Double descent phenomenon

1.Fay Elhassan 2. Vongai Mitchell 3. Bahati Kilongo

April 10, 2023

1 Bias - variance Tradeoff

Model complexity

The complexity of a model is its capacity to capture a complex underlying process. It refers to the number of predictor or independent variables or features that a model needs to take into account in order to make accurate predictions.

The following are key factors that govern the model's complexity and impact the model's accuracy:

- The Number of parameters.
- The norm used.
- The number of training examples.
- For the neural networks, we can add the number of hidden layers, increase number of neurons in each layer, and alter the form of activation functions.

Bias-variance Tradeoff

In the context of machine learning and model evaluation, the generalization error refers to the overall error of a model when applied to unseen data, and it is a key concept in evaluating the performance of a model. The generalization error is composed of three components: bias, variance, and irreducible error.

Bias: Bias refers to the error introduced by the simplifying assumptions or limitations of a model in capturing the true underlying relationship between the data points. It quantifies the deviation of the model predictions from the true values. A high bias indicates that the model is too simplistic and unable to capture the true complexity of the data, leading to underfitting. On the other hand, a low bias means that the model is more capable of capturing the true relationship between data points. Mathematically, bias can be calculated as:

$$\text{Bias} = E[f(x)] - f\,\text{true(x)}$$

where:

f(x) represents the predictions of the model for a given input x
E[f(x)] represents the expected value or average of the model predictions over different datasets
$f_{\text{true}}(x)$ represents the true underlying relationship between the data points

Variance: Variance refers to the variability or spread in the predictions of a model for different datasets. It quantifies how much the predictions of a model change when trained on different subsets of data. A high variance indicates that the model is sensitive to the training data and may overfit, capturing noise or random patterns in the data. On the other hand, a low variance means that the model is more stable and consistent in its predictions. Mathematically, variance can be calculated as:

$$\text{Variance} = E[(f(x) - \text{E[f(x)]})^2]$$

where:

f(x) represents the predictions of the model for a given input x
E[f(x)] represents the expected value or average of the model predictions over different datasets

Irreducible error: Irreducible error represents the inherent noise or randomness in the data that cannot be reduced by any model. It is the minimum error that any model would have, regardless of its complexity or performance. The relationship between bias, variance, and the generalization error can be expressed by the following equation:

$$\text{Generalization error} = Bias^2 + \text{Variance} + Irreducible error$$

This equation highlights that the generalization error is the sum of the squared bias, variance, and irreducible error. The goal in model training is to find a balance between bias and variance, as reducing one may increase the other. The bias-variance tradeoff aims to select a model that strikes the right balance between bias and variance to minimize the generalization error and achieve a model that can effectively generalize to unseen data. A model with high bias may underfit the data, while a model with high variance may overfit the data. The ideal model should have an appropriate level of complexity that captures the underlying patterns in the data without being too simplistic or too sensitive to noise. Now we can represent the test error according to the model complexity as follows:

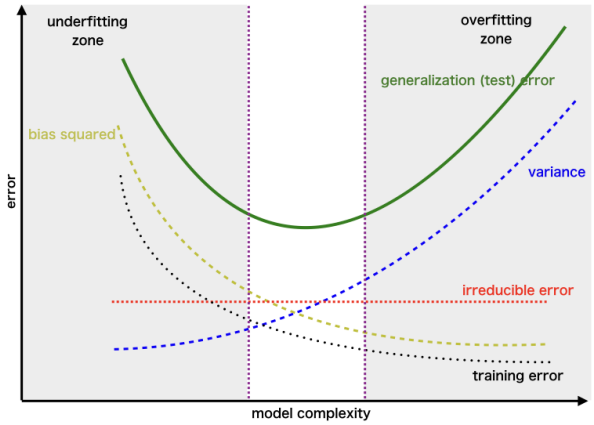

Figure 1: Bias-Variance Tradeoff

2 The double descent phenomenon

Figure 2 shows that if the model complexity is increased, the test error first decreases and afterwards increases to a peak point. In many cases it is empirically observed that the test error can start to decrease again to a very minimum point confirming the intuition that bigger/complex models are better. This phenomenon is called **double descent phenomenon**.

This second decrease happens at the **interpolation threshold** whereby model complexity is just sufficient enough to fit the training data and thereby causing a very small training error approximately equal to zero. This splits the graph into overparameterized (underconstrained) regime whereby number of model parameters, P, is larger than number of training data, N, and underparameterized (overconstrained) regime whereby N is greater than P.

The second decrease happens in the overparameterized regime and is due to the different factors of the model complexity as stated earlier.

The phenomenon can be also illustrated by this animated image [Web2, 2023] https://mlu-explain.github.io/double-descent/

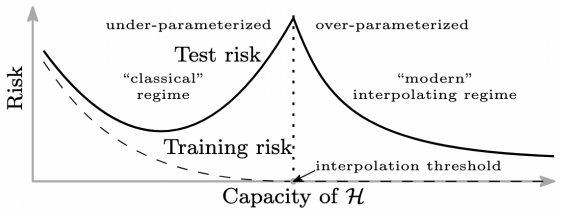

Figure 2: Double descent.

Model-wise double descent

In modern machine learning it is shown that the test error doesn't keep increasing as you increase model complexity, however, as you increase model complexity, the test error reaches a critical point(interpolation threshold) and starts to decrease again. Since this second decrease is caused by increase in model complexity, this type of double descent is called model-wise double descent.

Sample-wise double descent

As the number of training examples (N) increases, model becomes more complex in-order to cater for the large data set. In some cases, when the number of samples is increased to almost/equal to number of parameters P this increase in training examples has been noted to cause a type of double descent called sample-wise double descent.

Epoch-wise double descent

In complex models, for a given large number of optimization steps (number of epochs) of training data, the test error decreases, increases, and decreases again due to what is called epoch-wise double descent. This double descent is due to the number of optimization steps.

Key point to note

It is to be noted that, model-wise and sample-wise are very similar; model-wise focuses on increasing number of parameters P until it reaches or exceeds number of samples (N) whereas sample-wise focuses on increasing number of samples such that they reach number of parameters P.

3 Double descent phenomenon and model generalization

The double descent phenomenon occurs at the interpolation threshold, where a model perfectly fits the training data but becomes sensitive to noise, resulting in a peak in test error. In the overparameterized regime, there are multiple models that can absorb noise and fit the training data well, and SGD as an optimizer can find the best model. Despite lacking explicit regularization, overparameterized models still exhibit good generalization performance on the testing set due to implicit regularization through SGD.

4 Double descent phenomenon with models

The double descent phenomenon refers to a peculiar behavior observed in certain models where the test error initially decreases as the model complexity increases, then reaches a minimum, and finally increases again as the model becomes more complex. This phenomenon challenges the traditional bias-variance tradeoff and suggests that adding more complexity to a model may not always lead to overfitting.

Currently, the double descent phenomenon has been observed in models such as:

- Linear regression

- Linear Discriminant Analysis
- Logistic regression

On the other hand, there are still ongoing research and investigation to determine whether or not the following models exhibit the double descent phenomenon:

- Quadratic Discriminant Analysis (work on Linear Discriminant Analysis may be extended)
- Random forests
- Support Vector Machines
- Neural networks with nonlinear activation

The double descent phenomenon has sparked significant interest in the machine learning community as it challenges our traditional understanding of model complexity and generalization error. Further research is being conducted to better understand the underlying causes and implications of this phenomenon in different models.

5 Python Code

The code aims to investigate whether the double descent phenomenon occurs for a synthetic dataset and how the regularization parameter affects the shape of the test error curve. Linear regression is used as a simple model to test for double descent, and a function for the L2 regularizer is defined to further analyze the phenomenon.

5.1 Define Function

Regularized linear regression solution:

$$a = (\mathrm{X}^T\mathrm{X} + \lambda I)X^T\mathrm{y}$$

To investigate the relationship between the model complexity (controlled by the number of samples N and the regularization parameter λ and the generalization error (measured by the test error) on a fixed size of training data. By varying the values of N and λ and observing the resulting test errors

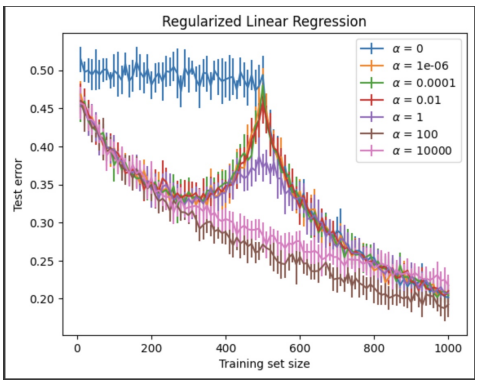

Figure 3: Linear Regression with L2 Regularizer

When λ is small, the regularization is weak and the model tends to fit the training data more closely, possibly leading to overfitting. When λ is large, the regularization is strong and the model tends to have smaller coefficients, which can help prevent overfitting. Double descent phenomenon is observed for smaller values of λ implying that using large regularization parameters might be able to stop the double descent phenomenon.

References

[Belkin et al., 2019] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

[Nakkiran et al., 2019] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt.

[Schaeffer et al., 2023] Schaeffer, R., Khona, M., Robertson, Z., Boopathy, A., Pistunova, K., Rocks, J. W., Fiete, I. R., and Koyejo, O. (2023). Double descent demystified: Identifying, interpreting ablating the sources of a deep learning puzzle.

[Web1, 2023] Web1 (Accessed April 2023). The double descent phenomenon in machine learning. Webots,https://math.gatech.edu/sites/default/files/images/reu2021_liao.pdf.

[Web2, 2023] Web2 (Accessed April 2023). Double descent animated image. Webots,https://mlu-explain.github.io/double-descent/.