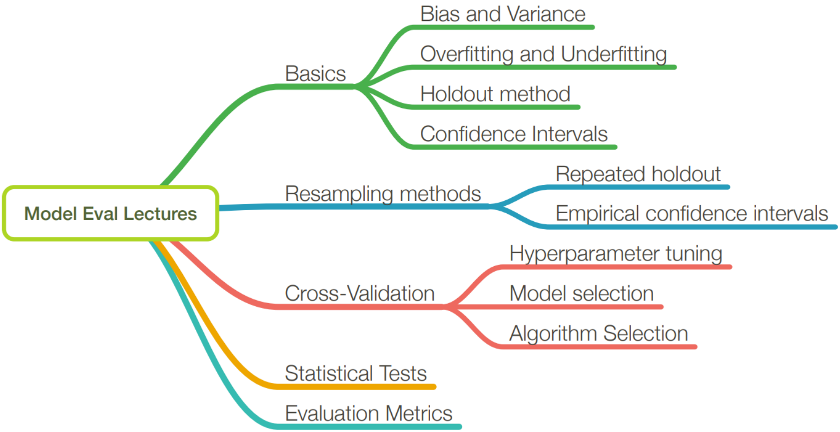
Generalization Error

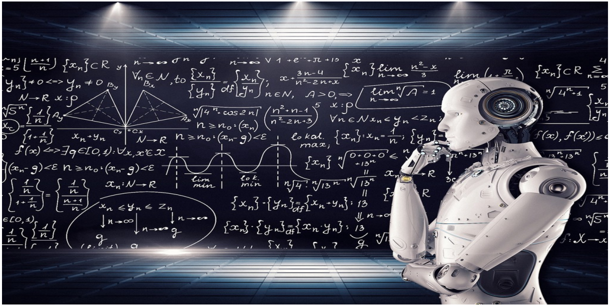
Ali Sharifi-Zarchi
Behrooz Azarkhalili
Peyman Naseri

CE Department
Sharif University of Technology

Fall 2022

Overview


Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Evaluation Metrics



When can we say that the machine has learned?

Generalization Performance

Generalization Performance

- When a model to ”generalize” well to unseen data (”high generalization accuracy” or ”low generalization error”)

Overfitting and Underfitting

Assumptions

- i.i.d. assumption: inputs are independent, and training and test examples are identically distributed (drawn from the same probability distribution)

Overfitting and Underfitting

Assumptions

- i.i.d. assumption: inputs are independent, and training and test examples are identically distributed (drawn from the same probability distribution)
- For some random model that has not been fitted to the training set, we expect both the training and test error to be equal

Overfitting and Underfitting

Assumptions

- i.i.d. assumption: inputs are independent, and training and test examples are identically distributed (drawn from the same probability distribution)
- For some random model that has not been fitted to the training set, we expect both the training and test error to be equal
- The training error or accuracy provides an (optimistically) biased estimate of the generalization performance

Overfitting and Underfitting

Model Capacity

- Underfitting: both training and test error are large

Overfitting and Underfitting

Model Capacity

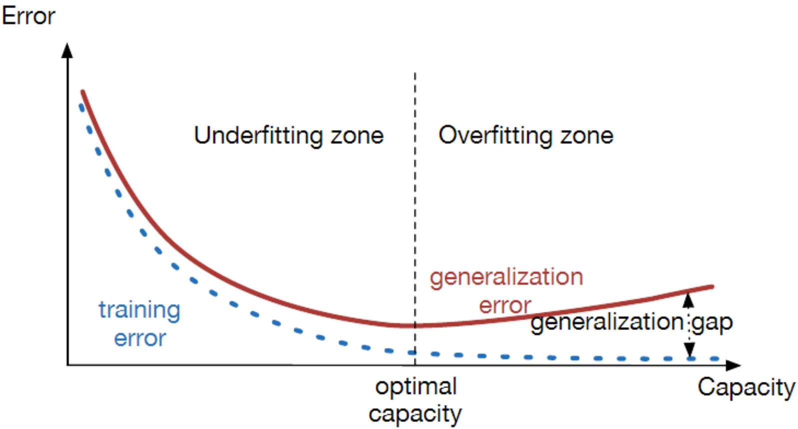
- Underfitting: both training and test error are large
- Overfitting: gap between training and test error (where test error is higher)

Overfitting and Underfitting

Model Capacity

- Underfitting: both training and test error are large
- Overfitting: gap between training and test error (where test error is higher)
- Large hypothesis space being searched by a learning algorithm
  - high tendency to overfit

Overfitting and Underfitting



Bias-Variance Trade-off

Bias-Variance Trade-off

Bias-Variance Decomposition

Bias-Variance Trade-off

Bias-Variance Decomposition

- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are correlated to underfitting and overfitting

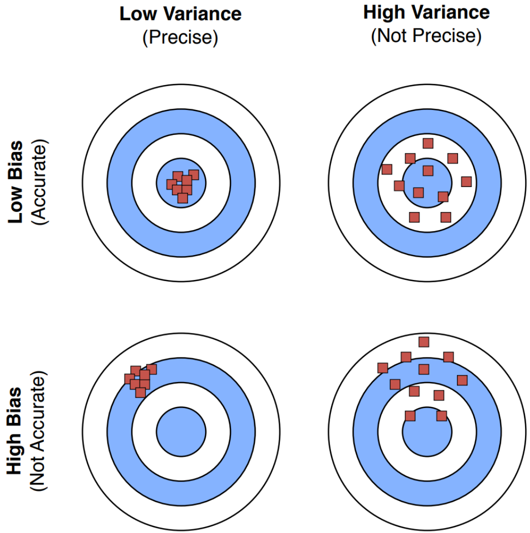
Bias-Variance Trade-off

Bias-Variance Decomposition

- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are correlated to underfitting and overfitting
- Helps explain why ensemble methods (last lecture) might perform better than single models

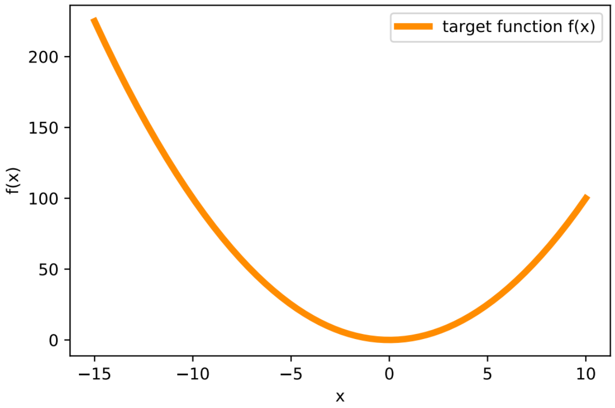
Bias-Variance Trade-off

Bias-Variance Intuition



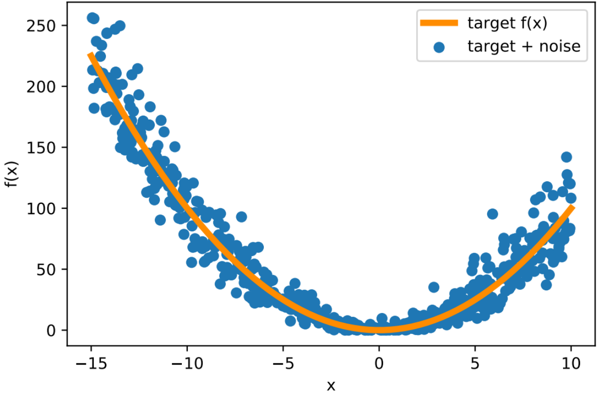
Bias-Variance Trade-off

Bias-Variance Intuition



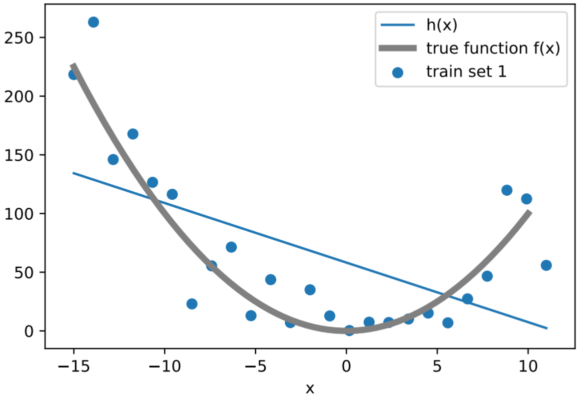
Bias-Variance Trade-off

Bias-Variance Intuition



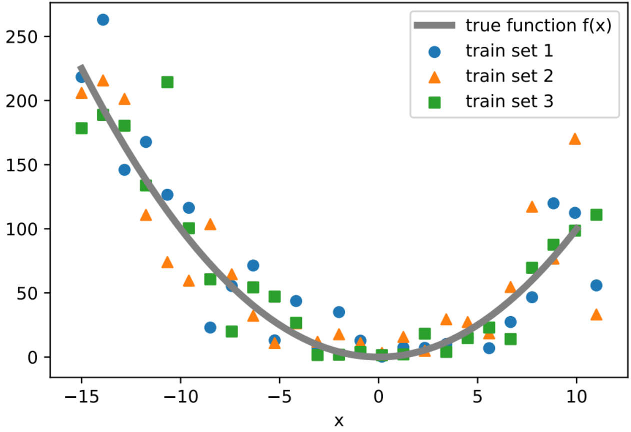
Bias-Variance Trade-off

Bias-Variance Intuition



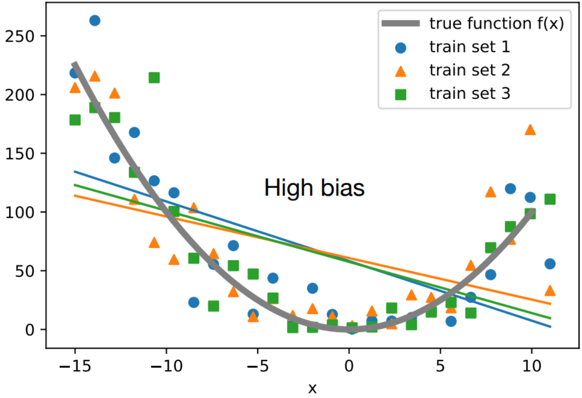
Bias-Variance Trade-off

Bias-Variance Intuition



Bias-Variance Trade-off

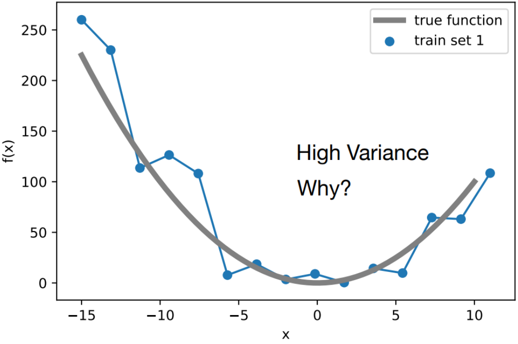
Bias-Variance Intuition



(There are two points where the bias is zero)

Bias-Variance Trade-off

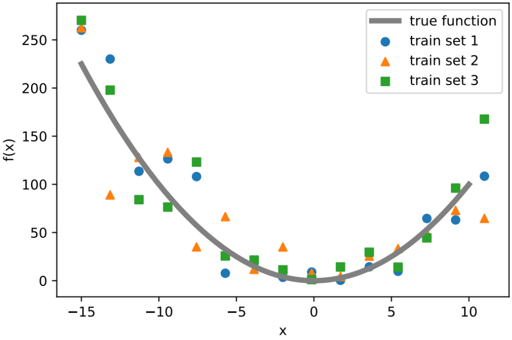
Bias-Variance Intuition



(here, I fit an unpruned decision tree)

Bias-Variance Trade-off

Bias-Variance Intuition

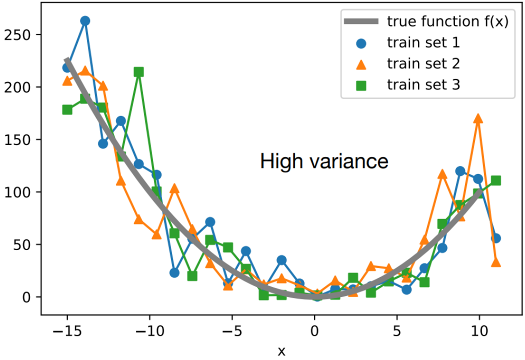


where f(x) is some true (target) function

suppose we have multiple training sets

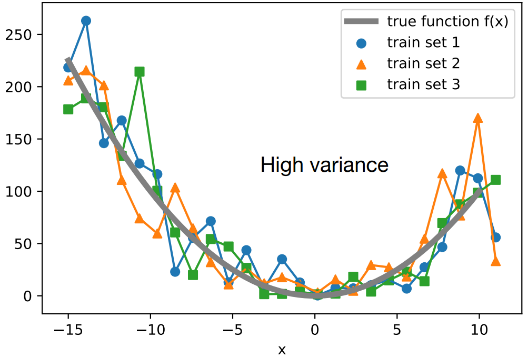
Bias-Variance Trade-off

Bias-Variance Intuition



Bias-Variance Trade-off

Bias-Variance Intuition



What happens if we take the average?
Does this remind you of something?

Bias-Variance Decomposition

Terminology
Point estimator θ of some parameter θ
(could also be a function, e.g., the hypothesis is an estimator of some target function)

$$\mathbf{Bias}(\theta) = E[\hat{\theta}] - \theta$$

$$\mathbf{Var}(\theta) = E[\hat{\theta}^2] - \left(E[\hat{\theta}]\right)^2$$

Bias-Variance Decomposition

$$Loss = Bias + Variance + Noise$$

Bias-Variance Decomposition of Squared Error

- the true or target function: $y = f(x)$

Bias-Variance Decomposition of Squared Error

- the true or target function: $y = f(x)$
- the predicted target value: $\hat{y} = \hat{f}(x) = \hat{h}(x)$

Bias-Variance Decomposition of Squared Error

- the true or target function: $y = f(x)$
- the predicted target value: $\hat{y} = \hat{f}(x) = \hat{h}(x)$
- the squared loss: $S = (y - \hat{y})^2$

(x is a particular data point e.g,. in the test set; the expectation is over training sets)

Bias-Variance Decomposition of Squared Error

$$S = (y - \hat{y})^2$$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2\,(y - E[\hat{y}])\,(E[\hat{y}] - \hat{y})\end{aligned}$$

Bias-Variance Decomposition of Squared Error

$$S = (y - \hat{y})^2$$

$$\begin{aligned}
(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\
&= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2\,(y - E[\hat{y}])\,(E[\hat{y}] - \hat{y})
\end{aligned}$$

$$E[S] = E[(y - \hat{y})^2]$$

$$\begin{aligned}
E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] \\
&= [\text{Bias}]^2 + \text{Variance}.
\end{aligned}$$

Bias-Variance Decomposition of Squared Error

$$S = (y - \hat{y})^2$$

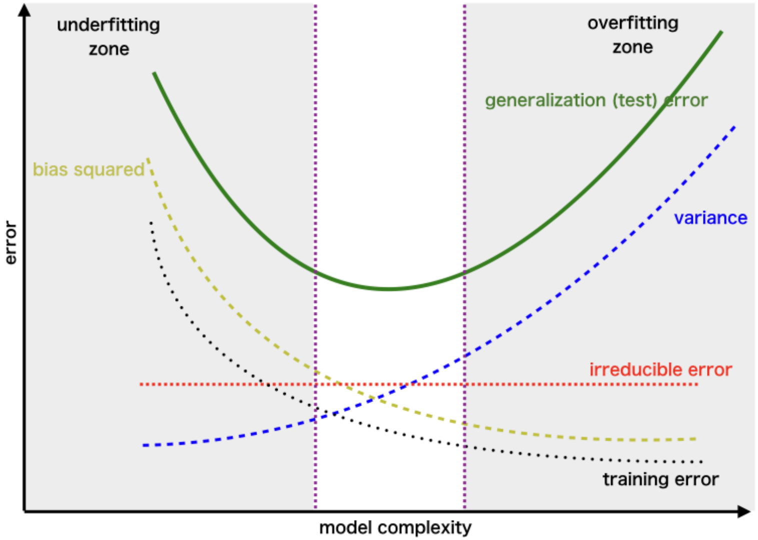
$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2\,(y - E[\hat{y}])\,(E[\hat{y}] - \hat{y})\end{aligned}$$

$$E[S] = E[(y - \hat{y})^2]$$

$$\begin{aligned}E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] \\ &= [\text{Bias}]^2 + \text{Variance}.\end{aligned}$$

$$\begin{aligned}E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] &= 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] \\ &= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})] \\ &= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}]) \\ &= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}]) \\ &= 0\end{aligned}$$

Bias-Variance Trade-off



Thank You!

Any Question?