

Overfitting, explained in three curves.

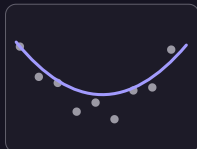
D8A.ACADEMY/RESOURCES/OVERFITTING

01 Three fits, same data

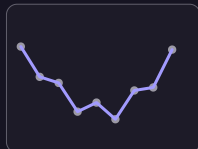
Same points. Three models. Only one learned the pattern.



UNDERFIT



GOOD FIT



OVERFIT

→ Underfit is too simple to see the shape. Overfit memorized every wiggle of noise. Good fit learned the trend.

02 Diagnose yours

Compare train error vs test error. The gap is the diagnosis.

UNDERFITTING

- Train error: high
- Test error: high too
- Model too simple
- Fix: more features, capacity

OVERFITTING

- Train error: near zero
- Test error: high
- Model memorized noise
- Fix: simplify, regularize

→ A student who memorized past exams aces practice tests and fails the real one. That is overfitting.

03 Five fixes that work

In rough order of how often they are the right answer.

more data	noise averages out with volume
simpler model	fewer parameters to memorize with
regularization	L1 / L2 penalties on extreme weights
early stopping	quit while validation still improves
cross-validation	5 splits for an honest estimate

△ COMMON TRAPS

04 Trap: test-set leakage

Scale or encode BEFORE the split and your test score is fiction.

the leak	preprocessing saw the test rows too
the symptom	great offline, embarrassing in production
the fix	split first; fit transforms on train only

→ This is exactly why sklearn Pipeline exists. Use it and the leak becomes impossible.