Overview

A. What is Machine Learning?

Machine learning is the branch of science that deals with algorithms and systems performing specific tasks using patterns and inference, rather than explicitly programmed instructions. There are a variety of different use cases for machine learning, from image recognition to text generation. Most machine learning tasks generalize to one of the following two learning types:

- **Supervised learning**: Using labeled data to train a model. The labels for the training dataset represent the class/category that each data observation belongs to. After training, the model should be able to predict labels for new data observations (from the same population distribution as the training data).
    - Example: Let's say you're training a machine learning model to predict whether a picture contains a lake or not. With supervised learning, you would train a model on a dataset of pictures where the label for each picture is "Yes" if it contains a lake or "No" if it doesn't. After training, the model will be able to take in a picture and determine whether or not it contains a lake.
- **Unsupervised Learning**: Using *unlabeled* data to allow a model to learn relationships between data observations and pick up on underlying patterns. Most data in the world is unlabeled, which makes unsupervised learning a very useful method of machine learning.
    - Example: Going back to the same picture dataset from above, but now assume the training dataset is unlabeled. Using unsupervised learning, a model will be able to pick up on the inherent differences between pictures with a lake and pictures without a lake, e.g. differences in pixel color or orientation. This allows the model to cluster the pictures into two separate groups.

If it is possible to get large enough labeled training datasets, supervised learning is the way to go. However, it is often difficult to get fully labeled datasets, which is why many tasks require unsupervised learning or semi-supervised learning (a mix of supervised and unsupervised learning). Deciding which type of learning method to use is only the first step towards creating a machine learning model. You also need to choose the proper model architecture for your task and, most importantly, be able to process data into a training pipeline and interpret/analyze model results.

B. ML vs. AI vs. Data Science

People often throw around the terms “machine learning”, “artificial intelligence”, and “data science” interchangeably. In reality, machine learning is a subset of artificial intelligence and overlaps heavily with data science. Artificial intelligence deals with any technique that allows machines to display “intelligence”, similar to humans. Machine learning is one of the main techniques used to create artificial intelligence, but other non-ML techniques (e.g. alpha-beta pruning, rule-based systems) are also widely used in AI.

On the other hand, data science deals with gathering insights from datasets. Traditionally, data scientists have used statistical methods for gathering these insights. However, as machine learning continues to grow, it has also penetrated into the field of data science.

In industry, any data scientist or AI researcher needs to have a good understanding of machine learning. Machine learning in industry has allowed us to create wonderful autonomous systems. These systems have matched, or sometimes even exceeded, the best human performance in their respective fields. A good example is AlphaGo, a machine-learning based system that has beaten the best human Go players in the world.

C. 7 Steps of the Machine Learning Process

1. Data Collection: The process of extracting raw datasets for the machine learning task. This data can come from a variety of places, ranging from open-source online resources to paid crowdsourcing. The first step of the machine learning process is arguably the most important. If the data you collect is poor quality or irrelevant, then the model you train will be poor

quality as well.

2. Data Processing and Preparation: Once you've gathered the relevant data, you need to process it and make sure that it is in a usable format for training a machine learning model. This includes handling missing data, dealing with outliers, etc.
3. Feature Engineering: Once you've collected and processed your dataset, you will likely need to transform some of the features (and sometimes even drop some features) in order to optimize how well a model can be trained on the data.
4. Model Selection: Based on the dataset, you will choose which model architecture to use. This is one of the main tasks of industry engineers. Rather than attempting to come up with a completely novel model architecture, most tasks can be thoroughly performed with an existing architecture (or combination of model architectures).
5. Model Training and Data Pipeline: After selecting the model architecture, you will create a data pipeline for training the model. This means creating a continuous stream of batched data observations to efficiently train the model. Since training can take a long time, you want your data pipeline to be as efficient as possible.
6. Model Validation: After training the model for a sufficient amount of time, you will need to validate the model's performance on a held-out portion of the overall dataset. This data needs to come from the same underlying distribution as the training dataset, but needs to be different data that the model has not seen before.
7. Model Persistence: Finally, after training and validating the model's performance, you need to be able to properly save the model weights and possibly push the model to production. This means setting up a process with which new users can easily use your pre-trained model to make predictions.

D. What this course will provide

After taking this course, you'll be able to take process and clean a raw dataset, train a machine learning model on the data, and validate the model's performance. Specifically, you will be able to:

- Take a raw dataset and process it for a given task. This means dealing with missing data and outliers, normalizing and transforming features

with missing data and outliers, normalizing and transforming features,

figuring out which features are the most relevant to the task, and picking out the best combination of features to use.

- Picking the correct model architecture to use based on the data. Many people will always default to using a large neural network for any machine learning task, but many times this is unnecessary and can even hurt the model's final performance if the dataset is not large enough.
- Code a machine learning model and train it on processed data. Validate the model's performance on held-out data and understand techniques to improve a model's performance.