Available in the course: 9/29 ○ Due Date: **10/6 at 11:59 PM ET** *(10/7; 3:59 UTC)*
Supervised Learning: Linear Regression and Logistic Regression	13. Linear Regression Via Least Squares (10/6) 14. Bayesian Linear Regression (10/8) 15. Advanced Topics in Bayesian Linear Regression (10/13) 16. Classification (10/15)	• **Homework 5** ○ Available in the course: 10/6 ○ Due Date: **10/22 at 11:59 PM ET** *(10/23; 3:59 UTC)*
Unsupervised Learning	17. Clustering and Density Estimation (10/20) 18. Dimensionality Reduction (10/22)	• **Homework 6** ○ Available in the course: 10/20 ○ Due Date: **10/27 at 11:59 PM ET** *(10/28; 3:59 UTC)*
State-Space Models	19. State-Space Models – Filtering Basics (10/29) 20. State-Space Models – Kalman Filters (11/3)	• **Homework 7** ○ Available in the course: 10/29 ○ Due Date: **11/6 at 11:59 PM ET** *(11/7; 4:59 UTC)*
Gaussian Process Regression	21. Gaussian Process Regression – Priors on Function Spaces (11/5) 22. Gaussian Process Regression – Conditioning on Data (11/10) 23. Bayesian Global Optimization (11/12)	• **Homework 8** ○ Available in the course: 11/5 ○ Due Date: **11/13 at 11:59 PM ET** *(11/14; 4:59 UTC)*
Neural Networks	24. Deep Neural Networks (11/17) 25. Deep Neural Networks Continued (11/19) 26. Physics-Informed Deep Neural Networks (11/24)	• **Homework 9** ○ Available in the course: 11/17 ○ Due Date: **11/30 at 11:59 PM ET** *(12/1; 4:59 UTC)*
Advanced Methods for Characterizing Posteriors	27. Sampling Methods (12/1) 28. Variational Inference (12/3)	• **Homework 10** ○ Available in the course: 12/3 ○ Due Date: **12/8 at 11:59 PM ET** *(12/9; 4:59 UTC)*

Course Help

To get help with course content, comment in the Piazza discussion forums. Each homework assignment and lecture has its own corresponding discussion in the Piazza course site. By commenting in these discussion forums, the course team will be able to respond to your question more quickly.
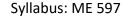
Discussion Guidelines

Please follow the Discussion Guidelines when contributing to discussions in this course. Here are a few of the key points you should remember:

- Do not use offensive language. Present ideas appropriately.
- Be cautious in using Internet language. For example, do not capitalize all letters since this suggests shouting.
- Avoid using vernacular or slang language. This could possibly lead to misinterpretation.
- Do not hesitate to ask for feedback.
- Be concise and to the point.
- Think and edit before you push the "Send" button.

Technical Help

If you experience technical difficulties with the edX platform, contact edX Support using:

- The email address: support-masters@edx.org
- edX's Contact Us form

Accessibility Information

Purdue University strives to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability, you are welcome to let me know so that we can discuss options. You are also encouraged to contact the Disability Resource Center at: drc@purdue.edu or by phone: 765-494-1247.

The **Office of Institutional Equity**, which is responsible for ensuring Americans with Disability Act compliance, can be contacted with any accessibility concerns at:

Phone: (765) 494-7253
Email: equity@purdue.edu
TTY: (765) 496-1343
Website

- Purdue's Disability Resource Center Website
- Purdue's Web Accessibility Policy
- Purdue Equal Access Frequently Asked Questions (FAQs)
- edX's Website Accessibility Policy
- Google Accessibility
- Jupyter Notebook a11y toolbar (from Microsoft research)

Academic Integrity

Academic integrity is one of the highest values that Purdue University holds. Individuals are encouraged to alert university officials to potential breaches of this value by either emailing or by calling 765-494-8778. While information may be submitted anonymously, the more information that is submitted provides the greatest opportunity for the university to investigate the concern.

The Purdue Honor Pledge

> *"As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue"*

Nondiscrimination Statement

Purdue University is committed to maintaining a community which recognizes and values the inherent worth and dignity of every person; fosters tolerance, sensitivity, understanding, and mutual respect among its members; and encourages each individual to strive to reach his or her own potential. In pursuit of its goal of academic excellence, the University seeks to develop and nurture diversity. The University believes that diversity among its many members strengthens the institution, stimulates creativity, promotes the exchange of ideas, and enriches campus life. Link to Purdue's nondiscrimination policy statement.