VIETNAM GENERAL CONFEDERATION OF LABOR
TON DUC THANG UNIVERSITY
INFORMATION TECHNOLOGY FACULTY

HOMEWORK 1

MACHINE LEARNING

Overfitting and Prevention Techniques

Instructor: **Prof. Lê Anh Cường**
Name: **Đỗ Phạm Quang Hưng**
Student ID: **520K0127**

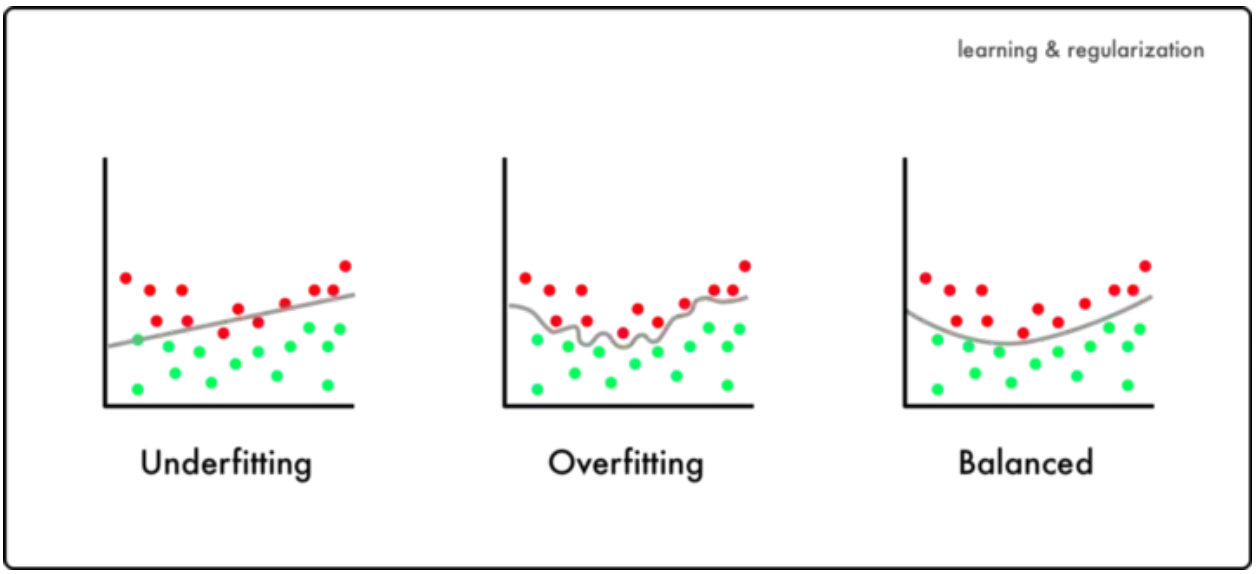
HO CHI MINH CITY, 2023

Table of content

What is overfitting?

It is a common pitfall in deep learning algorithms in which a model tries to fit the training data entirely and ends up memorizing the data patterns and the noise and random fluctuations. These models fail to generalize and perform well in the case of unseen data scenarios, defeating the model's purpose.

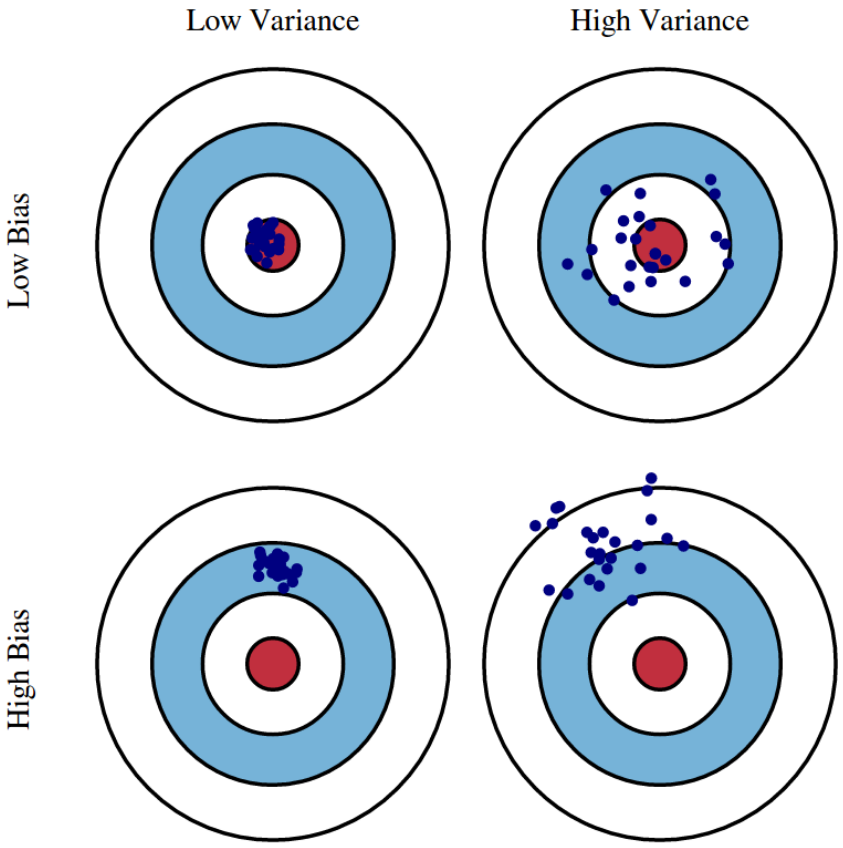
When machine learning algorithms are constructed, they leverage a sample dataset to train the model. However, when the model trains for too long on sample data or when the model is too complex, it can start to learn the “noise,” or irrelevant information, within the dataset. When the model memorizes the noise and fits too closely to the training set, the model becomes “overfitted,” and it is unable to generalize well to new data



Overfitting: Key definitions.

- **Bias:** Bias measures the difference between the model’s prediction and the target value. If the model is oversimplified, then the predicted value would be far from the ground truth resulting in more bias.
- **Variance:** Variance is the measure of the inconsistency of different predictions over varied datasets. If the model’s performance is tested on different datasets, the closer the prediction, the lesser the variance. Higher variance is an indication of overfitting in which the model loses the ability to generalize.
- **Bias-variance tradeoff:** A simple linear model is expected to have a high bias and low variance due to less complexity of the model and fewer trainable parameters. On the other hand, complex non-linear models tend to observe an

opposite behavior. In an ideal scenario, the model would have an optimal balance of bias and variance.



Graphical illustration of bias and variance.

- **Model generalization:** Model generalization means how well the model is trained to extract useful data patterns and classify unseen data samples.
- **Feature selection:** It involves selecting a subset of features from all the extracted features that contribute most towards the model performance. Including all the features unnecessarily increases the model complexity and redundant features can significantly increase the training time.

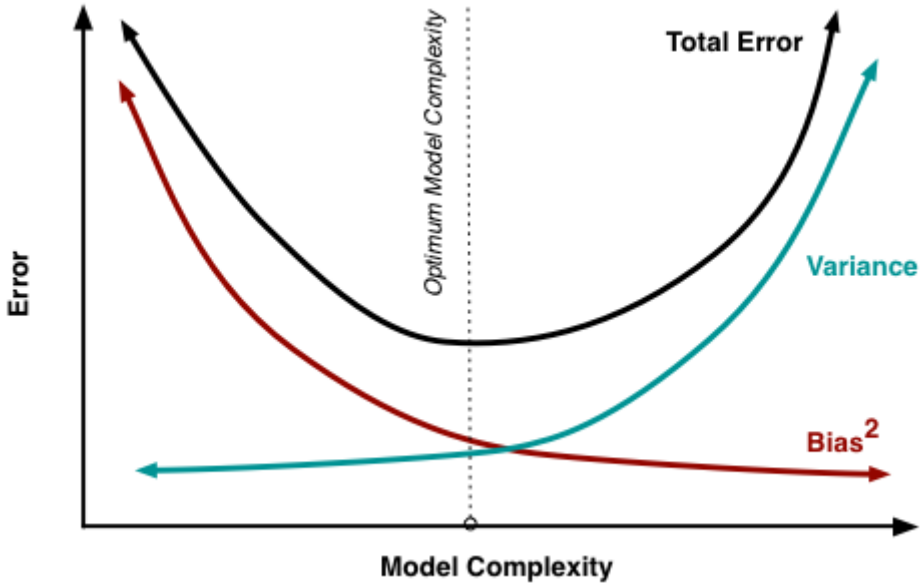


Fig 2: The variation of Bias and Variance with the model complexity. This is similar to the concept of overfitting and underfitting. More complex models overfit while the simplest models underfit.

How to detect overfit models?

Low error rates and a high variance are good indicators of overfitting. In order to prevent this type of behavior, part of the training dataset is typically set aside as the "test set" to check for overfitting. If the training data has a low error rate and the test data has a high error rate, it signals overfitting.

Loss and Accuracy of train and validation set tend to diverge is a sign that your model is overfitting

Overfitting Prevention Techniques

1. Validation

1.1. Validation

We are still used to dividing the data set into two small sets: training data and test data. And one thing I still want to reiterate is that when building the model, we must not use test data. So how to know the quality of the model with unseen data.

The simplest method is to extract a small subset from the training data set and perform model evaluation on this small subset. The small subset extracted from this training set is called the validation set. At this point, the training set is the remainder of the original training set. Train error is calculated on this new training set, and there is another concept that is defined similarly to validation error, i.e. error is calculated on the validation set.

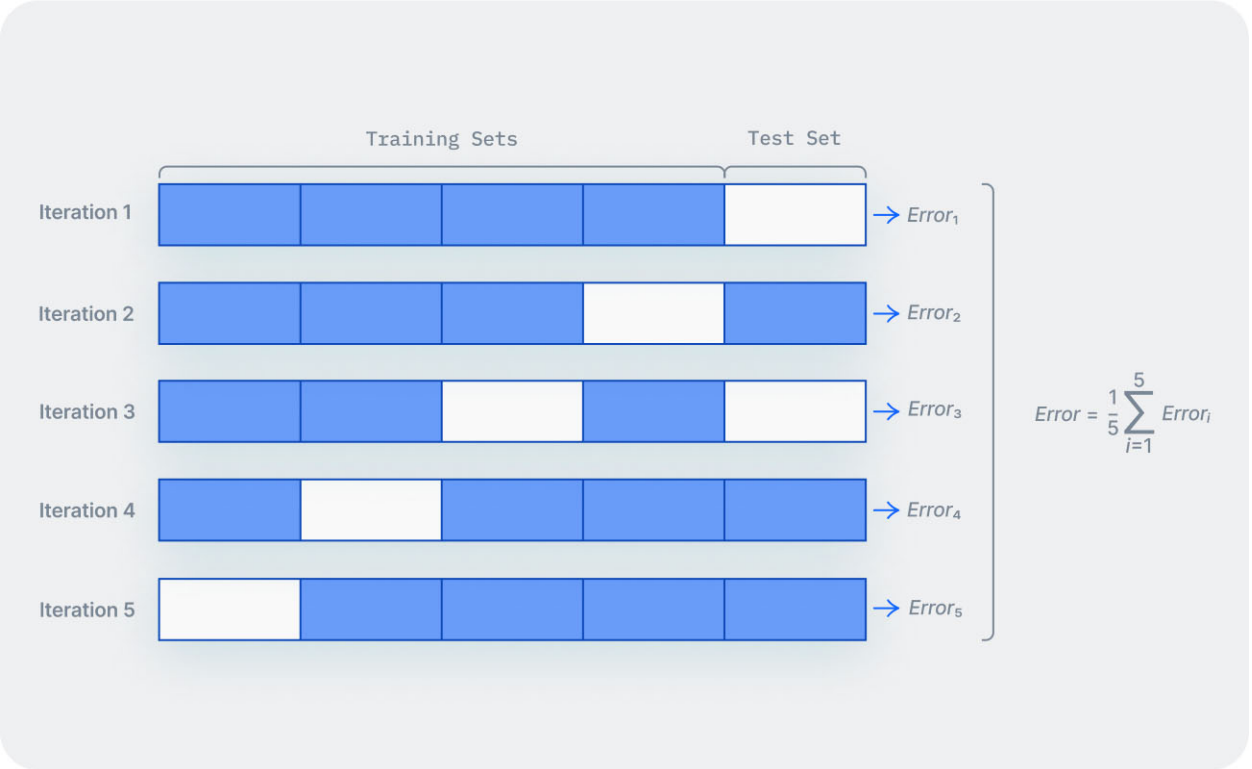
1.2. Cross-validation

In many cases, we have a very limited amount of data to build a model. If too much data in the training set is taken out as validation data, the remaining data of the training set is not enough to build the model. At this point, the validation set must be very small to keep the training data large enough. However, another problem arose. When the validation set is too small, overfitting can occur with the remaining training set. Is there any solution for this situation?

The answer is **cross-validation**.

Cross validation is an improvement of validation with a small amount of data in the validation set, but the model quality is evaluated on many different validation sets. A common way to use it is to divide the training set into ***k*** subsets with no common elements, of roughly equal size. At each test run, called run, one of the k subsets is taken as the validate set. The model will be built based on the union of the remaining **k-1** subsets. The final model is

determined based on the average of the train errors and validation errors. This approach is also known as k-fold cross validation.



K-fold cross-validation

2. Early stopping

This method aims to pause the model's training before memorizing noise and random fluctuations from the data. There can be a risk that the model stops training too soon, leading to underfitting. One has to come to an optimum time/iterations the model should train.

3. Train with more data

With the increase in the training data, the crucial features to be extracted become prominent. The model can recognize the relationship between the input attributes and the output variable. The only assumption in this method is that the data to be fed into the model should be clean; otherwise, it would worsen the problem of overfitting.

4. Addition of noise to the input data

Another similar option as data augmentation is adding noise to the input and output data. Adding noise to the input makes the model stable without affecting data quality and privacy while adding noise to the output makes the

data more diverse. Noise addition should be done in limit so that it does not make the data incorrect or too different.

5. Data augmentation

An alternative method to training with more data is data augmentation, which is less expensive and safer than the previous method. Data augmentation makes a sample data look slightly different every time the model processes it.

6. Feature selection

Every model has several parameters or features depending upon the number of layers, number of neurons, etc. The model can detect many redundant features or features determinable from other features leading to unnecessary complexity. We very well know that the more complex the model, the higher the chances of the model to overfit.

7. Regularization

If overfitting occurs when a model is too complex, reducing the number of features makes sense. Regularization methods like Lasso, L1 can be beneficial if we do not know which features to remove from our model. Regularization applies a "penalty" to the input parameters with the larger coefficients, which subsequently limits the model's variance.

8. Ensemble methods

It is a machine learning technique that combines several base models to produce one optimal predictive model. In Ensemble learning, the predictions are aggregated to identify the most popular result. Well-known ensemble methods include bagging and boosting, which prevents overfitting as an ensemble model is made from the aggregation of multiple models.

9. Other methods

In addition to the methods mentioned above, for each model, many other overfitting avoidance methods are also used. Typically, **Dropout** in Deep Neural Networks was recently proposed. In a nutshell, dropout is a method of randomly shutting down units in Networks. Turn off instant for zero value units and calculate normal feedforward and backpropagation during training. This not only reduces the amount of computation, but also prevents overfitting.

Pruning (avoiding overfitting in Decision Trees), **VC dimension** (measuring the complexity of the model, the greater the complexity, the more prone to overfitting)

References

1. What is Overfitting? | IBM
2. Overfitting | Machine Learning Cơ Bản
3. What is Overfitting in Deep Learning [+10 Ways to Avoid It]
4. 8 Simple Techniques to Prevent Overfitting | by David Chuan-En Lin | Towards Data Science
5. Lecture 12: Bias Variance Tradeoff | Cornell University