What is Machine Learning? The Core Components

20.06.2025

What is Machine Learning?

* ML is the science of teaching computers to learn patterns from data and make decions or predictions, without being explicitly programmed for every single rule. * Imagine we want spam detection.

* Traditional programming:

```
if "free_money" in email_subject:
    return spam
if "lottery_winner" in email_subject:
    return spam
...
```

This approach is brittle and will always miss new types of spam.

* ML Approach: You don't write rules, you show computer examples (data). You provide thousands of emails that have been labeled by humans as "Spam" and "Not Spam". The computer learns, it automatically identifies the patterns associated with spam (e.g. certain words, unusual sender addresses, specific grammar) on its own. The result is a "model" that can now look at a new, unseen email and predict whether it's spam or not.

The core components:

To achieve this ML model needs two things from our data:

1. Features (X)

* The input variables, predictors, or attributes that you "show" to the model. They are like "clues" or "evidence" you give the model.
* has_word_free, len_of_email, is_sender_in_contacts, etc.
* The collection of all features is represented by a capital X.

2. Target (y)

* Output variable, the "answer", the value we're trying to predict.
* is_spam (which could be 1 for yes and 0 for no.)

---

* ML is the science of teaching computers to learn patterns from data and make predictions or decisions without being explicitly programmed.
* An ML model needs 2 things from our data: ① Features (the input variables for the model to learn from) ② Target (the value it will try to predict – only for supervised learning)

The Main Types of Machine Learning

20.06.2025

* ML problems are typically grouped into 3 broad categories based on the kind of data they use and the goal they are trying to achieve.

1. Supervised Learning

* Most common type of ML.
* "Supervised" b/c we provide the model with a "supervisor" — the labeled target variable y.
* Model examines X and corresponding y's to learn the mapping between them.
* Dataset must include target variables.
* "Is this email spam or not?", "Is this image a cat, dog, or a bird." "What will the price of this house?"

2. Unsupervised Learning

* No target variable. The model is given only the features and try to find patterns or structure within the data of its own, without a "supervisor" providing the right answers.
* "What interesting patterns or groups exist within my data?" "Can we segment our customers into different groups based on their behavior?"
* Dimensionality reduction is a type of unsupervised learning the goal of which is to simplify the data by reducing the number of features while retaining important information.

3. Reinforcement Learning

* This is about training an "agent" to make a sequence of decisions in an "environment" to maximize a cumulative "reward". The agent learns through trial and error.

---

* 3 categories of ML problems
  1. Supervised Learning: The model is provided both the features and correct answers.
  2. Unsupervised Learning: The model is provided with features only. It tries to discover patterns/structures within the data
  3. Reinforcement Training: No static data provided at all. The "agent" makes a sequence of decisions in an environment to learn through trial and error by receiving rewards or penalties.

* There is no static dataset. It requires an interactive environment where the agent can take actions and receive feedback (rewards or penalties).
* Training an AI to play chess. A robot learning how to walk.
* Reinforcement Learning is a more advanced field and is beyond the initial scope of our Kocyigit-ML project, but it's important to know it exists as a major pillar of AI.

Note on Deep Learning

* Deep learning is not a separate category of problem like Supervised or Unsupervised learning. Instead, it's a powerful set of techniques, based on deep neural networks, that can be used to solve problems within all 3 categories. It is particularly effective for complex tasks involving unstructured data like images, text, and sound.

Parameters vs. Hyperparameters

20.06.2025

* When we talk about "values" in a ML model, they fall into two very different categories:

1. Model Parameters

* The values that the model learns on its own from the training data. They are internal to the model and direct output of the training process.
* Think of these as the "knowledge" or the "memories" that the model gains during its "study session" (the .fit() process) You don't tell the model these values; it discovers them.
* These are the values that the model adjust during training to minimize the cost function.
* In linear regression ($y = mx + b$), m and b are model parameters. In scikit-learn these are found in model.coef_ and model.intercept_

2. Model Hyperparameters

* The configuration settings that you, the data scientist, must set before the training process begins. They control the behaviour of the learning algorithm itself.
* Think of these as the study strategy you give the model before it starts learning. Are you telling it to study for 5 hours or 50 hours? To focus on the bigger picture or memorize every detail.
* They are chosen based on experience, intuition, or a systematic tuning process.

---

* Parameters: The values that the model learns on its own from the training data and adjusts by itself to minimize the cost function.
* Hyperparameters: The configuration settings that the data scientist must set BEFORE the training process begins, chosen based on experience, intuition, or a systematic tuning process

* The number of neighbours to consider in K-Nearest Neighbors, the number of trees to build in Random Forest, etc.