Learning Curves and Model Complexity

Learning Curves

Let us explore the relationship between bias and variance on a model's ability to learn from data through visual graphs.

A learning curve in machine learning is a visual graph that compares the metric performance of a model on training and testing data over a number of training instances.

When we look at the relationship between data and error we should generally see a downward trend of error as the number of training points increases. This should make sense since we are trying to build models that learn from experience.

We separate training and testing sets so we can get a better idea whether the model can generalize to unseen data rather than fit to the data just seen.

You can verify in a learning curve when a model has learned as much as it can about the data when both the training and testing curves plateau and no difference between the two gaps change.

Learning Curves

Bias

When the training and testing errors converge and are quite high this essentially means the model is biased. No matter how much data we feed it, the model cannot represent the underlying relationship and therefore has systematic high errors.

Variance

When we have a large gap between the training and testing error this essentially means the model suffers from high variance. Unlike a bias model, models that suffer from variance can generally improve if we have more data to learn from or we can simplify the model representing the most important features of the data.

Ideal Learning Curve

The ultimate goal for a model is one that has low errors and generalizes pretty well for unseen data (testing data). We can see this when both the testing and training curves converge and where the error is extremely low. That is the model is very accurate on unseen data.

Model Complexity

Unlike a learning curve graph, a model complexity graph looks at how the complexity of a model changes the training and testing curves rather than the number of data points to train on. The general trend of is that as a model increases, the more variability exists in the model for a fixed set of data.

Learning Curves and Model Complexity

So what is the relationship between learning curves and model complexity?

If we were to take the learning curves of the same machine learning algorithm with the same fixed set of data, but create several graphs for the increase model complexity, all the learning curve graphs would represent the model complexity graph. That is if we took the final testing and training errors for each model complexity and visualized them along the complexity of the model we would be able to see how well the model performs as the model increases.

Practical use of Model Complexity

Knowing that we can identify issues with bias and variance by analyzing a model complexity graph, we now have a visual tool to help identify ways to optimize our models. In the next section we will go over gridsearch and how to fine tune our models for better performance.

Summary

Congratulations on completing this course!

To recap: We reviewed some basic statistics and looked at several metrics to evaluate how well a model is learning based on the problem at hand.

Next we went over several data types and how to split data to verify that our models are indeed learning to generalize from unseen data rather than to the training set we gave it.

Then we looked at two common errors that our models can suffer from: Bias due to under-representing or underfitting the underlying data and variance; and model complexity which overfits the training data and no longer generalizes well.

Finally we looked at model complexity and used gridsearch to identify optimal parameters for our models.