Lecture 05: Bias-Variance Trade-Off

EE7209: Machine Learning (TE)

Mr. M.W.G Charuka Kavinda Moremada

Lecturer, Department of Electrical and Information Engineering, Faculty of Engineering, University of Ruhuna.

Lecture Overview

1. Model Generalization
   i. Info Slide: Data Snooping
   ii. Model Overfitting and Underfitting - Revisit
2. Bais-Variance Tradeoff
   i. Linear Fit
   ii. Fit With 5th Order Polynomial
   iii. Bulls-Eye Diagram
   iv. Summary

A text generated image from Bing image creator powered by DALLE-3 which illustrates the power of modern AI tools.

ML Model Generalization

Model Generalization

- What should be the ultimate goal in machine learning model training?

Minimizing the training data loss/error	or	Minimizing the testing data/unseen loss/error

- Minimizing the training loss is our approach towards the goal of learning a predictive model; since we cannot perform model optimization directly using test data (Note: When implementing the ML models, we have to keep the usage of test set at the minimum and avoid leakage of information from test set to train set).
- The most important evaluation metric of a model is the loss on unseen test examples, which is oftentimes referred to as the test error.

Info Slide

Data Snooping

- **Data snooping** refers to statistical inference that the researcher decides to perform after looking at the data (as contrasted with pre-planned inference, which the researcher plans before looking at the data) [3].
- In simple terms, data Snooping is a statistical mistake that happens when someone decides to look at the data before having his statistical guidelines or building hypothesis. He/She is being influenced by what he is seeing and the hypothesis that he will start working on might already be too much data driven.
- **When you decide to take the entire the dataset to develop your ML model and ignore the split step of it this can ultimately leads to data snooping.**
- This can cause under different phases throughout the ML model development process.

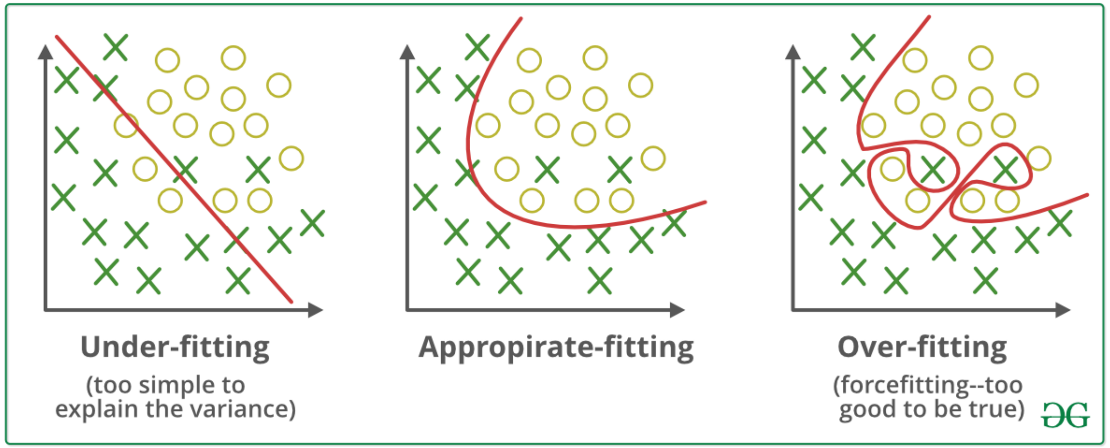
Model Generalization

- The key difference between the training and testing datasets is **the testing data are unseen in the sense that these data have not been utilized under the training procedures (if utilized, it is fundamentally inaccurate).**
- However, under the classical statistical learning settings, the training and testing examples are drawn from the **same distribution**. However, with respect to the training procedure, **the test set is still unseen to the model.**
- Even though both training and testing data have been drawn from the same distribution, **the test error is not always close to the training error** due to the fact that test data are unseen while training data are seen to the model during the training procedure.
- As a result, **successfully minimizing the training error may not always lead to a small test error.**

Model Generalization: Model Overfitting and Underfitting - Revisit

- **Overfitting:** If a model predicts accurately on the training dataset but doesn't generalize well to other test examples, that is, if the training error is small but the test error is large.
- **Underfitting:** A model underfits the data if the training error is relatively large; and in this case, typically the test error is also relatively large.

Underfitting, Appropriate Fit and Overfitting for a Classification Problem



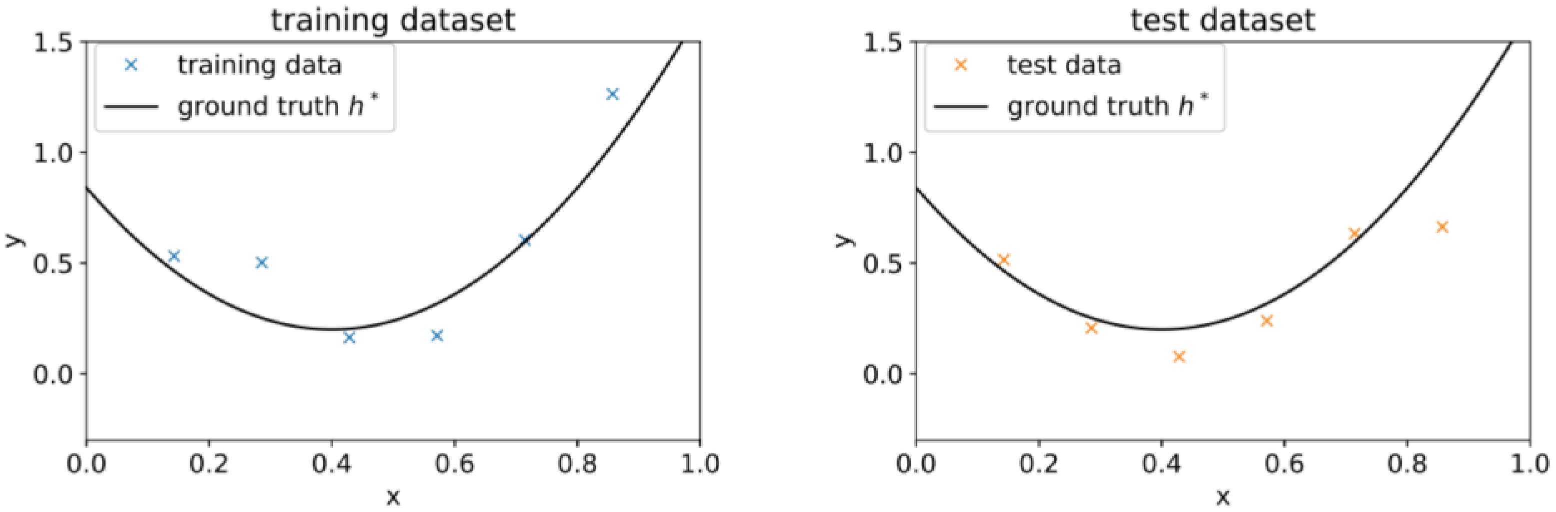
A text generated image from Bing image creator powered by DALLE-3 which illustrates the power of modern AI tools.

Bias-variance Tradeoff

Bais-Variance Tradeoff

An example taken from, A. Ng and T. Ma, CS229 Lecture Notes. 2023.

Ground Truth Function



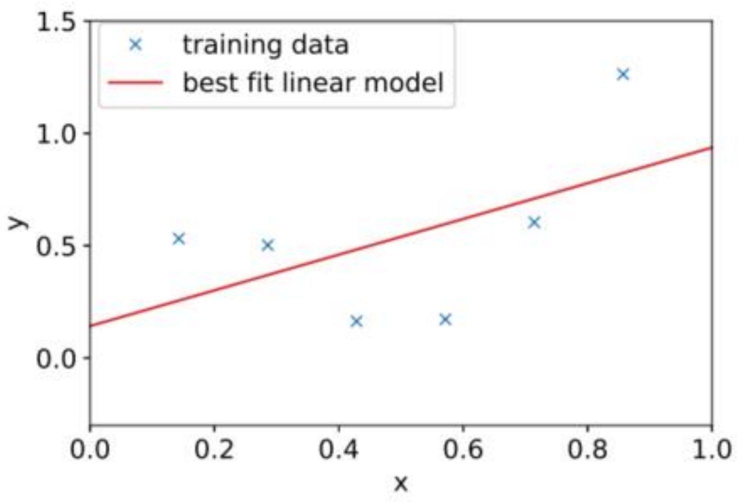
$$y^{(i)} = h^*(x^{(i)}) + \varepsilon^{(i)}$$
$$; where\ \varepsilon \sim N(0, \sigma^2)$$

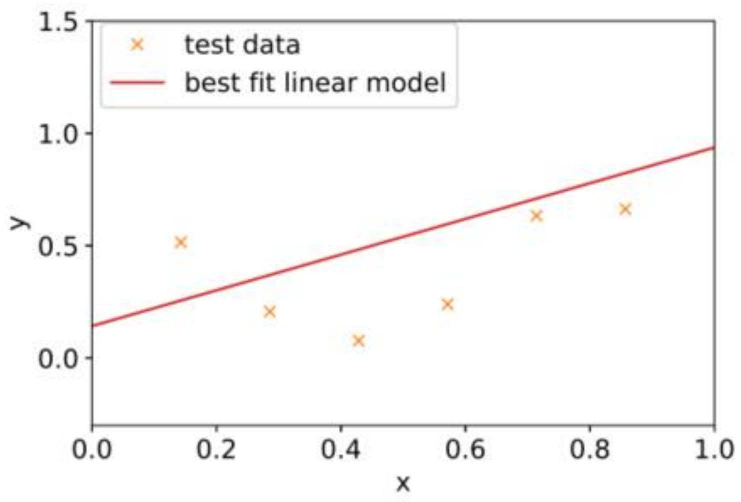
- Similar to the training data the testing data having the same input-output relationship.

Since it is impossible to predict the observation noise ε we are trying to recover the quadratic function $h^*(.)$; the solid line.

Bias-Variance Tradeoff: Linear Fit

Best-Fit Linear Model



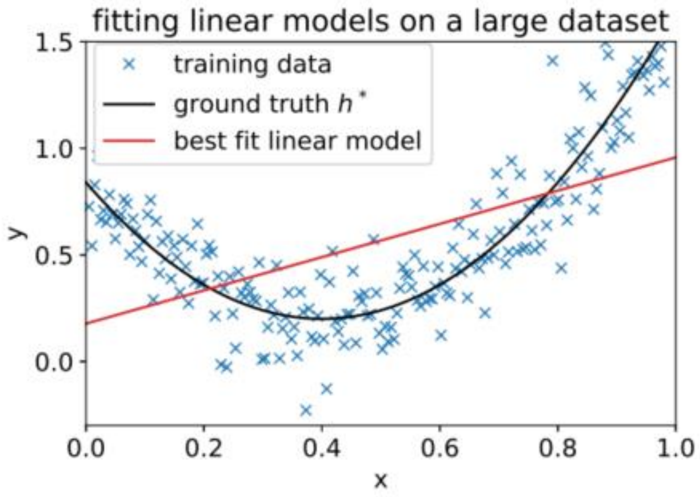


The best fitted linear model cannot predict y from x accurately even on the training dataset, let alone on the test dataset. This is because the true relationship between y and x is not linear - any linear model is far away from the true function h^*.

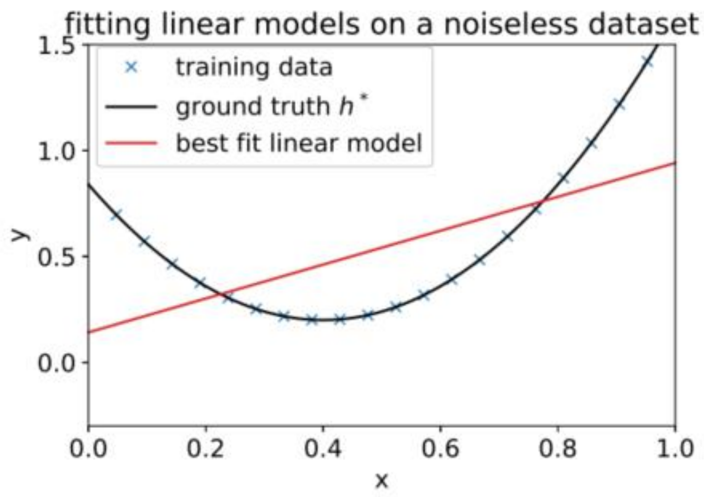
Large Training and Test Errors

Underfitting

Bias-Variance Tradeoff: Linear Fit - With More Data



The best fit linear model on a much larger dataset still has a large training error.



The best fit linear model on a noiseless dataset also has a large training/test error.

Even with more/infinite training example the issue persisted and even with data that do not present noise the same issue can be identified.

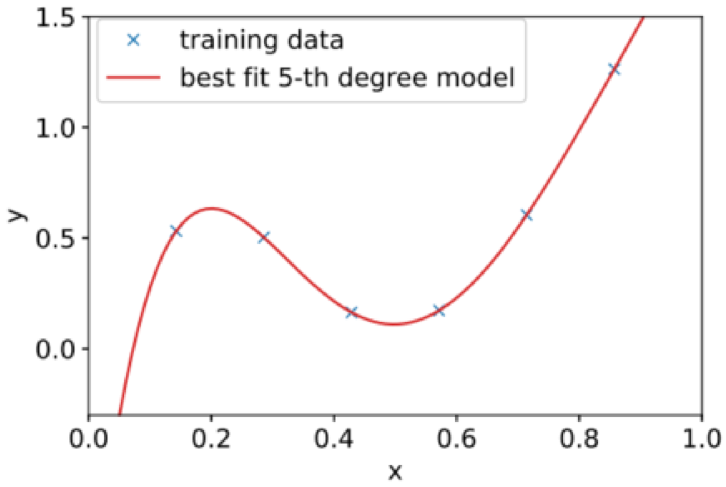
Main Bottleneck: The linear model family's inability to capture the structure in the data; linear models cannot represent the true quadratic function

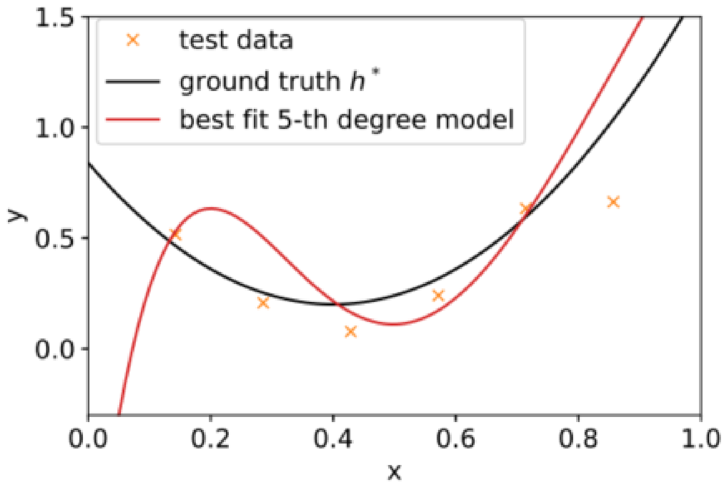
Bias-Variance Tradeoff: Linear Fit

- We define the bias of a model (informally) to be the test error even if we were to fit it to a very (say, infinitely) large training dataset.
- **Thus, in this case, the linear model suffers from large bias, and underfits (i.e., fails to capture structure exhibited by) the data.**

Bias-Variance Tradeoff: Fit With 5th Order Polynomial

Best Fit 5th Order Polynomial



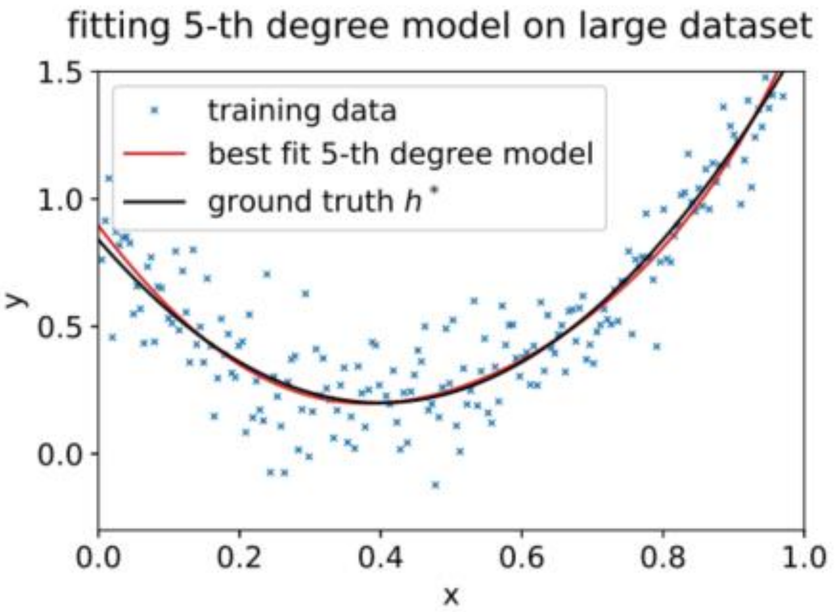


Best fit 5th degree polynomial has zero training error, but still has a large test error and does not recover the ground truth. This is a classic situation of overfitting.

The model learnt from the training set does not generalize well to other test Examples; the test error is high.

Overfitting

Bias-Variance Tradeoff: Fit With 5^{th} Order Polynomial-With More Data

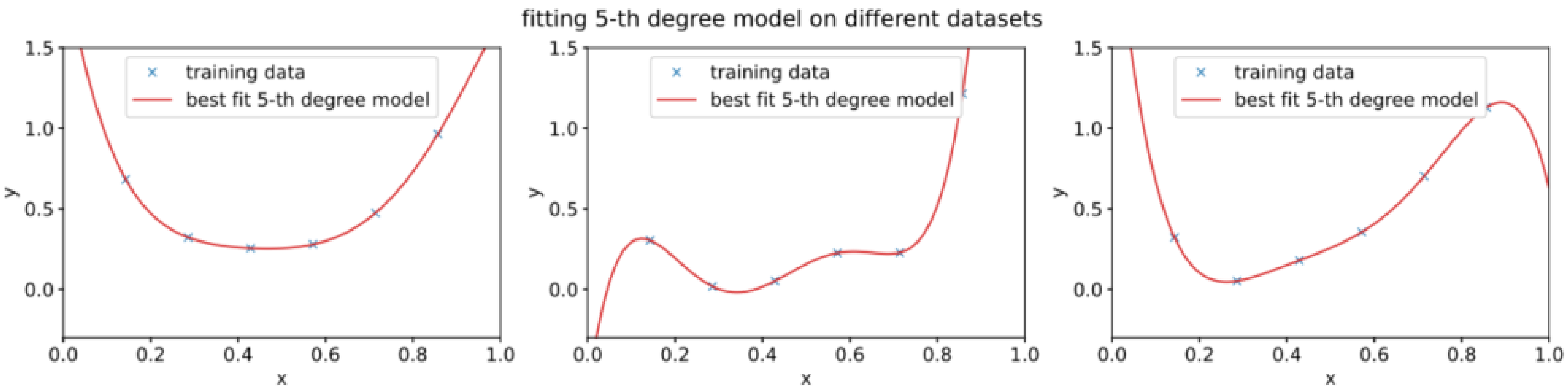


- If we were to fit a 5-th degree polynomial to an extremely large dataset, the resulting model would be close to a quadratic function and be accurate.
- This is because the family of 5-th degree polynomials contains all the quadratic functions (setting $\theta_5 = \theta_4 = \theta_3 = 0$ results in a quadratic function), and, therefore, 5^{th} degree polynomials are in principle capable of capturing the structure of the data.

The best fit 5-th degree polynomial on a huge dataset nearly recovers the ground-truth

In this case, the test error indicated in the previous slide occurred due to variance not bias

Bias-Variance Tradeoff: Fit With 5th Order Polynomial



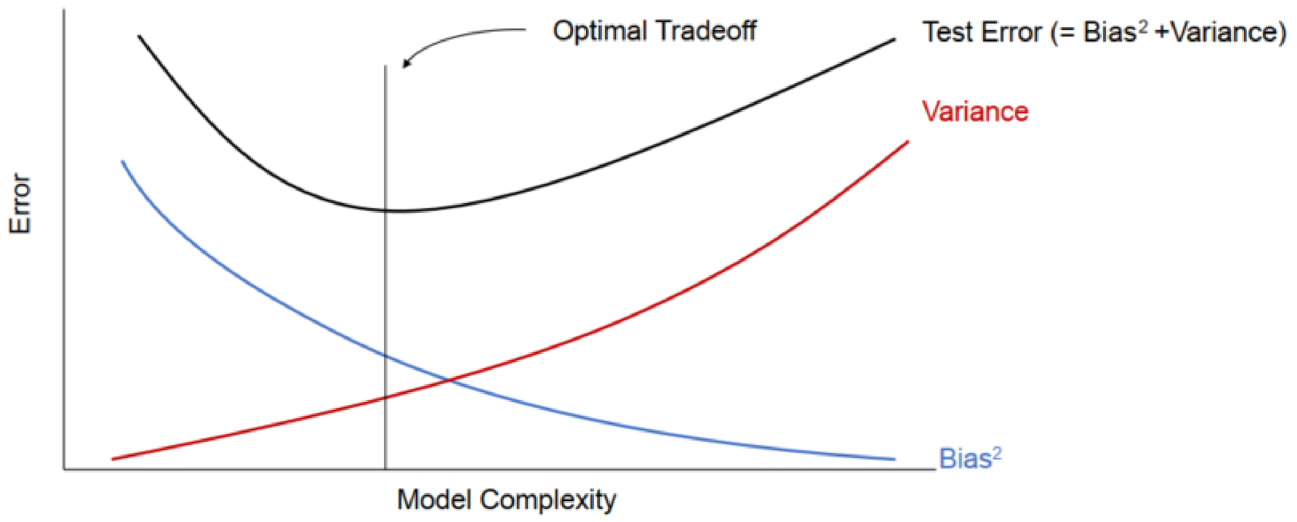
The best fit 5-th degree models on three different datasets generated from the same distribution behave quite differently, suggesting the existence of a large variance.

Bias-Variance Tradeoff: Fit With 5th Order Polynomial

- Under the previous cases, we are we're fitting patterns in the data that happened to be present in our small, finite training set, but that do not reflect the wider pattern of the relationship between x and y.
- The **"spurious" patterns** in the training set are (mostly) due to the **observation noise** $\varepsilon^{(i)}$.
- By fitting these spurious patters results in a model with large test error. In this case, we say the model has a **large variance**.
- The "spurious patterns" **are specific to the randomness of the noise (and inputs) in a particular dataset**, and thus are different across multiple training datasets.

Bias-Variance Tradeoff

Illustration of Bias-Variance Tradeoff

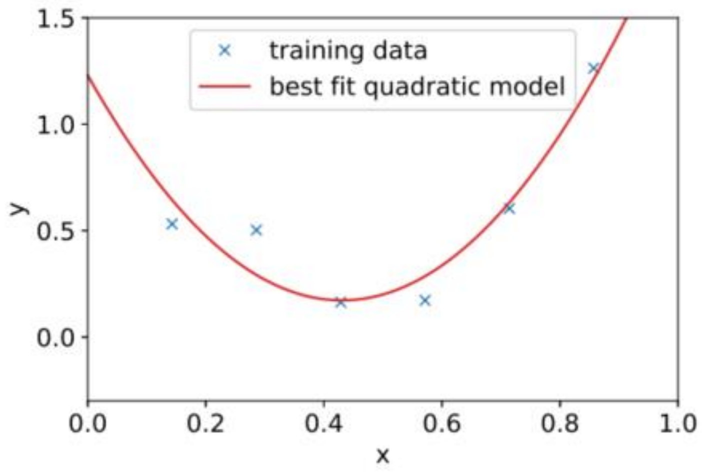


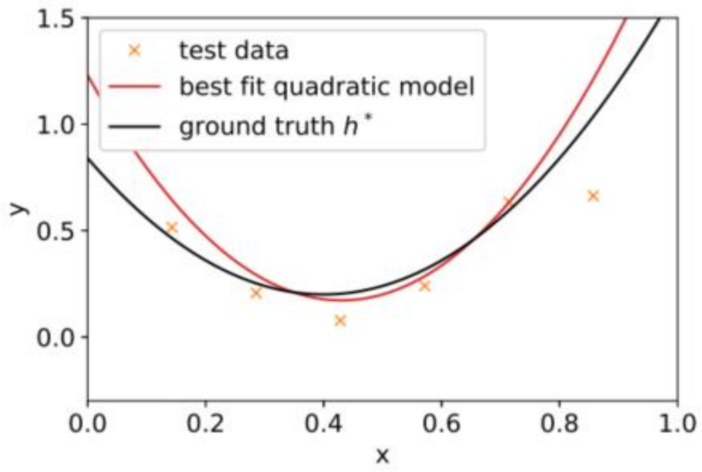
Note: The test error can be decomposed as a summation of bias and variance.

If our model is too "simple" and has very few parameters, then it may have large bias (but small variance), and it typically may suffer from underfitting.

If it is too "complex" and has many parameters, then it may suffer from large variance (but have smaller bias), and thus overfitting.

Bias-Variance Tradeoff



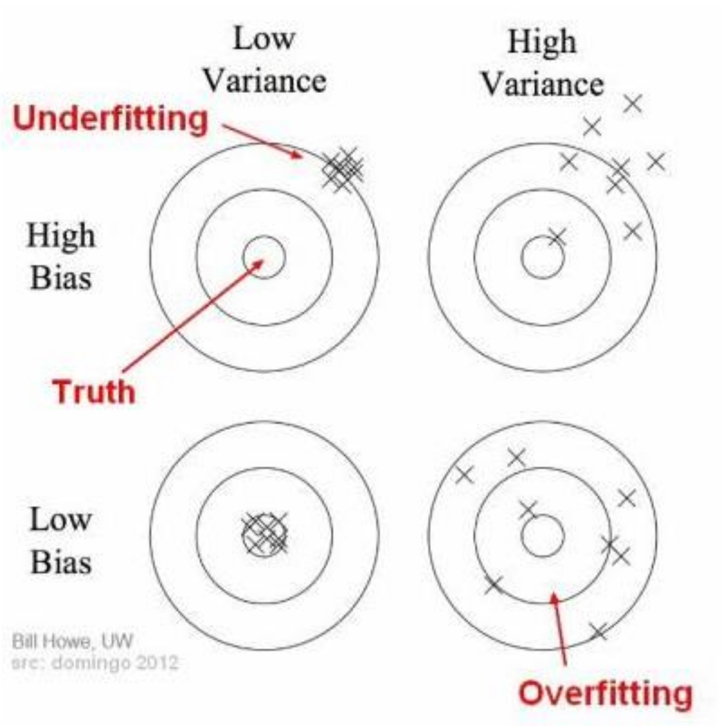


In practice we should tune the model complexity to achieve the best tradeoff.

Best fit quadratic model has small training and test error because quadratic model achieves a better tradeoff.

Bias-Variance Tradeoff: Bulls-Eye Diagram

- In the above diagram, center of the target is a model that perfectly predicts correct values. As we move away from the bulls-eye our predictions become get worse and worse.
- In supervised learning, underfitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance.
- Overfitting happens when our model captures the noise along with the underlying pattern in data. These models have low bias and high variance.



Bias-Variance Tradeoff: Summary

- **Bias:**
  - Bias is the difference between the average prediction of our model and the expected value which we are trying to predict.
  - Model with high bias pays very little attention to the training data and oversimplifies the model.
  - High bias leads to underfitting.
- **Variance:**
  - Variance is the variability in the model prediction, meaning how much the predictions will change if a different training data set is used.
  - Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before.
  - High variance leads to overfitting.

References

1. A. Ng and T. Ma, CS229 Lecture Notes. 2023.
2. D. Nautiyal, “Underfitting and Overfitting in Machine Learning,” GeeksforGeeks, Nov. 23, 2017. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/
3. “Data Snooping,” web.ma.utexas.edu. https://web.ma.utexas.edu/users/mks/statmistakes/datasnooping.html
4. N. Pogeant, “Data Snooping in Data Science — A Not Enough Known Issue,” Geek Culture, May 10, 2022. https://medium.com/geekculture/data-snooping-in-data-science-a-not-enough-known-issue-b6936e3552b7#:~:text=Data%20snooping%20refers%20to%20statistical (accessed Feb. 11, 2024).
5. “Lecture 12: Bias Variance Tradeoff,” www.cs.cornell.edu. https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html
6. “Chapter 2 Assessing Performance; Bias + Variance Tradeoff Slides (pdf) Video (Panopto) 2.1 Linear Regression Model.” Accessed: Feb. 13, 2024. [Online]. Available: https://courses.cs.washington.edu/courses/cse416/22sp/lectures/2/lecture_2.pdf
7. S. Singh, “Understanding the Bias-Variance Tradeoff,” Medium, May 21, 2018. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229.

Thank You