Supervised vs Unsupervised Learning

Supervised learning trains on examples with answers; unsupervised learning finds hidden

Introduction

In the last lesson you learned that machine learning learns from examples. Now comes the first big split in the field: did those examples come with answers, or not? That one question divides almost all machine learning into two families.

Supervised learning is like studying with an answer book. Every training example comes with the correct answer — the label. You show the model a photo labelled "cat", another labelled "dog", and it learns to tell them apart. When you later show it a new photo, it predicts a label. The teacher (you, the data) supervised the learning.

Unsupervised learning is like being handed a big box of mixed items with no labels at all, and being asked to sort them into sensible groups. Nobody tells the model the right answer, because there is no answer yet — the job is to *find* the structure hiding in the data. Shop customers grouped by buying habits is the classic example: nobody labelled them, the groups emerge from the data itself.

Most beginner projects are supervised, because labelled data gives you a clear way to check your work. But unsupervised learning is how we explore data when we do not yet know what we are looking for.

There is also a practical cost hiding behind supervised learning: labels take human time. Every labelled photo, email or student record was checked by a person, or by a system a person trusted. That is why labelled data is expensive and why companies guard it. Unsupervised learning avoids that cost — it works on raw, unlabelled data — but it pays a different price: nobody can tell you for sure whether its groups are right, because there are no answers to check against.

A useful habit is to ask two questions before any project. First: do I already have answers for my examples? Second: if not, can I label a few hundred examples myself, or do I need to find groups first and label later? Many real projects start unsupervised — group the customers, look at the groups — and only then label the groups and train a supervised model. The two families are not rivals; they are two steps of the same workflow.

(Roman Urdu: Supervised mein har misaal ke saath jawab bhi hota hai — jaise answer book ke saath parhna. Unsupervised mein koi jawab nahi hota, model khud pattern dhoondta hai.)

Key points

- **Supervised = labelled data.** Each example has an input and the correct output (label). The model learns the mapping from input to label.
- **Unsupervised = unlabelled data.** Examples have inputs only. The model finds groups, patterns or structure on its own.

- **Supervised splits into two jobs: classification** (predict a category, like spam/not spam or pass/fail) and **regression** (predict a number, like a price or a temperature).
- **Unsupervised's most common job is clustering:** sorting similar items into groups, called clusters, without being told the groups in advance.
- **How to tell them apart:** ask "do my examples already have the answer attached?" If yes, supervised. If no, unsupervised.
- **Supervised is easier to grade.** Because you know the true answers for test data, you can count how often the model is right. Unsupervised results are harder to judge — a grouping can be sensible without being "correct".
- **Real projects often mix both.** A shop might first cluster customers without labels (unsupervised), then train a supervised model to predict which group a new customer belongs to.
- **Labels cost time and effort.** Someone must check every label, so labelled data is expensive; unlabelled data is cheap and plentiful, but harder to judge.
- **Semi-supervised learning mixes both.** It trains on a small labelled set plus a large unlabelled set — a common real-world compromise when labelling everything is too costly.

Examples

Example 1 — Supervised: pass or fail (classification)

A teacher has records of 10 students: hours studied and whether each passed or failed. Those are labelled examples — every row has the answer.

Hours studied	Result
1	Fail
2	Fail
3	Fail
5	Pass
6	Pass
8	Pass

A supervised model learns the pattern: low hours go with Fail, higher hours with Pass. Give it a new student who studied 7 hours, and it predicts Pass. The labels made the learning possible — and testable.

Example 2 — Supervised: predict a number (regression)

A fruit seller records the weight of 8 mangoes and their prices. Weight is the input, price is the label — a number, not a category. The model learns how price rises with weight, and can then predict the price of a new mango from its weight alone. Number prediction is regression, and you will build one in lesson ML.4.

Example 3 — Unsupervised: grouping shoppers (clustering)

A shop has 12 customers, each described by how much they spend per visit and how often they visit. Nobody labelled them. A clustering method groups them anyway: one group spends little but visits often, another spends a lot but visits rarely, a third spends little and visits rarely. The shop did not know these groups existed — the data revealed them. That is unsupervised learning, and lesson ML.8 goes deeper into it.

Example 4 — Worked example: supervised, predicting marks from hours

A teacher has six labelled rows — hours studied and the exam marks actually scored:

Hours	Marks
2	35
4	50
6	65
8	80

The pattern is steady: about 7 to 8 extra marks per extra hour. A supervised model learns that pattern, so for a new student who studied 5 hours it predicts roughly 57 to 58 marks. The labels (marks) made this possible, and they also make it checkable: when the real marks come in, you can count how close the predictions were. If the new student's real mark is 90, the model missed by a lot — and that miss is visible, measurable evidence, not a feeling.

Example 5 — Worked example: unsupervised, grouping songs by listening habits

A music app has no labels for its listeners, only numbers: how many hours of pop, rock and classical music each person played last month.

| Listener | Pop hours | Rock hours | Classical hours | |---|---|---|---| | A | 20 | 2 | 0 | | B | 18 | 3 | 1 | | C | 1 | 19 | 2 | | D | 2 | 20 | 0 | | E | 3 | 2 | 15 | | F | 2 | 1 | 18 |

By eye, three groups appear: A and B (pop lovers), C and D (rock lovers), E and F (classical lovers). Nobody labelled them; the groups came from the numbers alone. The app can now recommend new pop songs to group one without ever being told what "pop lover" means. But note the limit: if the app later calls group one "young listeners", that label is a human guess, not something the data proved — age was never in the table.

Real-world case — How a shop uses both, in order

An online shop first clusters its customers without labels: frequent small buyers, rare big spenders, and one-time visitors. That is unsupervised — exploration. Then the shop labels each group by hand ("loyal", "premium", "passing through") and trains a supervised model: given a brand new customer's first two orders, which group will they belong to? That model then decides who gets a loyalty offer and who gets a premium offer. Step one found the groups; step two turns new customers into decisions. Most real projects look like this, not like one pure method.

Quick recap

- Supervised learning uses labelled examples; unsupervised uses unlabelled data.
- Supervised jobs: classification (a category) and regression (a number).
- Unsupervised jobs: clustering (finding groups) and finding hidden patterns.
- The test: do my examples have answers attached? Yes = supervised, no = unsupervised.
- Supervised is easier to grade because true answers exist to check against.

Common mistakes

1. **Calling everything "AI learning" without naming the type** — the type decides which method and which checks you use. Fix: name it first — supervised or unsupervised — before choosing a model.
2. **Using supervised methods on unlabelled data** — without labels there is nothing to learn from or check against. Fix: either label some data, or switch to a clustering method.
3. **Thinking clustering gives "correct" groups** — clusters are suggestions from the data, not facts. Fix: check whether the groups make sense for your real problem before acting on them.

1. **Labelling data carelessly and trusting the model anyway** — wrong labels teach wrong patterns, and the model will repeat them confidently. Fix: check a sample of labels by hand before training, and fix disagreements first.
2. **Assuming an unsupervised grouping is the final answer** — clusters suggest, they do not prove. Fix: show the groups to someone who knows the real problem and ask whether they make sense before acting on them.

Practice

1. For each of these, write S (supervised) or U (unsupervised) and one line why: (a) predicting tomorrow's temperature, (b) grouping news articles by topic, (c) detecting spam email, (d) finding types of music listeners.
2. Take the pass/fail table above and add two more rows of your own. Then predict, by eye, the result for a student who studied 4 hours — and say why you are unsure.
3. Think of one grouping problem in your own life (books, friends, expenses). Write down what "similar" would mean there.

1. A hospital wants to (a) predict which patients may return within a month, and (b) find hidden groups among patients by symptoms. Write S or U for each, and name the label (if any) in one line.
2. You have 500 photos with no labels and time to label only 50. Write a short plan: which 50 would you label first, and what would you do with the other 450?
3. Take the song-listener table in Example 5 and add a seventh listener: pop 10, rock 10, classical 0. Which group would they join, and why is this case harder to place than the others?

Practice quiz

1. What is supervised learning?

- Learning without any data
- Learning from labelled examples, where the right answers are given
- Learning by watching videos
- Learning with a teacher in the room

Answer: Learning from labelled examples, where the right answers are given

2. What is unsupervised learning?

- Learning with answer keys
- Finding groups and patterns in data without labels
- Learning slowly
- Learning from the internet

Answer: Finding groups and patterns in data without labels

3. What is the difference between regression and classification?

- Regression is faster
- Regression predicts a number, classification predicts a category
- Classification predicts numbers
- There is no difference

Answer: Regression predicts a number, classification predicts a category

---

Watch the video lesson: https://youtu.be/SISxMosaAgk

Read online: https://scopetreasure.com/notes/ml-supervised-unsupervised · More free tools and notes at https://scopetreasure.com