FINTECH 540 - Machine Learning for Fintech

Fall Semester 2023

Second Lecture

Introduction to Machine Learning Paradigms

Duke

PRATT SCHOOL of ENGINEERING

Today's Learning Objectives

Today we will explore the basic concepts of machine learning modeling.

- Define what *learning* means in the context of machine learning.
- Understand how machine learning differs from standard statistical approaches.
- Explore different paradigms of machine learning, i.e., *different ways of learning*.

We will cover these **essential notions** to prepare for diving into different modeling techniques.

Learning: A Fundamental Concept

The concept of learning is central to human development. Specifically, learning *refers to the known ability of humans and animals to acquire the capability to carry out a task and to generalize when the context changes.*

So, what is **machine learning** all about?

Machine learning involves:

- Using a set of algorithms that learn a model from data.
- Limiting the number of required assumptions.

What is a model? What is an assumption?

Overlaps with Econometrics

Both machine learning (ML) and econometrics, which is the general approach in the financial industry, aim to build predictive models using explanatory variables. However, they often use different methodologies to accomplish this.

So, what is the critical difference between ML and Econometrics?

In the context of ML, **learning** is not the end goal. Instead, it is a process or method used to achieve the ultimate task, such as generalization in prediction or classification.

Difference with Econometrics

Here are some key differences between the two approaches:

- **Focus**: Econometrics emphasizes understanding relationships and causal inference, while ML prioritizes prediction accuracy.
- **Model Structure**: Econometrics often assumes a specific functional form, while ML can capture complex, non-linear relationships.
- **Variable Importance**: Econometrics values the interpretability and significance of variables, while ML focuses on overall predictive power.
- **Validation**: ML uses techniques like cross-validation for model assessment, while Econometrics often relies on hypothesis tests and confidence intervals.
- **Data Size**: ML techniques are often better suited to large high-dimensional, unstructured datasets.

Why is ML Useful?

Why can't we stick with the standard inferential statistics?

Econometrics:

1. Choose a statistical model (e.g., linear regression).
2. Estimate the model's parameters using the available data.
3. Draw conclusions from the estimate.

ML:

1. Make fewer or no assumptions based on prior knowledge of the problem.
2. Estimate a statistical model whose form is driven by the data.

The central aspect to consider is the difference between *choosing a model based on prior assumptions* (as in Econometrics) and *learning a model from the data* (as in ML).

Focusing on these different approaches

Let us try to frame what has been described so far.

Focusing on these different approaches

Let us try to frame what has been described so far.

How many in the class are familiar with or have heard of **Econometrics**?

Focusing on these different approaches

Let us try to frame what has been described so far.

How many in the class are familiar with or have heard of **Econometrics**?

Can someone make an example of a common *econometric problem* one can face in economics/finance?

Focusing on these different approaches

Let us try to frame what has been described so far.

How many in the class are familiar with or have heard of **Econometrics**?

Can someone make an example of a common *econometric problem* one can face in economics/finance?

Let us discuss how the problem could be tackled differently using **ML**.

Data-driven models $\longrightarrow$ the kind of data that shapes the problem that an algorithm should solve

Let's consider a general two-dimensional case: we have a dataset $\mathcal{D}$ composed by a number M of vectors $\mathbf{x} \in \mathbb{R}^K$ where x_i represents the i-th feature. These vectors are generally referred to as **examples** or **data points**.

$\mathcal{D}$ is a two-dimensional matrix with M rows and K columns. ML could also work with less structured data, as we will see...

Depending on $\mathcal{D}$ we define different **learning paradigms**

- **Supervised Learning**: Known problem answers exist. The goal is to learn the relationship between the response and data, i.e., **learn a mapping**.

- **Unsupervised Learning**: No a priori answer exists. The aim is to learn valuable dataset characteristics, e.g., **learn the data generating process (DPG)**.

- **Reinforcement Learning**: Used for sequential decision-making problems. The goal is to **learn the optimal way to act** in a given context.

Supervised learning is the most common machine learning paradigm and is involved in many commercial applications.

The dataset $\mathcal{D}$ is made of data points $\mathbf{x}$, which includes the values x_i up to $K-1$ and the labels or targets y, which are the values x_K.

The goal is to learn a function $f(\mathbf{x}) = y$ where $\mathbf{x} \in \mathbb{R}^{K-1}$. The type of label defines the supervised problem

- if $y \in \mathbb{R}$, we solve a **regression** problem, e.g., predicting house prices.

- if $y \in \{1 \dots k\}$, we solve a **classification** problem, e.g., identifying wether an email is spam or not.

Supervised comes from the supervision applied by the researchers who used to manually label the dataset to specify its context for the models, e.g., stating which objects are included in an image.

The label choice is the most crucial part in the setting of a supervised problem because it affects the type of mapping that the algorithm learns from the data, i.e., the objective of the learning itself.

Examples of Supervised Learning

- **Bioinformatics**, given some information about our face (the vectors $\mathbf{x}$), smartphones are able to unlock by recognizing us (the label y).
- **Spam detection**, given certain structure on the received emails (the vectors $\mathbf{x}$), a software detects which are spam (the label y).
- **Credit scoring**, given some information about a possible borrower (the vectors $\mathbf{x}$), a bank decides if to grant a loan to him (the label y).
- **Volatility forecast**, given price data of S&P500 index and other relevant macroeconomic indicators (the vectors $\mathbf{x}$), predict the future level of market volatility (the label y).

Examples of Supervised Learning(II)

- **Medical Diagnosis**, given a patient's symptoms and medical history (the vectors $\mathbf{x}$), a machine learning model can predict the likelihood of a particular disease (the label y).

- **Sentiment Analysis**, given the text of a product review (the vectors $\mathbf{x}$), a model can predict whether the sentiment expressed is positive, negative, or neutral (the label y).

- **Fraud Detection**, given transaction details (the vectors $\mathbf{x}$), a model can predict whether the transaction is fraudulent or not (the label y).

- **Stock Price Prediction**, given historical stock prices and other financial indicators (the vectors $\mathbf{x}$), a model can predict future stock prices (the label y).

Unsupervised learning is a type of machine learning that aims to discover patterns in the dataset $\mathcal{D}$ without the use of labels. It infers the probability distribution $p(\mathbf{x})$ that generates the data points.

In this approach, the algorithm does not receive any label y. Instead, it learns directly from the features of the data, without trying to predict a specific output.

Unsupervised learning searches for patterns in the data to provide a different representation and learn some data characteristics. For example, it can be used to group customers into different segments based on their purchasing behavior.

Examples of Unsupervised Learning

- **Recommendation systems**, given information about a set of customers (the vectors $\mathbf{x}$), one can learn how to group people in clusters by inferring similarities from the data
- **Social Network Analysis**, given a network of users (the vectors $\mathbf{x}$), a model can identify communities or clusters of users with similar interests or behaviors.
- **Money Laundering detection**, given a set of bank transactions (the vectors $\mathbf{x}$), the algorithm identifies unusual activities, i.e., anomalies in the data.
- **Natural Language Processing**, given a large corpus of text (the vectors $\mathbf{x}$), a model can identify topics or themes in the text.

Remark: the difference between SL and UL is sometimes subtle

Reinforcement Learning (RL) is a machine learning paradigm where an agent learns to make decisions by interacting with an environment. Unlike Supervised or Unsupervised Learning, RL focuses on trial-and-error learning to perform optimal control tasks.

In RL, the agent's actions and the environment's responses dynamically shape the dataset. The data vectors $\mathbf{x}$ encapsulate the environment's state, the agent's action, and the received reward.

For instance, RL can train a self-driving car to navigate roads efficiently and safely by learning from its interactions with traffic and road conditions.

An agent interacts with an environment when solving a decision-making problem with no fixed dataset.

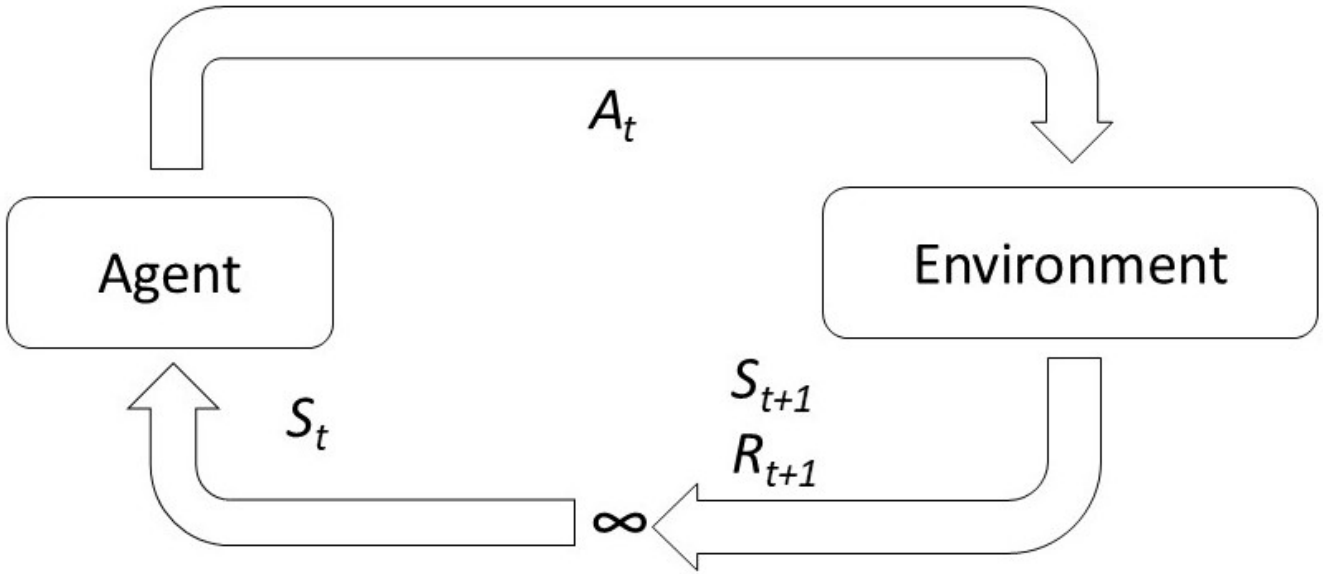

Figure 1: RL loop

Examples of Reinforcement Learning

- **Self-driving cars**, where a car (the agent) interacts with a set of external factors (environment) to make the right driving decision.
- **Energy Management**, where a system (the agent) learns to manage energy usage in a building or city by observing and responding to patterns in energy consumption and environmental conditions.
- **Gaming**, AlphaGo speaks for itself. It learned how to play such a complicated board game at a superhuman level.
- **Financial Trading**, where a trader (the agent) observes the financial market (the environment) and plans the optimal way to construct profitable strategies.
- **Online Advertising**, where an advertising system (the agent) learns to display the most effective ads by observing user behavior (the environment).

Test your understanding

We have talked about these algorithmic approaches in the ML space so far. Try to think about these questions:

- Which of the paradigms is more similar to a standard econometric approach?

Test your understanding

We have talked about these algorithmic approaches in the ML space so far. Try to think about these questions:

- Which of the paradigms is more similar to a standard econometric approach?

- Can you perform the same task using different ML paradigms? Can you blend them?

Test your understanding

We have talked about these algorithmic approaches in the ML space so far. Try to think about these questions:

- Which of the paradigms is more similar to a standard econometric approach?
- Can you perform the same task using different ML paradigms? Can you blend them?
- Which paradigm is more similar to the human approach to learning new concepts?

Test your understanding (II)

- Can you blend two paradigms to solve a specific problem?

Test your understanding (II)

- Can you blend two paradigms to solve a specific problem?
- Can you tell why ML has become so ubiquitous in finance?

Test your understanding (II)

- Can you blend two paradigms to solve a specific problem?
- Can you tell why ML has become so ubiquitous in finance?
- Is recognizing images the same as carrying our profitable trades in a financial market?

Questions? Comments?