Machine Learning

Let’s learn something!

PIERIAN DATA

Python and Spark

- It is now time to begin with the Machine Learning Sections of the course!
- This introduction section will discuss a general introduction to machine learning and how Spark's MLlib library works for Machine Learning.

Python and Spark

- Most Machine Learning Sections have:
    - Suggested Reading Assignment
    - Basic Theory Lecture
    - Documentation Walkthrough
    - More realistic custom code example
    - Consulting Project
    - Consulting Project Solutions

Python and Spark

- The Consulting Projects are looser, more realistic projects for you to attempt with the skills you just learned.
- A dataset, some background, and a problem is described, and you are free to solve it however you want.

Python and Spark

- If you prefer a more guided approach to problems, that’s totally okay!
- We have the custom code examples before each Consulting Project.
- Plus, you can treat the Consulting Project Solutions as an additional “code-along”!

Python and Spark

- Because different students have different backgrounds in math, we will keep the mathematics behind the machine learning algorithms light.

Python and Spark

- If you are interested in reading more about the math behind the algorithms we discuss, we will be using **Introduction to Statistical Learning** by Gareth James as a companion book.
- It's freely available online.

Companion Book

- Students who want the mathematical theory should do the suggested reading assignment that will appear for each machine learning section.
- Otherwise, feel free to watch the Intro Theory Lectures for the fundamentals.

Companion Book

- First Suggested Reading Assignment:
  - Read Chapters 1 & 2 to gain a background understanding before continuing to the Machine Learning Lectures.

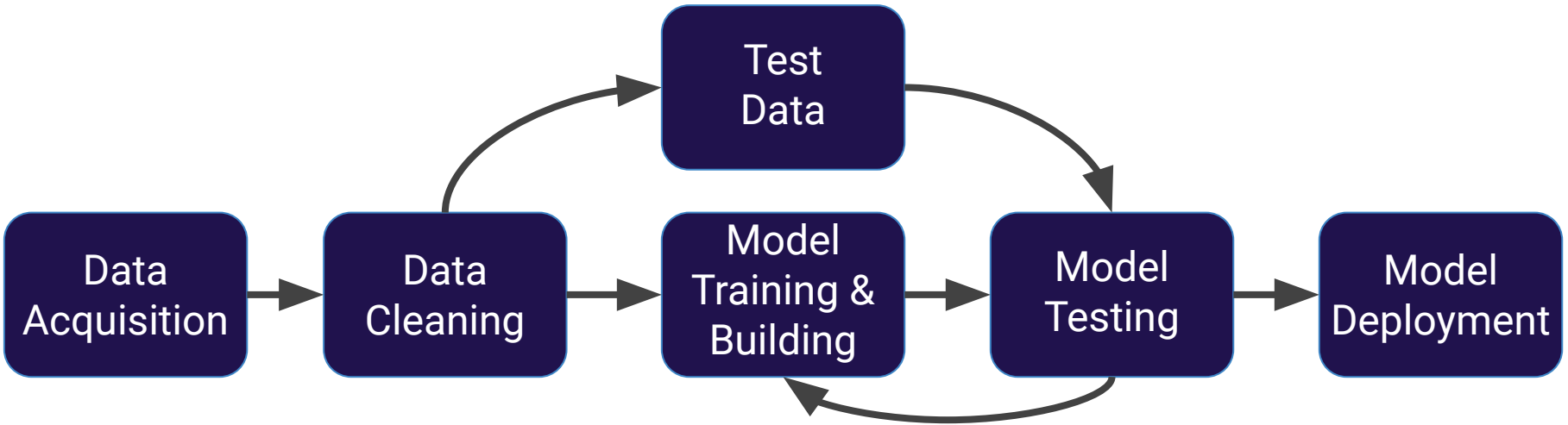
What is Machine Learning?

- Machine learning is a method of data analysis that automates analytical model building.
- Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

What is it used for?

- Fraud detection.
- Web search results.
- Real-time ads on web pages
- Credit scoring and next-best offers.
- Prediction of equipment failures.
- New pricing models.
- Network intrusion detection.
- Recommendation Engines
- Customer Segmentation
- Text Sentiment Analysis
- Predicting Customer Churn
- Pattern and image recognition.
- Email spam filtering.
- Financial Modeling

Machine Learning Process



Supervised Learning

- Spark's MLlib is mainly designed for Supervised and Unsupervised Learning tasks, with most of its algorithms falling under those two categories.
- Let's discuss them in more detail and describe how they are different!

Supervised Learning

- **Supervised learning** algorithms are trained using **labeled** examples, such as an input where the desired output is known.
- For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs).

Supervised Learning

- The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors.
- It then modifies the model accordingly.

Supervised Learning

- Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data.
- Supervised learning is commonly used in applications where historical data predicts likely future events.

Supervised Learning

- For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.
- Or it can attempt to predict the price of a house based on different features for houses for which we have historical price data.

Unsupervised Learning

- **Unsupervised learning** is used against data that has no historical labels.
- The system is not told the "right answer." The algorithm must figure out what is being shown.
- The goal is to explore the data and find some structure within.

Unsupervised Learning

- For example, it can find the main attributes that separate customer segments from each other.
- Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition.
- One issue is that it can be difficult to evaluate results of an unsupervised model!

Final Thoughts

- Machine Learning takes time to learn.
- Be patient with yourself and feel free to post to the QA forums.
- No one course can be a reference for all Machine Learning topics, but I'm always happy to point you in the right direction!

Machine Learning with Spark

PIERIAN DATA

Python and Spark

- Spark has its own MLlib for Machine Learning.
- The future of MLlib utilizes the Spark 2.0 DataFrame syntax.

Python and Spark

- One of the main “quirks” of using MLlib is that you need to format your data so that eventually it just has one or two columns:
    - Features, Labels (Supervised)
    - Features (Unsupervised)

Python and Spark

- This requires a little more data processing work than some other machine learning libraries, but the big upside is that this exact same syntax works with distributed data, which is no small feat for what is going on “under the hood”!

Python and Spark

- When working with Python and Spark with MLlib, the documentation examples are always with nicely formatted data.
- However, we'll have our own custom examples that have messier, more realistic data!

Python and Spark

- We will also have consulting projects, which set you loose on a real world data project with a data set and a problem to solve, without explicitly telling you what to do!

Python and Spark

- A huge part of learning MLlib is getting comfortable with the documentation!
- Being able to master the skill of finding information (**not** memorization) is the key to becoming a great Spark and Python developer!

Python and Spark

- Fortunately, the Spark MLlib documentation is quite good, and we'll constantly teach you how to refer to it during each Machine Learning Algorithm Section.
- Let's jump to it now!

spark.apache.org