The Erdős Institute
May-Summer 2024 Data Science Boot Camp
Syllabus

Instructor Information

Name: Steven Gubkin, Ph.D.
Email: steven@erdosinstitute.org
Preferred Form of Contact: The Erdős Institute Slack

Boot Camp Aim

The goal of our Data Science Boot Camp is to provide you with the skills and mentorship necessary to produce a portfolio worthy data science/machine learning project while also providing you with valuable career development support and connecting you with potential employers.

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
  - 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
- Supervised learning:
  - Regression,
  - Classification
  - Ensemble learning.
- Unsupervised Learning:
  - Dimension Reduction
  - Clustering
- Neural Networks
  - Perceptrons,
  - Dense Neural Networks,
  - Convolutional Neural Networks
  - Recurrent Neural Networks.

Boot Camp Information

Setup

After setting up your Erdős Institute profile make sure you have completed the *First Steps* on the Data Science Boot Camp website. These steps cover what you need to do in order to create a GitHub profile, clone a GitHub repository and open a `jupyter notebook` on your computer.

Prerequisites

Coding

We assume that you already have a basic working knowledge of python. You should understand the basics of control flow, data types, classes and methods. You should also have some familiarity with the pandas, numpy, and matplotlib libraries.

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

You can get a lot out of this bootcamp by just learning to apply the set of tools we cover. You do not *need* to have a math/stats background to succeed in the bootcamp.

However, your understanding of the content will be much deeper and more robust if you invest in learning the mathematical and statistical underpinnings of the techniques. For this, the bare minimum would be familiarity with the following topics at the level indicated by the linked review slides.

- Multivariate differential calculus (https://rb.gy/a6b6i)
- Some linear algebra (https://rb.gy/bl7qh)
- Basic probability theory (https://rb.gy/9uyw5)
- Basic statistics (https://rb.gy/tdebs)

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 notebooks` 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 2023 boot camp, note that not every notebook has a live lecture version.

Importantly, while it is preferable for you to attend live, you are also able to attend 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