

Training vs Testing

Data: ML's Secret Sauce



The Core Idea

- Machine learning models learn patterns from data.
- This process is split into two critical phases.
- The training phase and the testing phase.
- Each phase uses a different set of data.
- Using separate data is the key to honesty.
- It prevents the model from cheating on the exam.



Meet Training Data

- This is the model's textbook and study guide.
- It is a large dataset used for learning.
- The model analyzes it to find patterns.
- It adjusts its internal parameters based on this data.
- The goal is to learn general rules.
- More high-quality training data usually means better learning.



The Learning Process

- The model makes predictions on training data.
- Its errors are used to correct itself.
- This correction is called "fitting" or "training".
- The process repeats many times for accuracy.
- Think of it as practicing with flashcards.
- The model memorizes the material in the textbook.



Meet Testing Data

- This is the model's final exam.
- It is a separate, unseen dataset held in reserve.
- The model never learns from this data.
- It is used only after training is complete.
- The model makes predictions on this new data.
- This evaluates how well it learned general lessons.



Why Split The Data?

- To test the model's performance on new information.
- It checks for a problem called overfitting.
- Overfitting is when a model memorizes the training data.
- It performs perfectly on training data but fails on new data.
- It hasn't learned general concepts, just specific answers.
- The test data reveals this critical weakness.



The Goldilocks Zone

- We want a model that generalizes well.
- It should perform well on both data sets.
- Good performance on test data means it learned.
- The split is often 70–80% for training, 20–30% for testing.
- The exact split can vary based on data size.
- The key is to keep the test data completely separate.



Common Pitfalls

- Using test data during training invalidates the results.
- It is like giving a student the exam answers beforehand.
- Data leakage undermines the entire evaluation process.
- Using too little test data gives unreliable scores.
- Always randomize the split to avoid hidden biases.
- A good split is foundational for trustworthy ML.



Key Takeaway

- Training data is for learning the material.
- Testing data is for the unbiased final exam.
- They must be strictly separated to ensure honesty.
- This split is a non-negotiable best practice.
- It is the only way to measure real-world performance.
- Mastering this split is a fundamental ML skill.



Follow for More



Like this post? Repost!

Share your thoughts in the comments!

I share little hacks and tips I actually use in
Data & AI — follow if you wanna see more!



Shoaib Akthar
@theshoaibakthar

Follow for more Data Analytics, Data Science and AI insights