CFT Day 1: Introduction to Data Analysis

Keywords: Python, Colab, Big Data, optional

Learning Outcomes

- Introduction to Python in the Colab environment
  - Learn to use Jupyter notebooks
  - Learn to use the Colab environment

- Introduction to Data Analysis Workflows
  - Learn how to construct a workflow to ingest, analyze, and visualize large data

Exercises

| Working with data in Python Colab environment |
| Working with big datasets and visualization |

https://go.osu.edu/CcEr
CFT Day 2: Python for Data Analysis

Keywords: Pandas, Numpy, Timing, Profiling, Colab, optional

Learning Outcomes

Introduction to Numpy
• Gain familiarity with Numpy for data analysis

Introduction to Pandas
• Practice using the Pandas package tools for datasets

Introduction to Timing & Profiling
• Work with introductory examples of timing code execution

Exercises

<table>
<thead>
<tr>
<th>Numpy basics</th>
<th>![Book Icon]</th>
<th>![CPU Icon]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pandas basics</td>
<td>![Book Icon]</td>
<td>![CPU Icon]</td>
</tr>
<tr>
<td>Code Timing &amp; Profiling</td>
<td>![Book Icon]</td>
<td>![CPU Icon]</td>
</tr>
</tbody>
</table>

https://go.osu.edu/CcEv
https://go.osu.edu/CcEw
https://go.osu.edu/CcEx
CFT Day 3: Use Cases in Data Science

Keywords: Data Science workflows, unstructured text, image data, GeoPandas

---

### Learning Outcomes

Highlight workflows for different kinds of data
- Understand how workflows are modified with different data types: text, images, maps
- Gain experience implementing the workflow examples

### Exercises

- Workflows with Image data
- Workflows with Unstructured text
- Working with Maps

https://go.osu.edu/CcEz
**CFT Day 4: Fundamentals of Machine Learning**

**Learning Outcomes**

- **Machine Learning Basics**
  - Gain familiarity with the types of machine learning: supervised, unsupervised, reinforcement

- **Optimization strategies**
  - Practice with Linear regression with gradient descent
  - Work with unsupervised clustering methods

**Exercises**

<table>
<thead>
<tr>
<th>Linear Regression</th>
<th><img src="/fs/ess/PZS1124/AI_BOOT_CAMP_OSC/materials" alt="Notebook" /></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering examples: Mixture of Gaussians, K-means</td>
<td><img src="/fs/ess/PZS1124/AI_BOOT_CAMP_OSC/materials" alt="Clustering" /></td>
</tr>
</tbody>
</table>
CFT Day 5: Probabilistic Modeling / Bayesian Networks

Learning Outcomes

Probabilistic Modeling
• Become familiar with basic concepts in probabilistic machine learning
• Either try out coding up, or explore code for, the Naïve Bayes algorithm to understand probability modeling

Bayesian Networks
• Explore the results of probabilistic cluster induction through the Latent Dirichlet Allocation algorithm

Exercises

Naïve Bayes classifier for sentiment analysis

Latent Dirichlet Allocation

/ fs/ ess/ PZS1124/ AI_ BOOT CAMP_ OSC/ materials
CFT Day 6: Neural Networks

Keywords: Neural Networks

Learning Outcomes

Basics of Neural Networks
- Understanding the "model zoo" of current neural network approaches
- Build a first neural network in Pytorch
- Gain familiarity with Multilayer Perceptrons (MLPs)
- Convolutional Neural Networks (CNNs)
- Become familiar with convolutional networks for image processing
- Compare Multilayer Perceptrons (MLPs) to Convolutional Neural Networks (CNNs)

Exercises

<table>
<thead>
<tr>
<th>Building a 4-2-4 autoencoder in Pytorch</th>
</tr>
</thead>
<tbody>
<tr>
<td>/fs/ess/PZS1124/Al_BOOT CAMP_OSC/materials</td>
</tr>
</tbody>
</table>

Explore multi-layer perceptrons vs convolutional networks
CFT Day 7: Exploiting HPC for ML, DL and Data Science

Keywords: MPI, NCCL

Learning Outcomes

Overview of NCCL and MPI Libraries
- Become familiar with Hardware Architectures like Interconnects and Processors
- Gain familiarity with Communication Middleware including: Message Passing Interface (MPI), CUDA-Aware MPI and NVIDIA NCCL

Benchmarking with MPI and NCCL
- Gain familiarity with OSU Micro Benchmarks for MPI, PGAS, CUDA and OPENACC
- Try out or explore the results from running the benchmarks

Exercises

Running OMB-based MPI Benchmarks

Running OMB-based NCCL Benchmarks

/fs/ess/PZS1124/Al_BOOT_CAMP_DK/intro_to_hpc/
CFT Day 8: Introduction to Data Parallel DNN Training

Keywords: Horovod, tensorflow, pytorch, data-parallel training

Learning Outcomes

Deep Learning Frameworks
- Review Deep Learning Frameworks
- Gain familiarity with DL Execution Environments
- Learn about the steps in Deep Learning training
- Learn about approaches to parallel DNN training
- Understand data-parallel training and trade-offs
- Learn about the evolution of Data-Parallel Training
- Compare training time for serial and parallel DNN training

Exercises

- Setting up Distributed DNN Training using Horovod
- Evaluate the performance of Data-Parallel training for MLP and ResNet-50

/fs/ess/PZS1124/Al_BOOT/CAMP_DK/intro_ddp/session8
CFT Day 9: ML and DL Frameworks for Analysis of Large Data Sets

Learning Outcomes

- Practical ML and DL workflows for text and images
  - Create workflows to analyze text and images
  - Realize workflows for texts and pictures on commonly used platforms
  - Gain a practical understanding of widely used machine learning and deep learning methods on real data
  - Learn to measure the effectiveness of machine learning methods
  - Learn methods of working with hyperparameters

Exercises

- Unsupervised Classification w/ Image Data
- Workflows for images and with multi-layer perceptrons and convolutional neural networks

Keywords: Image classification, Naïve Bayes, SVM, Logistic regression

https://go.osu.edu/CcE2
https://go.osu.edu/CcE3
https://go.osu.edu/CcE4
SDFT Day 10: Data Science using Dask

Keywords: spark, hadoop, dask

**Learning Outcomes**

Software for Data Science on HPC Systems
- Gain familiarity with data growth
- Become familiar with the Apache Spark Project
- Become familiar with DASK and task graphs
- Learn about Parallel and Distributed Data Science using MPI4Dask

**Exercises**

- Run parallel and distributed data science applications using Dask on HPC systems
- How to use multi-node GPUs for Dask-based applications

/fs/ess/PZS1124/AI_BOOT/CAMP_DK/dask_session/
Key to Icons for exercises

- Command Line
- Notebook
- CPU
- GPU