Prince Genel R. Umali

BSCS-3B

Learning Curve

Building a model in machine learning is the first step, but the problem is understanding how well the model is actually learning from the data provided. Whether the model is just memorizing the dataset or the model is not learning properly. To track this progress, we use a learning curve. This report explores what learning curves are, how they visualize the critical balance between bias and variance, and how they are used in real-world applications to optimize model performance.

Learning curve is essentially a graph that plots a model's performance over time as it gains experience. The experience is usually represented by the number of training examples. These curves help us determine if a model is learning or not. The graph displays two distinct lines: the training score or error, which shows how well the model predicts on data it has already seen, and the validation score, which shows how well it performs on new, unseen data. By comparing these two lines, we can identify specific problems with the model's fit.

Interpreting a learning curve is about diagnosing the tradeoff between bias and variance. When we look at the behavior of the training and validation lines, we can identify three main states of the model. First, we might encounter underfitting, also known as high bias. This occurs when the model is too simple to capture the underlying patterns in the data. On the graph, this appears as a convergence to a poor score. The training loss and validation loss will be high and sit close to each other at the end of the training process. Often, there might be a sudden dip in losses, but they never reach a satisfactory low level because the model just doesn't have the complexity to learn. On the contrary, we have an overfitting, or high variance. This happens when the model doesn't just learn the patterns but memorizes the training data, including its noise. It can be spotted on a learning curve when there is a large gap between two lines. The training loss will be very low, indicating near-perfect performance on known data, but the validation loss will remain high or decrease very gradually without flattening out. The last state is the good fit, in this scenario, both the training and validation losses decrease initially and then flatten out. At the end of the training, the two lines should be close to each other, with the validation loss being only slightly higher than the training loss. This indicates that the model has learned the patterns well and can generalize that knowledge to real-world situations.

One major application of the learning curve is in medical image analysis. For example, when training a model to analyze medical scans, collecting labeled data is often difficult and expensive. Engineers look at the learning curve to decide if they need more data. If the validation accuracy keeps rising as more data is added, the curve suggests that collecting more samples will improve the model. Another application is fraud detection in finance, using learning curves to assess whether their fraud detection models

are overfitting to past fraud patterns, ensuring they catch new types of fraud rather than just memorizing old ones.

A recent study by Branson et al (2023), published in Bioinformatics Advances, highlights the continued relevance of learning curves in complex fields like drug response prediction. The researchers used learning curves to compare two different approaches of neural networks and XGBoost for predicting how cancers respond to drugs using biological data. The study found that learning curves were essential for tracking performance as the dataset size increased. The results showed that XGBoost achieved higher stability and accuracy with less data, whereas neural networks required significantly more data to improve. Furthermore, the curves helped the researchers identify specific signs of overfitting, allowing them to determine how data size directly affected the reliability of their predictions. This proves that learning curves remain an effective method for optimizing data usage in modern biomedical research.

In conclusion, learning curves are crucial because they act as a diagnostic tool. Without them, I wouldn't know if a model's poor performance was due to a lack of data or a poor choice of algorithm. Understanding how to interpret these curves allows me to make decisions like knowing when to stop training or when to look for more data or model.

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.