Learning Curves

Bias vs. Variance Tradeoff

What is learning curve?

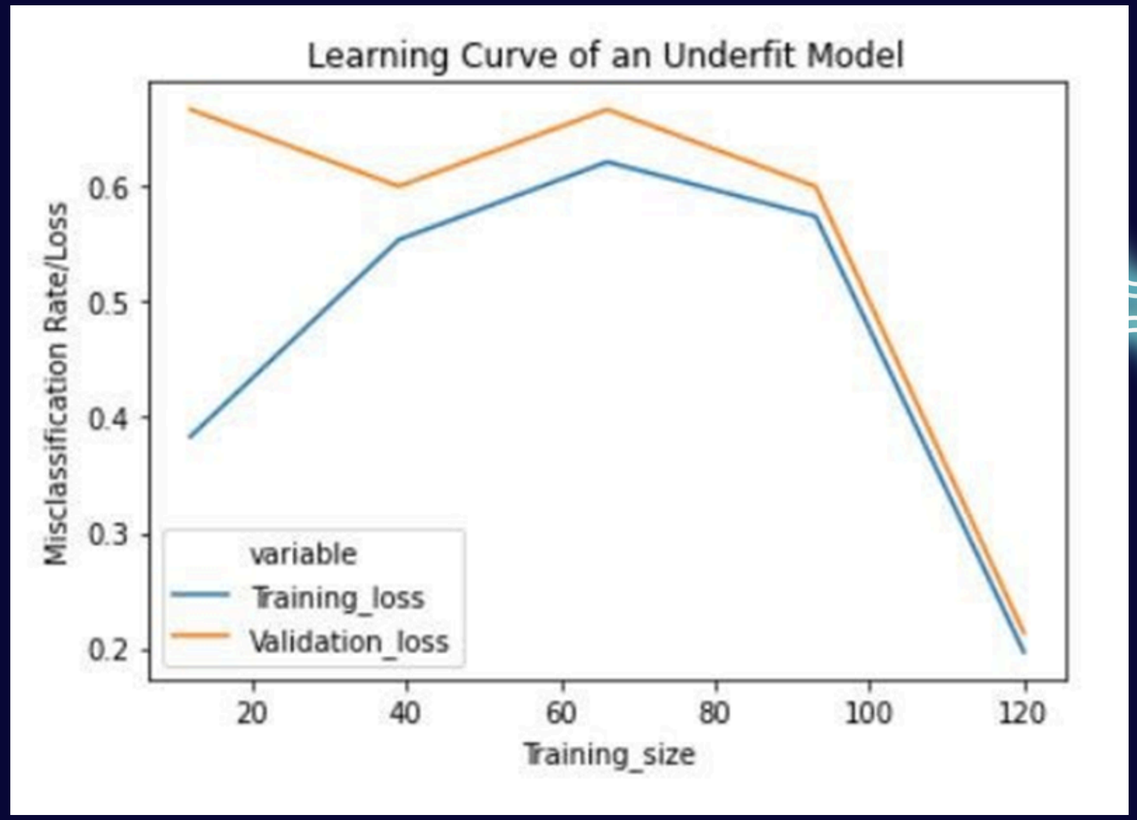
- Learning curves are graphs that show how a model improves as it learns from more data over time. They help us understand if the model is learning well or struggling. It works by evaluating a model on the training and validation datasets, then plotting the measured performance.

- Used to diagnose whether a model suffers from underfitting (high bias) or overfitting (high variance).

How it works

1. Underfitting (High Bias)

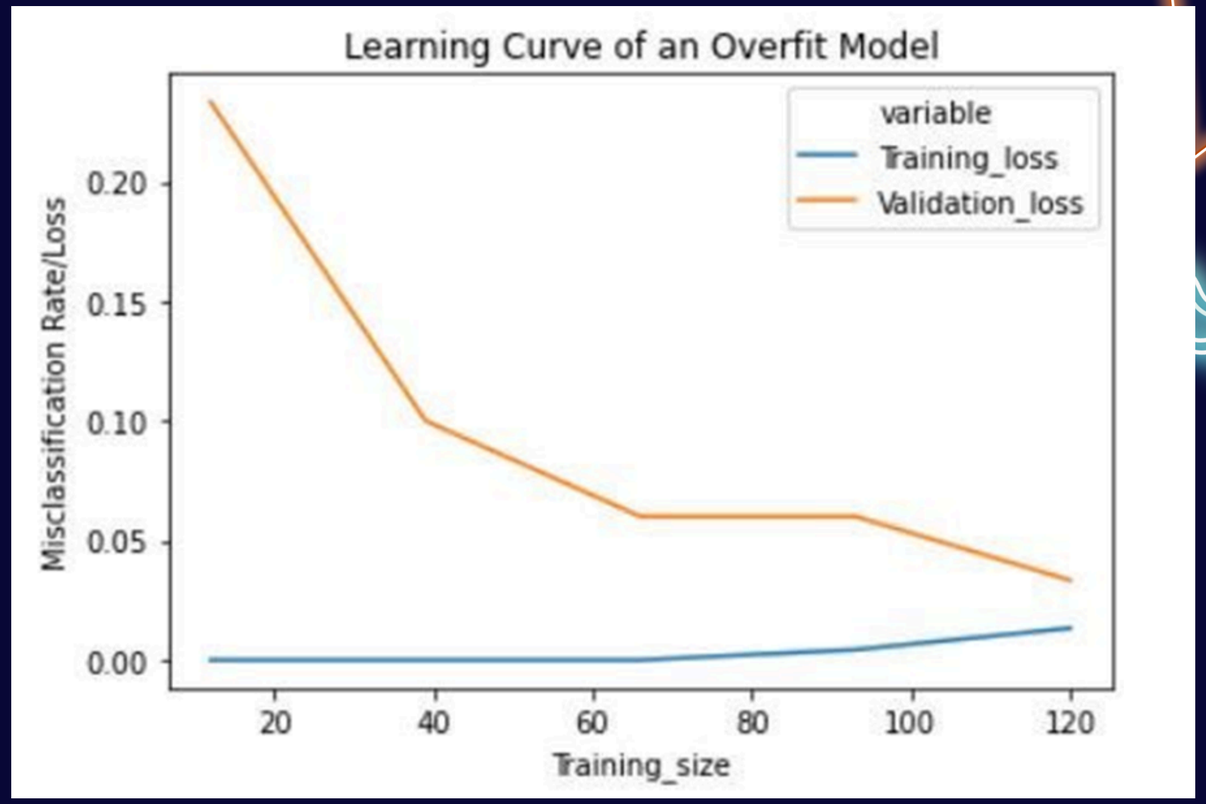
- Increasing training loss upon adding training examples
- Training loss and validation loss are close to each other at the end.
- Sudden dip in the training loss and validation loss at the end (not always).



How it works

2. Overfitting (High Variance)

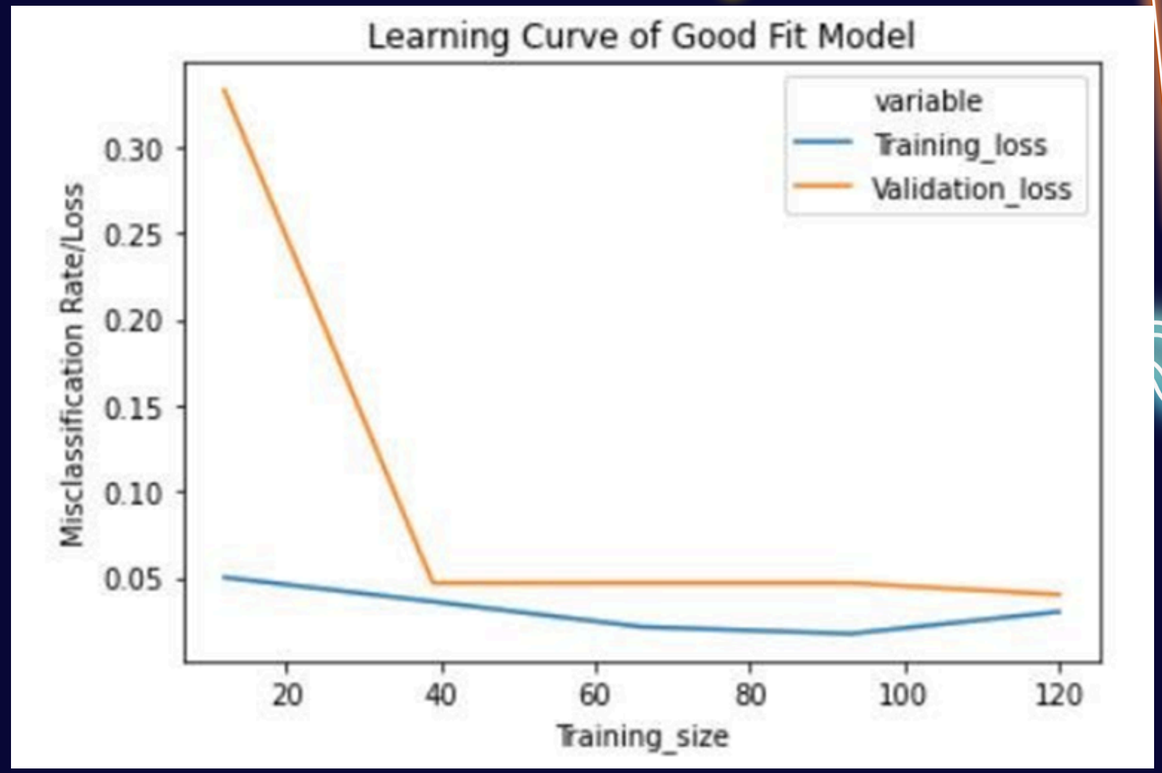
- Training loss and Validation loss are far away from each other.
- Gradually decreasing validation loss (without flattening) upon adding training examples.
- Very low training loss that’s very slightly increasing upon adding training examples.



How it works

3. Good Fit

- Training loss and Validation loss are close to each other with validation loss being slightly greater than the training loss.
- Initially decreasing training and validation loss and a pretty flat training and validation loss after some point till the end.



Real-World Applications

1.Healthcare Diagnostic

Learning curves help determine if a model predicting disease (e.g., from X-rays) needs more patient data or a simpler architecture to generalize better across hospitals.

2. Autonomous Vehicles – Object Detection

Learning curves help engineers track how model performance changes with more training images. They use it to decide if the system needs more data or simpler models for real-time accuracy.

3. Fraud Detection in Finance

Used learning curves to assess whether their fraud detection models are overfitting to past fraud patterns, ensuring they catch new types of fraud, not just memorizing old ones.

Research Trend:
Comparison of multiple modalities for drug response prediction with learning curves using neural networks and XGBoost

Authors:

- Nikhil Branson
- Pedro R. Cutillas
- Conrad Bessant.

Country: United Kingdom

Created at: December 23 2023

A 2023 study, “Comparison of multiple modalities for drug response prediction with learning curves using neural networks and XGBoost” explored how different biological data types can predict cancer drug responses. The researchers used Neural Networks and XGBoost with transcriptomics, proteomics, and phosphoproteomics data, and applied learning curves to track model performance as data size increased. The curves revealed which datasets improved faster and which ones needed more data to perform well. Learning curves also helped detect overfitting and underfitting, showing how data quantity affects accuracy. Overall, the study showed that learning curves are valuable tools for analyzing model behavior and guiding better data and model decisions in real-world machine learning research.

Reflection

- Learning curves help us see how well a machine learning model is learning. They show two lines, one for how the model does on the data it was trained on, and another for how it does on new, unseen data. If both lines are high and close together, the model is underfitting it’s too simple and can’t learn the patterns. If the training line is very low but the other line is much higher, the model is overfitting it’s memorizing the training data instead of learning to make good predictions on new data. When both lines are close and low, and they flatten out, that means the model is doing a good job and will likely work well in real situations.

References

- Muralidhar, K. (2025, March 5). Learning Curve to identify Overfitting and Underfitting in Machine Learning. Towards Data Science. https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5/

- Datacamp. (2022, March 9). https://www.datacamp.com/tutorial/tutorial-learning-curves

- Branson, N. (2023). Comparison of multiple modalities for drug response prediction with learning curves using neural networks and XGBoost. Bioinformatics Advances, 4(1), vbad190.

Thank You!!

Prince Genel R. Umali