SML 301: Data Intelligence: Modern Data Science Methods

Spring 2023

Description

This course provides the training and skill set for students who want to be independent in modern data analysis projects. The course emphasizes the rigorous treatment of data, and the programming skills and conceptual understanding required for dealing with modern datasets. The course examines various data applications through the lens of statistics and machine learning methods. Students verify their understanding of the concepts by working with real datasets. The course also covers other components of quantitative data science research, such as experiment design, the ethical use of data, best practices when using statistical and machine learning methods, reproducible research, writing a quantitative research paper, and presenting research results.

Prerequisites

- SML 201 or POL 346 or other equivalent project-based courses.
- SML 201 or ORF 245 or other equivalent statistics courses.
- One semester of calculus or discuss with the course instructor.
- Students are expected to have taken at least one introductory course that explores the fundamental statistical concepts in data science. Familiarity with R or Python programming is assumed.

Course Instructor

- Daisy Huang daisyhuang@princeton.edu

Preceptors

- Shanka Subhra Mondal smondal@princeton.edu
- Chengzhuo Ni chengzhuo.ni@princeton.edu
- Yu Wu yuw@princeton.edu

Office Hours

See the `Office Hours|Contacts` tab on Canvas.

Reading List

For Python (recommended)

Python for Data Analysis by Wes McKinney, 2nd Edition (Available online through Princeton University Library)

Machine learning with Python cookbook : practical solutions from preprocessing to deep learning by Chris Albon (Available online through Princeton University Library)

For data science

The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (Available online)

Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence by Jon Krohn with Grant Beyleveld and Aglaé Bassens (Available online through Princeton University Library)

All of the textbooks above can be accessed online free of cost.

Learning Goals

In this course, students will:

1. Develop critical thinking skills, and learn how to identify potential bias in a dataset and limitations in a study.
2. Acquire the intuition behind several machine learning algorithms and the ways to enhance the performance of these algorithms.
3. Learn common techniques for handling missing data.
4. Best practices in machine learning research
5. Master the ability to clearly communicate one's research in a professional format

Topics

- Model building and evaluations
- Treatments for missing data and imbalanced datasets
- Ethics and bias in data and research studies
- Common pitfalls in ML-based science
- Writing a scientific professional report
- Linear Classifiers
- Nonlinear Classifiers
- Additional topics in week 12

Structure of the course

Please see Course Schedule on Canvas

Grading

Your grade will be determined by the following:

- Pre-lecture exercises and precepts: 15%

- 1 Problem sets: 10%
- Project 1: 20% (individual work)
- Project 2: 20%
- Final Project: 35%

up to 1% extra credit for providing correct answers to other students' questions on Ed discussion; only the top 10 contributors on Ed will be considered for extra credit. An endorsed answer will be counted as 1.5 times of an answer without an endorsement.

Discussion Forum

Please use *Ed Discussion* on Canvas to ask and answer other students' questions about the course. Questions are anonymous to other students but not to the instructors, so please apply proper etiquette when posting. Instead of emailing the instructors with questions, *we prefer that you post on Ed* because other students may have similar issues or have insightful contributions.

Please keep *Ed Discussion* questions general. If your question suggests a possible approach for a question on a project, please post the question *privately.*

Precepts

The completion of weekly precept exercises is mandatory. No late precept will be accepted.

All precepts will be graded based on both effort and correctness.

You have two options to get credits for the precept:

1. You can attend your assigned precept **in-person**; at the end of each precept meeting, your **preceptor will check your work** and give you credit for the precept. You **do not need to submit your work** after that. Please make sure that you write down your name for your preceptor. You will form a group of 2 students (one of the groups can have 3 students if there is an odd number of students in the precept).

2. You can work on the precept on your own or with someone of your choice. your group is responsible to submit your group work (one submission per group) on Gradescope by **Friday 5p.m.** each week to get credit.

In case your preceptor has checked your work, but fails to record the score for you, please **contact your preceptor directly**.

Missing a precept

If for any reason you cannot attend your assigned precept that week, you can contact another preceptor for an alternative precept that you can attend at least 24 hours in advance. If they have room in their precept, they might let you attend the precept that they lead. Remember to either show your work to your preceptor, or upload your solutions to Gradescope before the deadline.

Pre-lecture Exercises (PLE)

You will be assigned some pre-lecture exercises throughout the course.

The PLE should be attempted *before* the respective lecture and are due by the *lecture date at 11:00pm.*

Working with Others

Except for Project 1, problem sets and projects may be completed in groups of up to 2 students. It is okay to work by yourself, if this is preferable.

For problem sets and projects **you may not work with the same student more than once for the entire semester.** Try to use precepts as your opportunities to test your working relationship with your potential group mate. In addition, you can post on Ed Discussion to recruit teammates.

You are responsible for the work your group turns in. You are not allowed to split work between members in a group; we suggest that you each work independently first and then compare your answers with each other once you all finish or help each other if someone in your group gets stuck.

We expect that the work on any given assignment contains approximately equal contributions from each member of the group. Each member is responsible for being able to explain the work that was done for the assignment. **Failing to make contributions and then putting your name on a project will be considered a violation of the honor code.**

Collaboration Policy and Academic Integrity

Please see `Modules > Collaboration policy` tab on Canvas.

You are allowed to read text books and resources online. You may not ask other individuals questions (e.g., you may not ask questions on Stack Exchange or the Python help discussion groups). In accordance with the honor code, you must cite all sources of external information used in your work. This can be a book or a web site. Part of being a successful data scientist is having the ability to leverage existing information and techniques, so it is okay to do so in this course as long as you cite the reference.

Submitting Assignments

You must submit your work electronically on Gradescope/Canvas. **Please do not email your work to any of the instructors in any case** (unless you have no access to Gradescope/Canvas at the time when the assignment is due–in this case please email the Head instructor your work before the deadline and upload the unaltered copies of the files to Gradescope/Canvas as soon as you regain access to Gradescope/Canvas); emailing instructors your work will only cause confusions and delay in grading. Please be sure to read the instructions for the submission procedure at the beginning of assignment too.

Late Submissions

No late precept will be accepted.

Late problem sets and projects will be penalized at intervals rounded up to multiples of 24 hours. For example, if you are 3 hours late, 10% off or if you are 30 hours late, 20% off.

Extensions

I understand that there might be family emergencies, personal sickness, conferences, interviews, or things that come up unexpectedly etc. To reduce stress related to the deadlines, you are given 2 late passes for the problem set and projects (no late pass for the final project). A *late passes* allows you to turn in an assignment up to 24 hours late without a penalty. **You cannot use more than one late pass toward an assignment**. A late pass cannot be used for the final project.

Each time you are late for submitting an assignment, a late pass will automatically be applied until you have used up all the late passes. Therefore, please try to save the late passes for the later assignments.

Note: **It is important that you *do not* email any of the instructors to just let them know that you are using a late pass**.

Extreme Cases

For severe illness or hospitalization, if after using your late pass you still cannot meet the deadline, please make your extension request through your college dean or director of study–in this case, you must also inform your college dean or director of study that you have already received a 24 hour extension. All extension requests must be received before the deadline.

Deadlines

All deadlines are set in EST (US Eastern Standard Time). If an assignment fails to mention the time zone for the deadline, please assume that it is in EST. Please still use EST even if you are traveling elsewhere.

Questions about Your Scores

Regrade requests must be made no later than a week after the project/problem set/midterm grades released; e.g., if the scores of a project were released on a Tuesday, your regrade request for that problem set must be made no later than the following Tuesday; regrade requests made after this time will be not considered.

Please note that a regrade request for a specific problem may result in a regrade of the entire problem set/project. Therefore, a regrade request may result in an increase or decrease of your overall score for the assignment.

If you and the grader cannot agree on the outcome please see the head instructor of the course.

Schedule

Week 1

- Intro to the course

- How do I know if my model can be generalized?
  - Model building: overfitting vs underfitting;
  - Model selection: cross validation.
- Review of Linear Regression, L1 and L2 Regularizations

Assignments:

- Precept: Intro to Python programming

Week 2

Intro to Linear Classifiers

- Linear classification
  - Regularization
  - Hyperparameter tuning, cross-validation

Assignments:

- Precept: Linear classifiers
- Problem set: Linear classifiers, cross-validation

Week 3

Prediction via Optimization

- More on the implementation of linear classification methods
- ROC curve
- Cost function

Intro to Nonlinear Classifiers

- Decision trees

Assignments:

- Precept: ROC curve, Decision trees, Cost functions

Week 4

More on Nonlinear Classifiers

- Random Forest (RF)
- RF applications
  - Implementation
  - Hyperparameter tuning

Assignments:

- Precept: RF applications, Implementation
- Assign Project at the end of the week: Logistic regression, cross-validation, RF, Imbalanced datasets

Week 5

- Imbalanced datasets
- Missing data
    - Common treatments
    - K-nearest neighbors algorithm
- Ethics and bias in data and studies:
    - Causality Inference
    - Observational study v.s. randomized controlled experiment

Assignments:

- Precept: K-nearest neighbors algorithm

Week 6

More on Nonlinear Classifier: Learning from the Data

- Introduction to Artificial Neural Network
- ANN applications
    - Implementation

Assignments:

- Precept: Artificial Neural Network toy examples

Week 7

Learning from the Data: Applications

- ANN applications
    - Hyperparameter tuning
- Dense layers in ANN

Assignments:

- Precept: Image recognition (MNIST dataset)
- Assign Project at the end of the week: Image recognition (CIFAR-10 or CIFAR-100 dataset)

Week 8

Learning from the Data: Applications (con't)

- ANN applications
- Convolutional layers in ANN

Assignments:

- Precept: Convolutional layers in ANN

Week 9

Methodological pitfalls in ML-based science

Reading and discussion: *Leakage and the Reproducibility Crisis in ML-based Science* by Sayash Kapoor, Arvind Narayanan Plus additional readings

Week 10

- Final project proposals (#1 & #2). We will collaborate with three faculty from fields related to Data Science. Each faculty will come to give a guest lecture to describe their project. Students will vote for a project that they want to work on after the third proposal.

Assignments:

- Precept: TBD

Week 11

- Final project proposals (#3)
- Writing a scientific professional report
- The do's and don'ts for making graphs

Assignments:

- Precept: Relate to the hints for the final project
- Project: Final project

Week 12

One of the following options:

Option 1: Unsupervised learning:

- Hierarchical clustering
- K means
- PCA

Option 2:

- Making animation maps for dynamic data

Assignments:

- Precept: TBD

Copyright

Audio or video recording of the lectures is prohibited without permission from the instructor.