Bias-Variance Tradeoff
and
Cross Validation

Maturity is the capacity to endure uncertainty.
- John Finley

Outline

Learning Objectives

In this video, you will learn to:

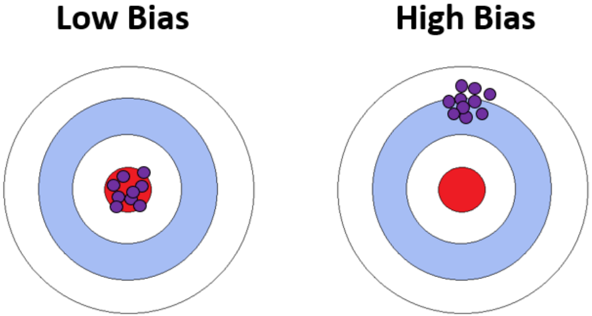
- Understand the three types of errors of prediction models, including bias and variance.
- Understand the relationship between bias, variance and the model complexity. In particular, understand the idea of bias-variance tradeoff.
- Understand the main idea of K-fold Cross Validation, and learn how it helps to overcome the Overfitting problem.

Introduction to Bias and Variance

Three Types of Prediction Errors

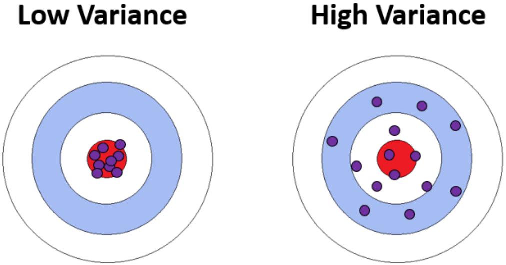
- ***Bias***: Error due to Bias.
- ***Variance***: Error due to Variance.
- ***Random error***: Error due to unavoidable randomness.

Bias



- Bias is the difference between the average prediction of our model and the true target value.
- In the diagram on the left, it has low bias.
- In the diagram on the right, it has high bias.
- A model with high bias may not do well in capturing the key information in the training dataset.
- The model may be oversimplified, namely, Underfitted, or make some overly simplistic assumptions.

Variance

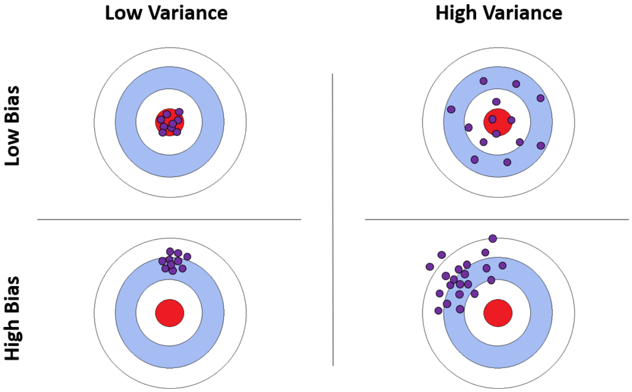


- Variance can be measured by the variability of a model prediction for a given data point.
- Imagine we could repeat the model building process multiple times. The variance measures how much variability these predictions have.
- In the diagram to the left, it has low variance.
- In the diagram to the right, it has high variance.
- A model with high variance is very sensitive to the input data.
- The model with high variance may have been overfitted to the training dataset, and hence, learn the "noise" of the training dataset.

Random Error

- Random error always exists, regardless of the models being applied.
- Our job is to find the optimal model that captures the actual relationship, while avoiding incorporating the random noise.

Bias vs. Variance

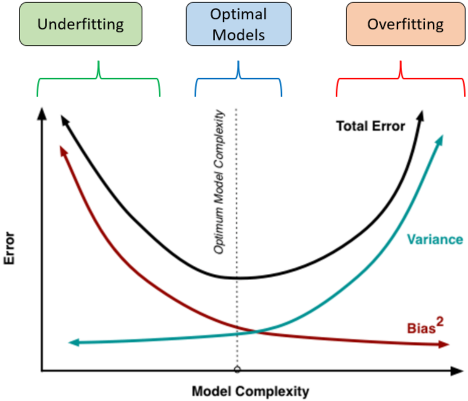


- Top left: ***Optimal models***, with low bias and low variance.
- Bottom left: ***Unfitted models***, with high bias and low variance.
- Top right: ***Overfitted models***, with low bias and high variance.
- Bottom right: Worst models, with high bias and high variance.

Bias-Variance Tradeoff

Bias-Variance Tradeoff

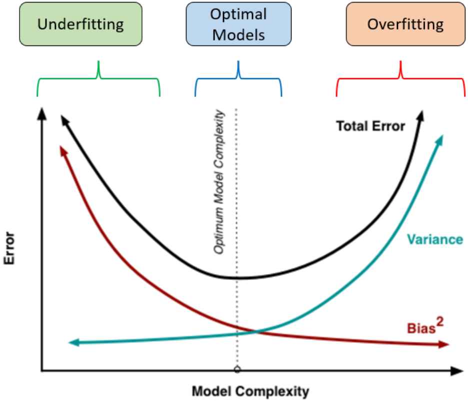
- The figure summarises the relationship between the model complexity, bias and variance.
- Underfitting: the model is too simple, and it is likely to have high bias and low variance.
- Overfitting: the model is too complex, and it is likely to have low bias and high variance.
- The overfitted models do not generalise well to any unseen data. In practice, they are likely to have a low error rate on the training dataset, but a high error rate on the test dataset.
- Only when the model has an appropriate complexity, it achieves low bias and low variance.



Source: `http://scott.fortmann-roe.com/docs/BiasVariance.html`

Bias-Variance Tradeoff

- In general, there is an inverse relationship between bias and variance.
- For example, as the model complexity increases, the bias decreases, and the variance increases.
- Our job is to find the optimal model that makes a good balance between bias and variance.



Source: `http://scott.fortmann-roe.com/docs/BiasVariance.html`

Regularisation: Adjust the Model Complexity

Regularisation: Adjust the Model Complexity

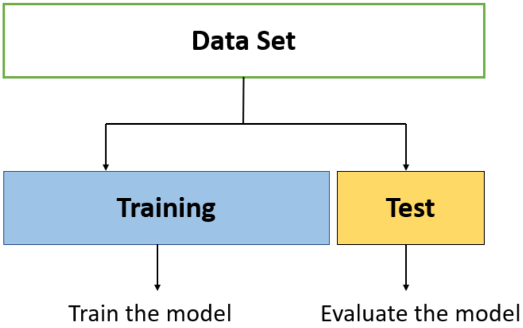
- Regularisation can adjust the model complexity, with an adjustable regularisation parameter.
- The parameter controls the degree of regularisation, and in turn, adjusts the coefficients of the model.
- A larger regularisation parameter will produce a model with smaller coefficients.
- By choosing a proper regularisation parameter, the regularised model will be able to have a decrease in variance, together with some affordable increase in bias.
- Regularisation solves the Overfitting problem.

Cross Validation

Cross Validation

- ***Cross Validation*** is a statistical method used to estimate the performance of some machine learning models.
- Cross Validation can help us to determine the optimal regularisation parameter.
- In the subsequent discussions, Cross Validation refers to K-fold Cross Validation.

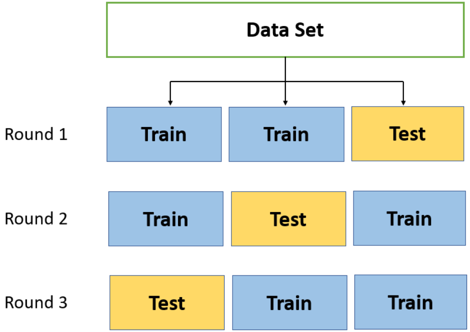
Train-Test Split



- The method of train-test split is also called the “holdout method”.
- It can be used for model selection.
- You can repeat the process for all the models, and record the test performance of the models.
- The optimal model can be selected as the one that has the least test error rate.

K-fold Cross Validation

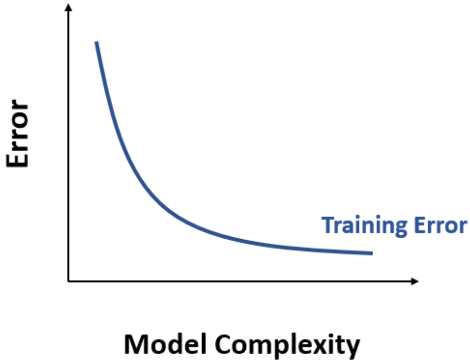
- For K-fold Cross Validation, the dataset is split into K number of equal sized folds.
- In round 1, the first two folds are the training dataset, and the 3rd fold is the test dataset.
- In round 2, the 1st and the 3rd folds are the training dataset, and the 2nd fold is the test dataset.
- In round 3, the 2nd and the 3rd folds are the training dataset, and the 1st fold is the test dataset.
- In the end, we will take the average of all rounds' error rates, and refer to it as the (mean) ***Cross Validation error rate***.
- The optimal model, or the optimal parameter, can be chosen as the one that results in the lowest Cross Validation error rate.



Model Complexity and Error

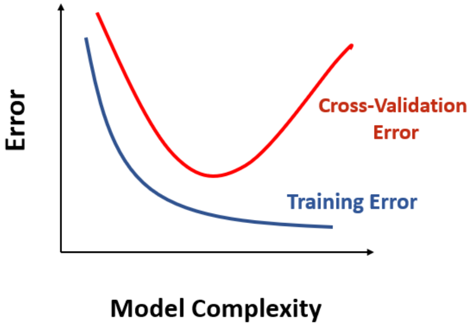
Model Complexity and Error

- As the model increases the complexity, the error rate of the model on the training dataset, denoted as ***training error***, continuously decreases.



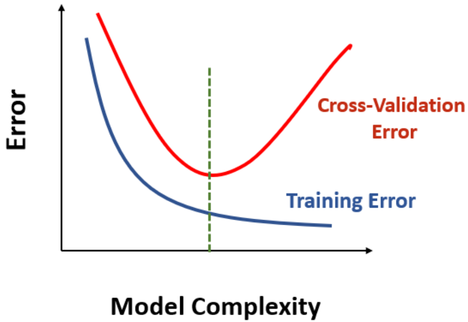
Model Complexity and Error

- As the model increases the complexity, the error rate of the model on the training dataset, denoted as ***training error***, continuously decreases.
- When the model is too simplistic, it has a high Cross Validation (***CV***) error rate.
- Before the inflection point, the CV error decreases, when the model complexity increases.
- After the inflection point, the CV error increases, when the model complexity increases.



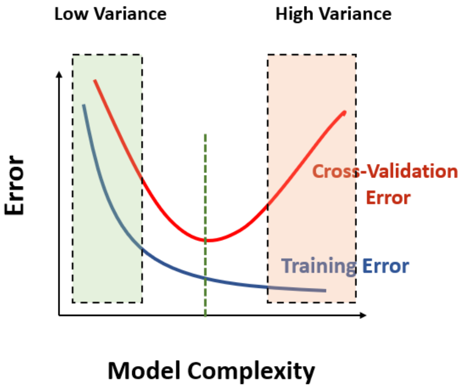
Model Complexity and Error

- As the model increases the complexity, the error rate of the model on the training dataset, denoted as ***training error***, continuously decreases.
- When the model is too simplistic, it has a high Cross Validation (***CV***) error rate.
- Before the inflection point, the CV error decreases, when the model complexity increases.
- After the inflection point, the CV error increases, when the model complexity increases.
- The plot of the Cross Validation error has a U shape.



Model Complexity and Error

- In general, when the model complexity is low, the model has low variance.
- When the model complexity is too high, the model has high variance, and it may not generalise well to any unseen data.
- The most appropriate model complexity is the one that achieves the lowest Cross Validation error.
- Such model makes a good balance between bias and variance.
- The Cross Validation method can help us to select the optimum models and the optimum parameters.



Summary

Summary

We have learnt to:

- Understand the relationship between bias, variance and the model complexity.
- Understand the bias-variance tradeoff.
- Understand how the K-fold Cross Validation works, and understand how it helps to select the optimal models and the optimal parameters.

In the next video,

We will introduce the theoretical model of Ridge Regression, and use a case study to assist you to build a Ridge Regression model in R.

References

- IBM, *Supervised Machine Learning: Regression*
  https://www.coursera.org/learn/supervised-machine-learning-regression/lecture/IlgJd/bias-variance-trade-off
  https://www.coursera.org/learn/supervised-machine-learning-regression/lecture/UYYeJ/cross-validation
- Scott Fortmann-Roe (2012) *Understanding the Bias-Variance Tradeoff*
  http://scott.fortmann-roe.com/docs/BiasVariance.html
- Seema Singh (2018) *Understanding the Bias-Variance Tradeoff*
  https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
- Great Learning Team (2020) *What is Cross Validation in Machine learning? Types of Cross Validation*
  https://www.mygreatlearning.com/blog/cross-validation/