The Erdős Institute
May 2023 Data Science Boot Camp
Syllabus

Instructor Information

Name: Matthew Osborne, Ph.D.
Email: matthewosborne71@erdosinstitute.org
Preferred Form of Contact: The Erdős Institute Slack

Boot Camp Aim

The goal of this boot camp is to endow you with the tools to complete end to end data science/machine learning projects with the help of Python.

Brief Overview of Content

In alignment with the aim of our boot camp our materials touch on the following content to varying degrees.

- Data collection:
  - Data competition websites,
  - Data repositories,
  - Databases queries with python and
  - Web Scraping.
- Data analysis and exploration:
  - Exploratory plotting,
  - Examining basic statistics and
  - Data manipulation with `pandas` and `numpy`.
- Data cleaning:
  - Cleaning data files,
  - Cleaning text data with str functionality,
  - Imputing missing values,
  - Creating new columns from existing columns and
  - More.
- Supervised learning:
  - Regression,
  - Classification and
  - Ensemble learning.
- Unsupervised Learning:
  - Dimension Reduction and
  - Clustering
- Neural Networks
  - Perceptrons,
  - Dense Neural Networks,
  - Convolutional Neural Networks and
  - Recurrent Neural Networks.

Boot Camp Information

Setup

After setting up your Erdős Institute profile go to this link, make sure you have completed the steps under *First Steps* on the Data Science Boot Camp website. These steps cover what you need to do in order to:

1. Create a GitHub profile,
2. Clone a GitHub repository and
3. Open a `jupyter notebook` on your computer.

Prerequisites

The content in this boot camp will cover both the math behind algorithms and how to practically implement them in python. The main reason for this is not because I think that you need to know the math perfectly in order to be a good data scientist, but rather because I like to know how things work. For data science and machine learning this involves knowing how to implement an algorithm with some kind of coding and/or knowing the math underlying an algorithm.

The point I want to stress here is that I do not expect you to know all of the math for every algorithm we cover. Content will be written so that you will be able to learn how to implement the algorithms even if you do not fully understand the math. That being said, I will present the math for those of you that do want to understand it.

Coding

In these notebooks we assume that you know the basics of Python. In particular you should be familiar with:

- `pandas`,
- `numpy` and
- `matplotlib`.

If you are not familiar with python work through our python prep content found here, https://www.erdosinstitute.org/programs/asynchronous/python-prep/

Math and Statistics

For those of you looking to understand the math presented in these notebooks an absolute minimum includes:

- Calculus (primarily derivatives),
- Some linear algebra,
- Basic probability theory and
- Basic statistics.

If you are unfamiliar with these concepts and would like to know more, check out our review slides:

- Probability Theory: https://rb.gy/9uyw5,
- Basic Statistics: https://rb.gy/tdebs,
- Calculus Refresher: https://rb.gy/a6b6i, and
- Linear Algebra: https://rb.gy/bl7qh.

Our GitHub Repository

All `jupyter notebooks` for this boot camp will be found at our GitHub repository. You can find the link to this repository under the "Program Content" section of the boot camp's website. In order to gain access to this repository you need to:

1. Add your GitHub profile information to your Erdős Institute profile and
2. Then be granted repository access by our community manager.

Boot Camp Format

Lectures

There will be live virtual lectures that work through the `jupyter notebook`s in the lecture folder of the repository. Not every lecture notebook will be touched upon, but the most important concepts will be covered in the live lectures. With that in mind, every lecture notebook has a pre-recorded lecture video available on the Data Science Boot Camp website. Even if you plan on attending the live lecture, I encourage you to watch these pre-recorded videos as needed. Watching the pre-recorded videos prior to the live lecture can help you prepare questions to ask during the live lecture.

At the start of the boot camp there will be multiple versions of each notebook:

- An empty version for you to fill with your own notes and coding attempts,
- A complete version that I completed while recording the lecture video and
- A live lecture version that I completed while giving a live lecture during our May 2022 boot camp, note that not every notebook has a live lecture version.

Importantly, you are able to consume the lecture content without having to come to the live lecture. You are welcome to consume the lecture videos asynchronously. If you take the asynchronous route you are always welcome to message me with questions.

Prep Notebooks & Problem Sessions

There will also be problem solving sessions in which you will form small groups to solve problems that touch on the concepts covered in that week's lectures. See the schedule for exact dates/times for the problem sessions.

Each problem session `jupyter notebook` will be found in the problem sessions folder of the repository. Each problem session notebook will also have an accompanying prep notebook. These notebooks are *completely optional*, but may provide a good refresher on some background material required to complete the problem session. Again these are optional, but past participants have found them helpful.

Group Projects

In order to receive the Erdős Institute Data Science Boot Camp certificate you must complete a group project by the end of the program. Each group will be assigned a project mentor to help them throughout the semester. Projects culminate in a five minute pre-recorded project presentation video.

Additional project coordination will be provided by Alec Clott, Ph.D., a former Erdős Institute alumni now working in industry. The best way to contact him is on the Erdős Institute slack.

We will provide you with additional details regarding projects during the boot camp.

Practice Problems

In addition to the lectures and problem sessions the repository has a folder of practice problems that are additional problems for you to use as practice, on your own time. These problems are **not** homework, but may be useful review as you prepare for job interviews or if you just want to explore more data science content.

Final Note

We look forward to having you participate in the our Data Science Boot Camp! If you have any questions or concerns, do not hesitate to contact us on slack or via email. We do our best to answer promptly.

References

The following is a list of references that are helpful for learning data science and machine learning. These are **not** required reading, but you may be interested in giving them a look. Many of these have at least one edition available for free online.

- Applied Predictive Modeling
- Python for Data Analysis
- Introduction to Machine Learning with Python
- Hands-On Machine Learning with Scikit-Learn and TensorFlow
- An Introduction to Statistical Learning
- Regression and Other Stories
- Elements of Statistical Learning
- The Hundred-Page Machine Learning Book
- Neural Networks and Deep Learning
- Deep Learning with Python
- Deep Learning with PyTorch