What is Overfitting?

Learn the causes and effects of overfitting in machine learning, and how to address it to create models that can generalize well to new data.

Contents

Aug 2023 · 5 min read

Abid Ali Awan
I am a certified data scientist who enjoys building machine learning applications and writing blogs.

TOPICS

Artificial Intelligence (AI)

Machine Learning

Overfitting is a common challenge in machine learning where a model learns the training data too well, including its noise and outliers, making it perform poorly on unseen data. Addressing overfitting is crucial because a model's primary goal is to make accurate predictions on new, unseen data, not just to replicate the training data.

Overfitting Explained

In machine learning, the ultimate goal is to create models that can generalize well to new data. Overfitting occurs when a model becomes too closely adapted to the training data, capturing even its random fluctuations. Imagine teaching a child to recognize birds, but instead of teaching general bird characteristics, you only show them pigeons. If they see an eagle, they might still think it's a pigeon because that's all they know.

The causes of overfitting can be numerous:

- **Complex models.** Using an overly complex model for a simple task can lead to overfitting. For instance, using a high-degree polynomial regression for data that's linear in nature.
- **Insufficient data.** If there's not enough data, the model might find patterns that don't really exist.
- **Noisy data.** If the training data contains errors or random fluctuations, an overfitted model will treat these as patterns.

The impact of overfitting is significant. While an overfitted model will have high accuracy on its training data, it will perform poorly on new, unseen data because it's not generalized enough.

How to Detect Overfitting

Detecting overfitting is a crucial step in the machine learning process. Here's how you can spot it:

- **Validation set.** Split your data into training and validation sets. If your model performs well on the training set but poorly on the validation set, it's likely overfitting.
- **Learning curves.** Plot the model's performance on both the training and validation sets over time. If the two curves start to diverge, it's a sign of overfitting.
- **Cross-validation.** Use cross-validation, where the training data is split multiple times and the model is evaluated on each split.

It's especially important to check for overfitting when:

- You're using a complex model.
- You have a small amount of data.
- The stakes are high, like in medical diagnoses.

How to Prevent Overfitting

Preventing overfitting is better than curing it. Here are some steps to take:

- **Simpler models.** Start with a simpler model and only add complexity if necessary.
- **More data.** If possible, collect more data. The more data a model is trained on, the better it can generalize.
- **Regularization.** Techniques like L1 and L2 regularization can help prevent overfitting by penalizing certain model parameters if they're likely causing overfitting.

- **Dropout.** In neural networks, dropout is a technique where random neurons are "dropped out" during training, forcing the network to learn more robust features.

See our full tutorial on **how to prevent overfitting in machine learning**.

Overfitting vs Underfitting

While overfitting is a model's excessive adaptation to training data, underfitting is the opposite. An underfitted model fails to capture even the basic patterns in the training data.

- **Overfitting:** High accuracy on training data, low accuracy on new data. Imagine a GPS that works perfectly in your hometown but gets lost everywhere else.
- **Underfitting:** Low accuracy on both training and new data. It's like a GPS that can't even navigate your hometown.

Both overfitting and underfitting lead to poor predictions on new data, but for different reasons. While overfitting is often due to an overly complex model or noisy data, underfitting might result from an overly simple model or not enough features.

Overfitting: The Ongoing Struggle for ML Engineers

As machine learning (ML) engineers, we are constantly seeking to build the most accurate models possible. However, overfitting is one of the major risks that comes with pursuing high accuracy.

Many companies fall into the trap of overfitting. They see that high training accuracy and assume they have developed an excellent model. Unfortunately, when they deploy the model in the real world, it breaks down completely. It's like acing all your practice exams, but then failing the actual test.

As ML engineers, we have to resist the temptation of chasing perfect accuracy on training data alone. You simply cannot get 100% accuracy on training data and expect that to translate to new data. We have to use techniques like cross-validation, regularization, data augmentation, and ensembling to ensure our models generalize well.

The journey of machine learning is often one of starting with an underfit model and slowly improving accuracy through iteration. But there comes a point where additional tweaks start leading to overfitting. We have to walk that fine line between under and overfitting to find a Goldilocks model that performs well in all situations.

Want to **learn more about AI** and machine learning? Check out the following resources:

- **Introduction to ChatGPT course**
- **Generative AI concepts course**

- 7 AI projects for all levels
- What is AI literacy?

FAQs

Why is overfitting bad?

Overfitting reduces a model's ability to generalize to new data, which is its primary goal.

How can I prevent my model from overfitting?

Use simpler models, gather more data, apply regularization techniques, and use dropout in neural networks.

Can overfitting be good in some scenarios?

In very rare cases, if the training data is extremely representative and comprehensive, overfitting might not be detrimental. However, in most real-world scenarios, it's undesirable.

AUTHOR

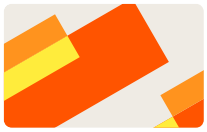
Abid Ali Awan

I am a certified data scientist who enjoys building machine learning applications and writing blogs on data science. I am currently focusing on content creation, editing, and working with large language models.

TOPICS

Artificial Intelligence (AI) Machine Learning

Related



The 10 Best Custom GPTs on the GPT Store

Nisha Arya Ahmed

Understanding and Mitigating Bias in Large Language Model...

Nisha Arya Ahmed

Inside Algorithmic Trading with Anthony Markham, Vice...

Richie Cotton

See More →

Grow your data skills with DataCamp for Mobile

Make progress on the go with our mobile courses and daily 5-minute coding challenges.

Download on the App Store

GET IT ON Google Play

LEARN

Learn Python

Learn R

Learn AI

Learn SQL