# Introduction to Machine Learning: Fundamental Concepts of Machine Learning

Universidad Politécnica de Yucatán

Robotics 9B

Marco Alejandro González Barbudo

Professor's Signature

Machine learning incorporates several hundred statistical-based algorithms and choosing the right algorithm or combination of algorithms for the job is a constant challenge for anyone working in this field. But before we examine specific algorithms, it is important to understand the three overarching categories of machine learning. These three categories are supervised, unsupervised, and reinforcement.

## I. Supervised and Unsupervised Learning

**Supervised learning** focuses on understanding patterns by establishing connections between variables and known outcomes while working with labeled datasets. In this paradigm, the machine is provided with sample data containing various features (referred to as "X") along with the corresponding correct output values (referred to as "y"). This knowledge of both the output and feature values classifies the dataset as "labeled." The algorithm then identifies underlying patterns within the data, ultimately constructing a model capable of applying the same rules to new data.

The dataset itself is a collection of labeled examples, denoted as {(xi, yi)}N i=1. Each xi is known as a feature vector, comprising multiple dimensions (j = 1,...,D) where each dimension, represented as x(j), describes the example in some way. The label yi can take various forms, including membership in a finite set of classes {1, 2,...,C}, a real number, or even more complex structures such as vectors, matrices, trees, or graphs.

Once the machine discerns the data's rules and patterns, it creates what we refer to as a model—a computational equation for generating outcomes based on the acquired rules from the training data. Once the model is prepared, it can be applied to new data and assessed for

accuracy. Only after successfully passing both the training and testing stages can it be confidently deployed for real-world applications.

In contrast**, Unsupervised learning** does not involve categorizing all variables and data patterns. Instead, the machine must uncover hidden patterns and generate labels using unsupervised learning algorithms. A prime example of this is the k-means clustering algorithm, commonly used in unsupervised learning.

In unsupervised learning, the dataset comprises unlabeled examples, expressed as {xi}N i=1. Again, each xi serves as a feature vector. The primary objective of an unsupervised learning algorithm is to construct a model that takes a feature vector x as input and either transforms it into another vector or derives a value suitable for solving practical problems.

One significant advantage of unsupervised learning lies in its ability to reveal previously unnoticed patterns within the data. Techniques like k-means clustering can serve as a steppingstone for further analysis once discrete groups have been identified.

In conclusion, the key difference between supervised and unsupervised learning lies in the availability of labeled data. Supervised learning relies on this labeled information to make predictions with precision and is well-suited for tasks with established outcomes. In contrast, unsupervised learning operates in scenarios where labeled data is scarce or absent, relying on the discovery of underlying patterns and structures. Both approaches have their unique strengths and are vital tools in the machine learning toolkit, catering to different problem domains and research objectives. The choice between the two depends on the specific goals and the nature of the data at hand.

## II. Probabilistic Model

Probabilistic models are an essential component of machine learning, which aims to learn patterns from data and make predictions on new, unseen data. They are statistical models that capture the inherent uncertainty in data and incorporate it into their predictions. Probabilistic models are used in various applications such as image and speech recognition, natural language processing, and recommendation systems. In recent years, significant progress has been made in developing probabilistic models that can handle large datasets efficiently.

These models can be classified into the following categories:

- **Generative models:**
  Generative models aim to model the joint distribution of the input and output variables. These models generate new data based on the probability distribution of the original dataset. Generative models are powerful because they can generate new data that resembles training data. They can be used for tasks such as image and speech synthesis, language translation, and text generation.
- **Discriminative models:**
  The discriminative model aims to model the conditional distribution of the output variable given the input variable. They learn a decision boundary that separates the different classes of the output variable. Discriminative models are useful when the focus is on making accurate predictions rather than generating new data. They

can be used for tasks such as image recognition, speech recognition, and sentiment analysis.

- **Graphical models:**
  These models use graphical representations to show the conditional dependence between variables. They are commonly used for tasks such as image recognition, natural language processing, and causal inference.

Probabilistic models are of paramount importance in the realm of machine learning as they offer a structured approach to comprehending intricate patterns and intricacies inherent in extensive datasets. These models serve to intuitively gauge the likelihood of diverse outcomes, shedding light on the inherent data structure. Additionally, probabilistic models empower researchers and practitioners to make sound judgments when confronted with uncertain scenarios. One notable advantage lies in their capacity to facilitate Bayesian inference, a potent technique for revising our convictions regarding a hypothesis as fresh data emerges. This capability proves especially valuable in scenarios demanding decisions amidst uncertainty.

**Advantages Of Probabilistic Models**

- Probabilistic models are an increasingly popular method in many fields, including artificial intelligence, finance, and healthcare.
- The main advantage of these models is their ability to take into account uncertainty and variability in data. This allows for more accurate predictions and decision-making, particularly in complex and unpredictable situations.
- Probabilistic models can also provide insights into how different factors influence outcomes and can help identify patterns and relationships within data.

**Disadvantages Of Probabilistic Models**

- One of the disadvantages is the potential for overfitting, where the model is too specific to the training data and doesn't perform well on new data.
- Not all data fits well into a probabilistic framework, which can limit the usefulness of these models in certain applications.
- Another challenge is that probabilistic models can be computationally intensive and require significant resources to develop and implement.

## III. Regression and Classification

Regression and classification are two fundamental techniques in machine learning, each serving distinct purposes based on their characteristics and applications.

**Regression** involves finding a model or function that predicts continuous real values, in contrast to classification, where data is categorized into discrete classes or discrete values. Regression is often used to analyze historical data and identify trends or patterns within it, with the model's performance assessed using error metrics since it predicts quantities.

Key characteristics of regression include:

- Target variables are continuous.
- It aims to find the best-fit line or model to represent data trends.

- Evaluation metrics such as Mean Squared Error, R2-Score, and MAPE are commonly used to assess performance.
- Regression tasks can be linear or non-linear, depending on the data relationships.

State-of-the-art regression algorithms that deliver optimal results by employing techniques like bagging and boosting include:

- Lasso Regression
- Ridge Regression
- XGBoost Regressor
- LGBM Regressor

In contrast, **classification** involves finding a model or function that categorizes data into multiple categorical classes or discrete labels, based on specific input parameters. The primary objective in classification is to predict discrete target variables, typically represented as class labels, using a set of independent features.

Key characteristics of classification include:

- Target variables are categorical labels.
- It aims to determine a suitable decision boundary that effectively segregates different classes within the target variable.
- Evaluation metrics such as Precision, Recall, and F1-Score are commonly used to evaluate performance.
- Classification deals with problems that can have binary or multiple discrete labels.

State-of-the-art classification algorithms designed to produce optimal results through techniques like bagging and boosting include:

- Decision Tree
- Random Forest Classifier
- K – Nearest Neighbors
- Support Vector Machine

In conclusion, regression and classification are fundamental techniques in machine learning with distinct purposes and characteristics. Regression predicts continuous numerical values, while classification categorizes data into discrete categories. The choice between them depends on the problem and the type of output variable. Utilizing state-of-the-art algorithms and techniques in both regression and classification enables accurate predictions and informed decision-making across various domains.

## IV. REFERENCES


[1] O. Theobald, *Machine Learning for Absolute Beginners: A Plain English Introduction*. Torraza Piemonte (TO), Italy: Jeremy Pedersen and Red to Black Editing's Christopher Dino, 2021.

[2] J. Hurwitz and D. Kirsch, *Machine Learning For Dummies, IBM Limited Edition*. John Wiley & Sons, Inc., 2018.

[3] O. Theobald, *Machine Learning for Absolute Beginners: A Plain English Introduction*. Torraza Piemonte (TO), Italy: Jeremy Pedersen and Red to Black Editing's Christopher Dino, 2021.

[4] "Probabilistic models in machine learning," GeeksforGeeks, https://www.geeksforgeeks.org/probabilistic-models-in-machine-learning/ (accessed Sep. 6, 2023).

[5] “Classification vs regression in machine learning,” GeeksforGeeks, https://www.geeksforgeeks.org/ml-classification-vs-regression/ (accessed Sep. 6, 2023).