MSCI 446: Introduction to Machine Learning

Introduction

Mehrdad Pirnia

WHAT IS MACHINE LEARNING (ML)?

- ML is a branch of artificial intelligence (AI), assisting humans by utilizing data and algorithms while improving their accuracy gradually. The term “machine learning” was originated by Arthur Samuel in reference to his research in the early 1960s, in which Robert Nealy played checkers against an IBM 7094 computer and lost.
- The global ML market size is predicted to grow from $21.17 billion (USD) in 2022 to $209.91 billion (USD) by 2029 (Fortune Business Insights, 2022). Deloitte AI Institute’s (2020) survey found that 67% of companies are using ML, and 97% are using or planning to use it in the future.
- This growing field is the engine behind modern technologies in healthcare, energy, finance, transportation, etc. Some concrete examples could be social media features, like Facebook Notes on your activities, virtual assistants (e.g., Siri, Amazon’s Alexa, and Google), product recommendations in ecommerce websites, and image recognition when tagging people.

Descriptive vs Predictive vs Prescriptive

The function of a machine learning system can be descriptive, meaning that the system uses the data to explain what happened; predictive, meaning the system uses the data to predict what will happen; or prescriptive, meaning the system will use the data to make suggestions about what action to take.

(Malone et al., 2020)

EXAMPLES OF MACHINE LEARNING Applications -1

COVID Moonshot is an open-source project to crowdsource potential medicines against COVID-19. The project has identified new anti-viral drugs in 2020, which were ready to advance to animal trials. It uses the submission of designs for molecules with potential to balk the virus and utilizes ML methods (semi-supervised deep learning) to analyze more than 14,000 submissions.

- For more information, read Crowdsourcing against coronavirus.

EXAMPLES OF MACHINE LEARNING Applications -2

A language model was trained using neural networks, by MIT researchers, to read the genes of viruses and predict potentially dangerous mutations. They discovered that an infection-causing virus possesses both an appropriate biological "grammar" and a semantic "meaning" of immune responses. They trained an ML model to guess a missing word in a sentence, using data from various infectious bugs.

For more information, read The language of viruses.

EXAMPLES OF MACHINE LEARNING Applications -3

To expedite the development of COVID-19 vaccines, Pfizer used an ML tool (Smart Data Query) to analyze clinical trial data. This approach helped data scientists inspect datasets for inconsistencies during the collection of tens of millions of datapoints, in only 22 hours rather than 30 days, with maintaining high data quality.

For more information, read How a novel ‘incubation sandbox’ helped speed up data analysis in Pfizer’s COVID-19 vaccine trial.

PREDICTING THE NEXT PANDEMIC

Scientists use ML to investigate the vast amount of information about viruses, genomes, and other features of the viruses to learn patterns and recognize certain factors. Dr. Carlson, a biologist, recently pointed at mousepox as a potential threat to humans using ML models. Similarly, Dr. Barbara Han, a disease ecologist, used ML to create a list of high-priority rodent species that carry disease-causing agents and discovered those species who die young carry more pathogens. Read more here.

Researchers at Glasgow utilized an ML model (gradient boosting) to find animal viruses which could be transferred to humans. The models used a database, consisting of proportion of human-infecting variants in a given virus family, features of carrier species, and viral genomes. It then identified 313 potentially dangerous animal viruses. Read more here.

In a 2022 study, researchers claim that ML models can be used to enhance wildlife sampling for undiscovered viruses. Researchers designed statistical models to predict host–virus associations, using the bat hosts of betacoronaviruses as a case study. They declare that worldwide, over 400 bat species could be undetected betacoronavirus hosts. Read more here.

TYPES OF MACHINE LEARNING METHODS

Machine learning (ML) is based on giving a large volume of data to algorithms and having them learn from the data to predict, find patterns, recommend solutions, and classify data. There are three ML types that work differently from each other:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

SUPERVISED LEARNING

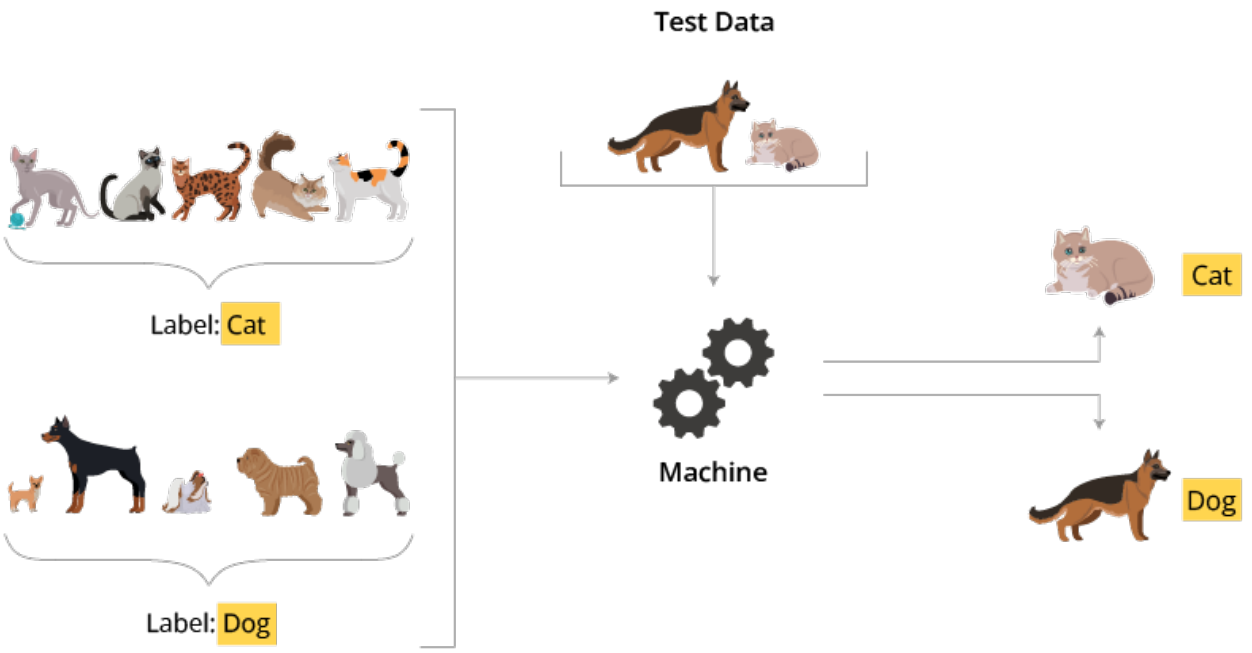
Supervised learning is one of the most popular types, in which machines are supervised or trained using **labelled data**. Labelled data contains observations with:

Input data, consisting of the properties of observations (e.g., shape, age, height, etc.)

Output data or the label of observation (e.g., cat/dog, diabetes/non-diabetes)

The properties of data are also called features. The label is also called the outcome. Supervised learning finds a mapping function to map the features to the output variable or outcome. It can be used in image classification, risk assessment, patient classification, fraud detection, and so on.

Supervised Learning



Characteristics of Supervised Learning Method

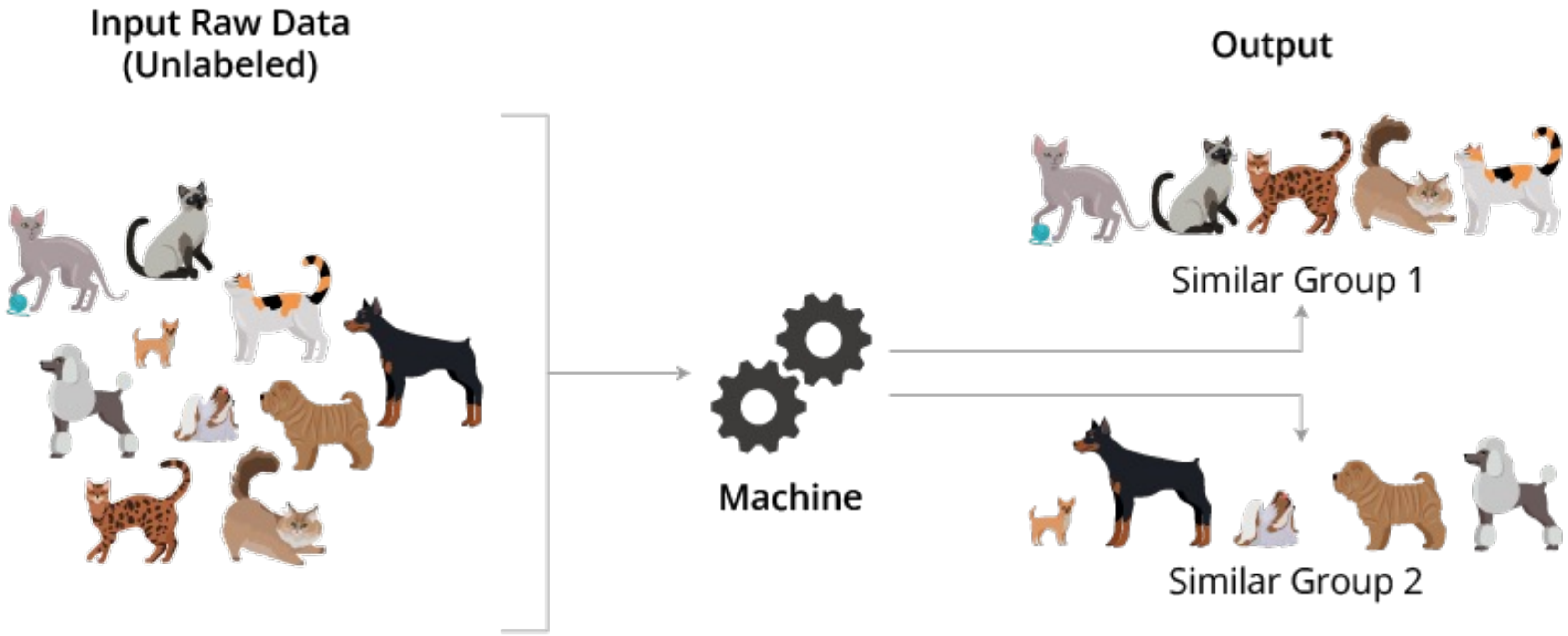
- Can predict output based on previous experiences
- There would be an exact idea about the classes of an object
- Widely used for solving real-word problems

- Not a suitable method for complex tasks
- Not able to predict with high accuracy if the test data is different from training dataset
- Execution time of training could be very high
- There should be sufficient knowledge about the classes of data

UNSUPERVISED LEARNING

- Unsupervised learning is used when we do not have labelled data, and therefore, cannot supervise the ML models with the relationships between data features and target variable. Such models find the hidden patterns of the data and group the data based on the found similarities.
- For example, unsupervised learning finds similarities among image features and then clusters them into different groups, such as dogs and cats. Unsupervised learning is very similar to how human brains learn to think by their own experiences.

UNSUPERVISED LEARNING



UNSUPERVISED LEARNING Characteristics

- Can be used for more complex tasks than supervised learning due to not requiring labelled data
- Easier to use due to no need for labelled data

- More difficult than supervised learning as there is no stated feature or target variable
- May have less accuracy as the algorithms do not know the exact output

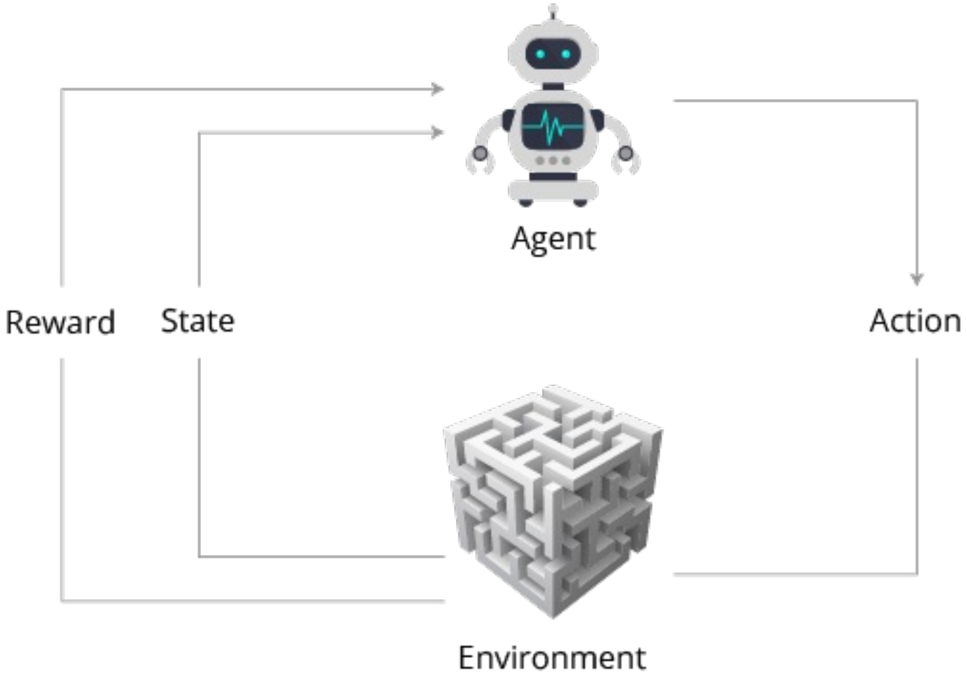
REINFORCEMENT LEARNING (RL)

This type of algorithm is based on an agent learning to behave properly in an environment in which it will be rewarded for a good outcome and punished for a bad one.

Therefore, it is mainly a feedback-based ML technique, where the agent learns by its own experiences. There is no labeled data in RL.

It is useful when the decision making is sequential and the goal is achieved in the long-term (e.g., playing games).

REINFORCEMENT LEARNING (RL)



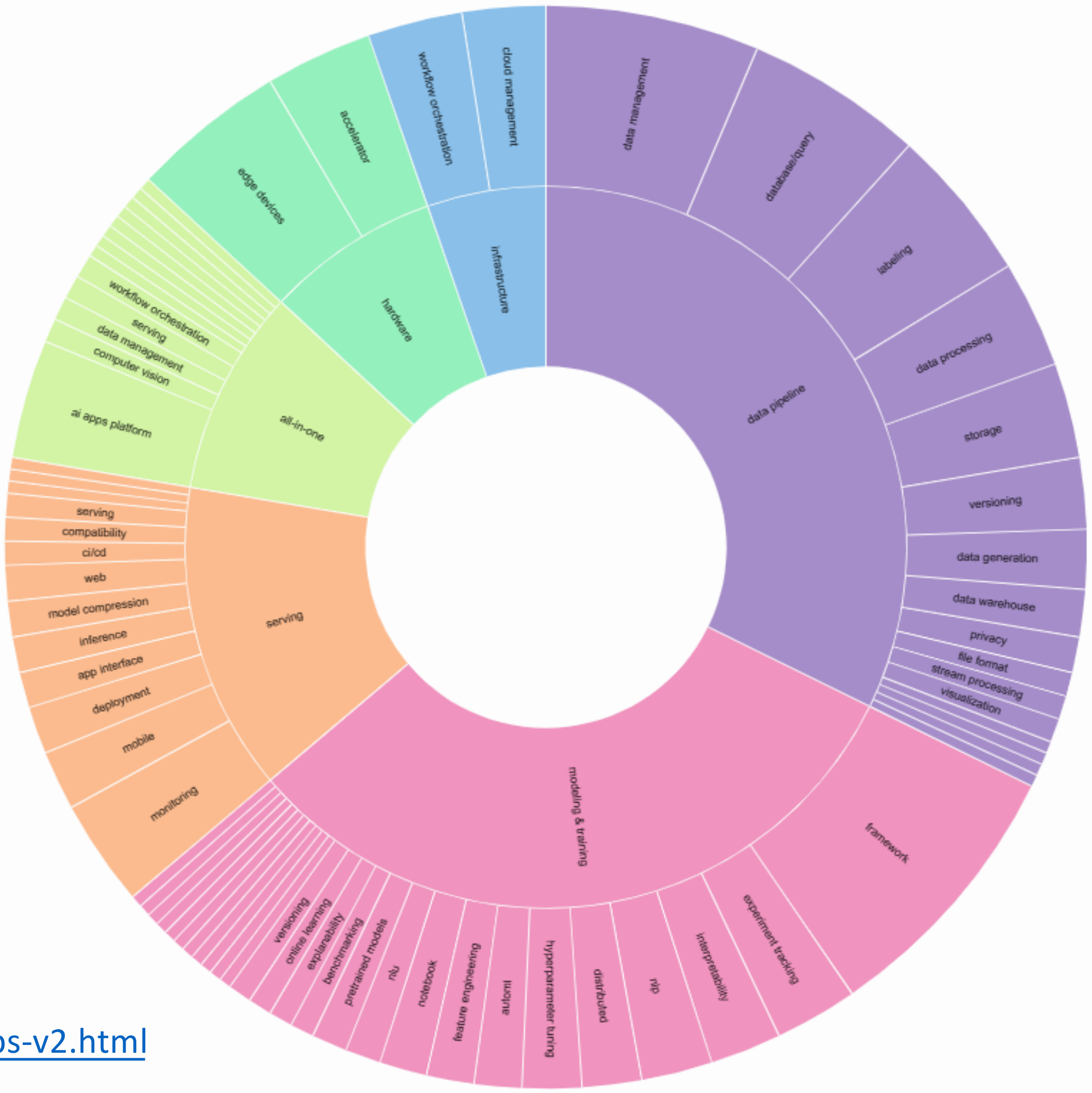
MACHINE LEARNING WORKFLOW

The success of an ML project relies on designing a valid research question and then following the below steps:

1. Gathering data: Using real-time data from Internet of things (IoT)-based devices or collecting it from other valid databases.
2. Pre-processing data: One of the most important requirements for a successful ML project. It is common to spend a great deal of time in this step.
3. Finding the model that best fits the data at hand: Selecting proper supervised/unsupervised learning or RL models.
4. Training and testing the model: Choosing a proper ratio of data for training ML models and testing their performance.
5. Evaluating the performance of the model: Selecting the best parameters for the model to achieve the highest accuracy.

ML Tools

- Increasing focus on deployment
- The Bay Area is still the epicenter of machine learning, but not the only hub
- MLOps infrastructures in the US and China are diverging
- More interests in machine learning production from academia



ML tools by Chip Huyen - source: https://huyenchip.com/2020/12/30/mlops-v2.html