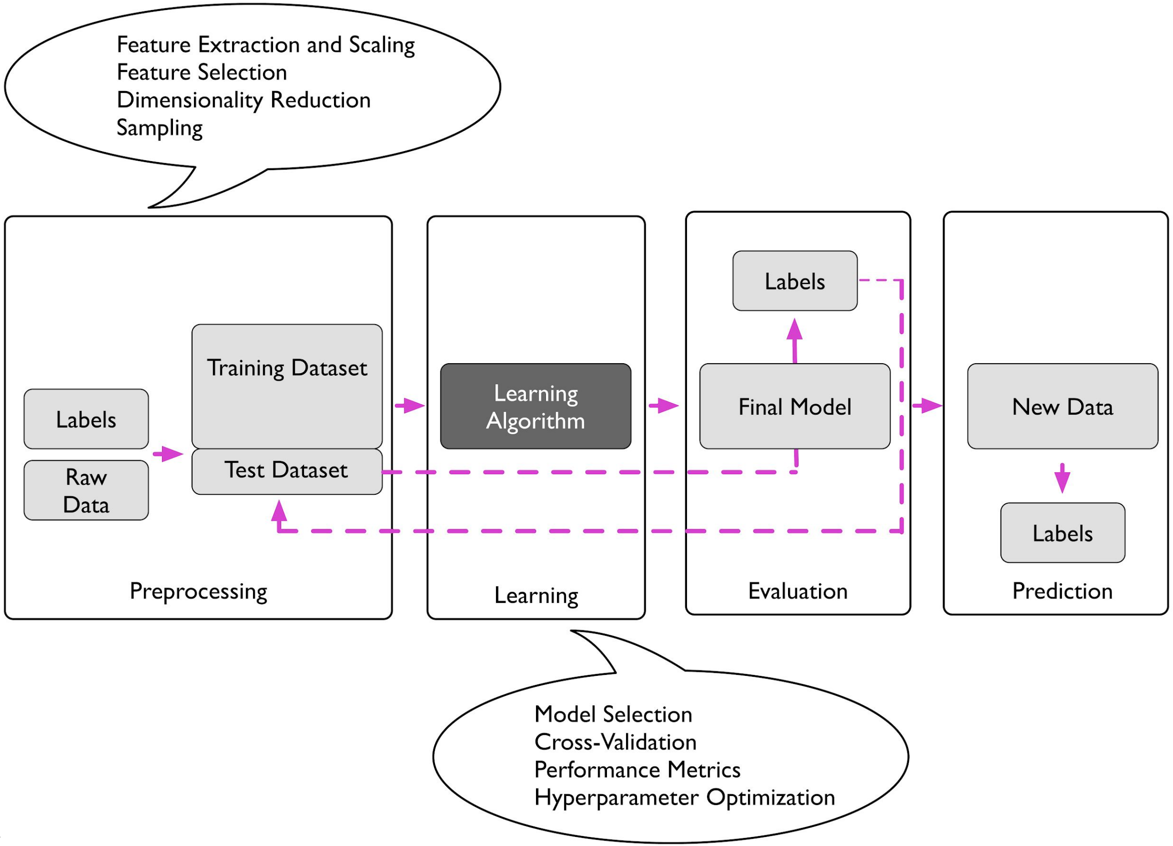
Lecture 14

Model evaluation

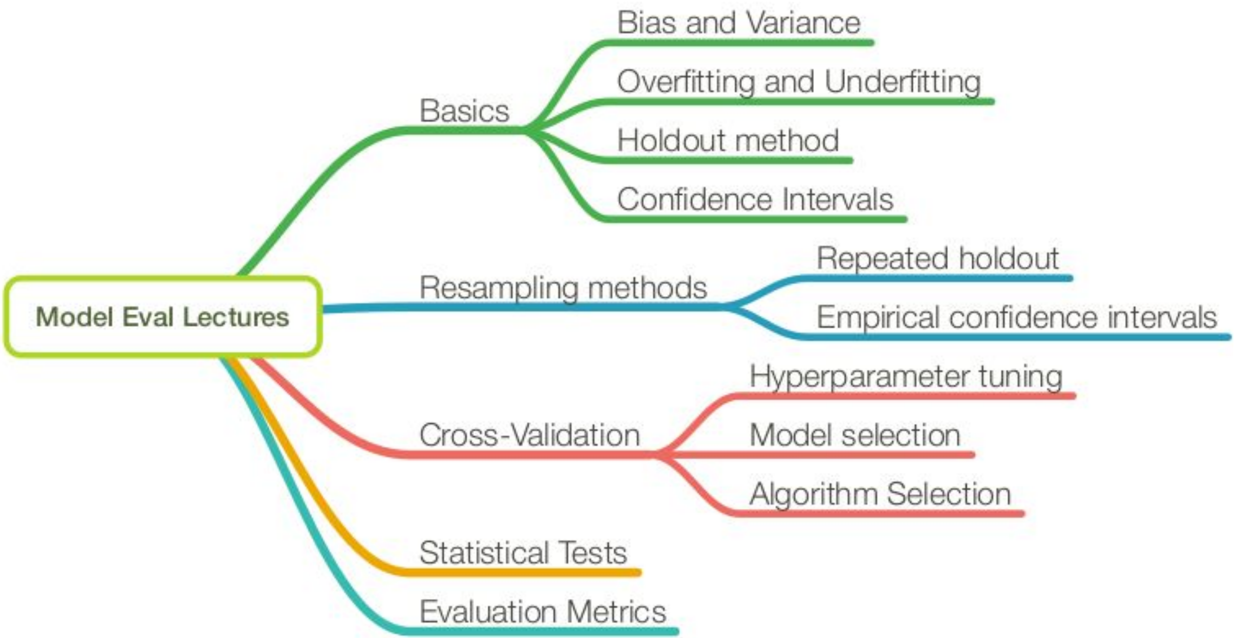
https://github.com/dalcimar/MA28CP-Intro-to-Machine-Learning

UTFPR - Federal University of Technology - Paraná
https://www.dalcimar.com/

Machine learning pipeline



Lecture overview



Lecturer Overview

- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- **Error(loss) in classification and regression problems**
- Bias-Variance Decomposition of the Squared Error
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the 0/1 Loss
- Other Forms of Bias

Generalization Performance

Want a model to "generalize" well to **unseen** data

- "high generalization accuracy" or
- "low generalization error"

Assumptions

i.i.d. assumption: training and test examples are independent and identically distributed (drawn from the same joint probability distribution, P(X, y))

For some random model that **has not been fitted to the training set**, we expect the training error is **approximately similar** the test error

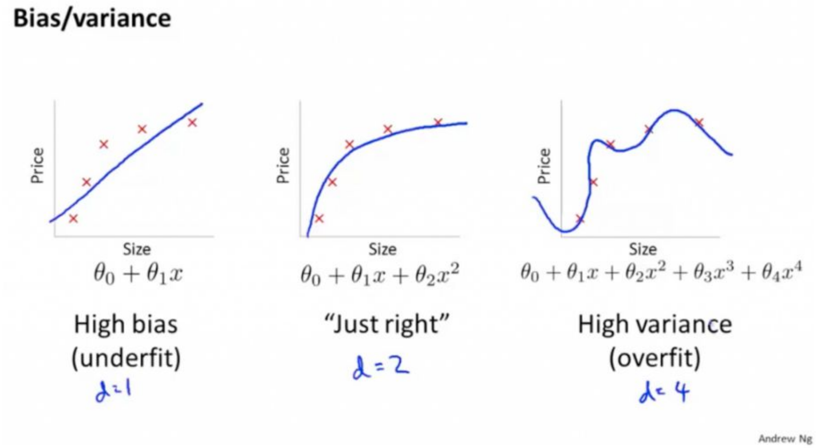
The training error or accuracy provides an **optimistically biased estimate** of the generalization performance

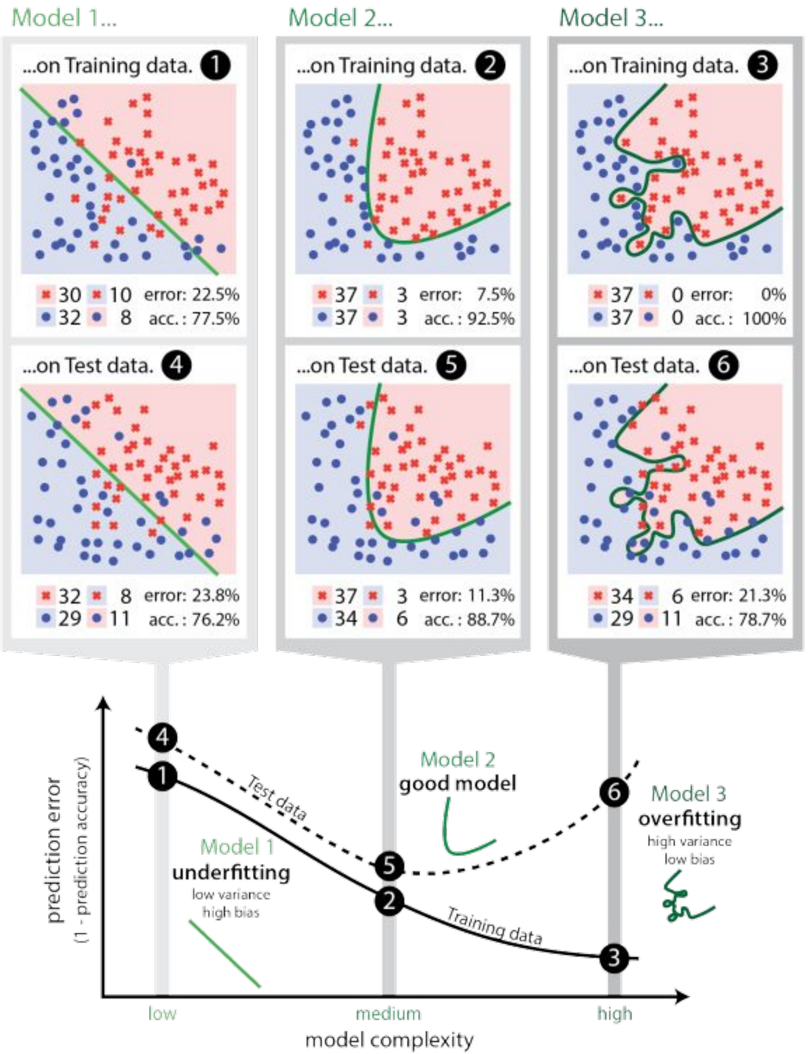
Model Capacity

Underfitting: both the training and test error are high

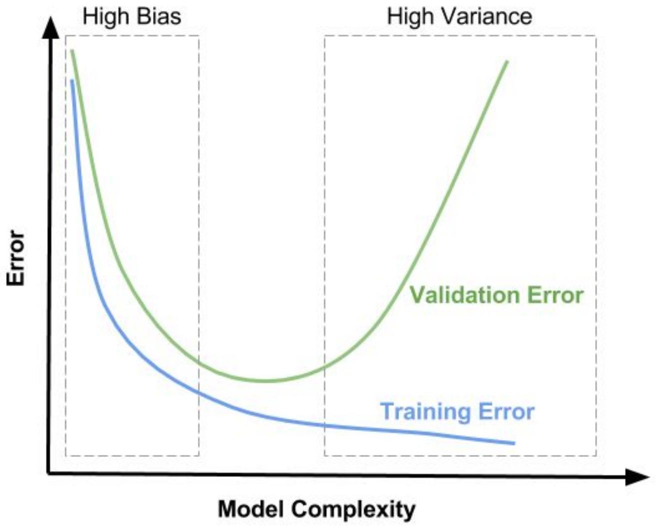
Overfitting: gap between training and test error (where test error is larger)

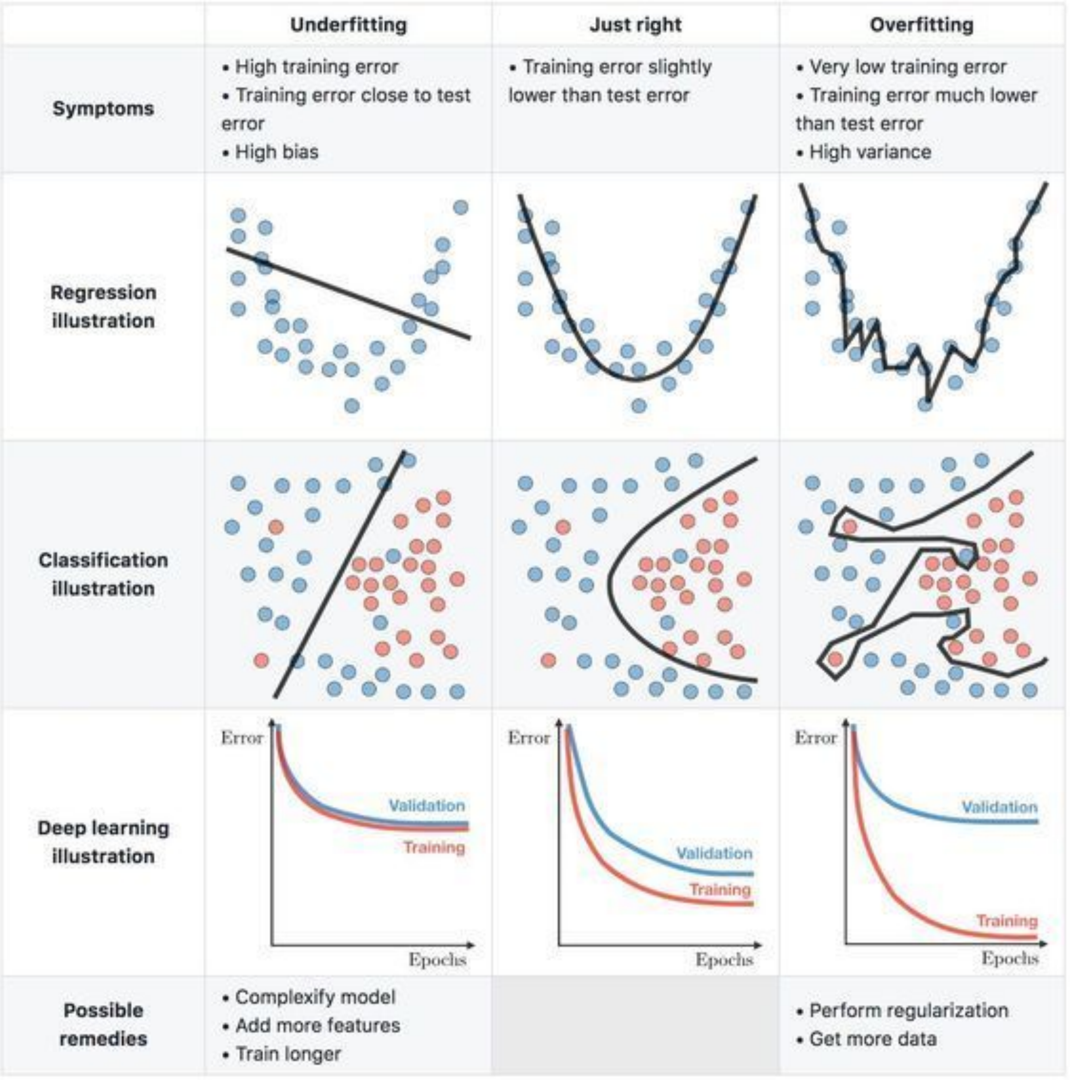
- Large hypothesis space being searched by a learning algorithm -> high tendency to overfit





Overfitting and Underfit





	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration	Error Validation Training Epochs	Error Validation Training Epochs	Error Validation Training Epochs
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

Lecturer Overview

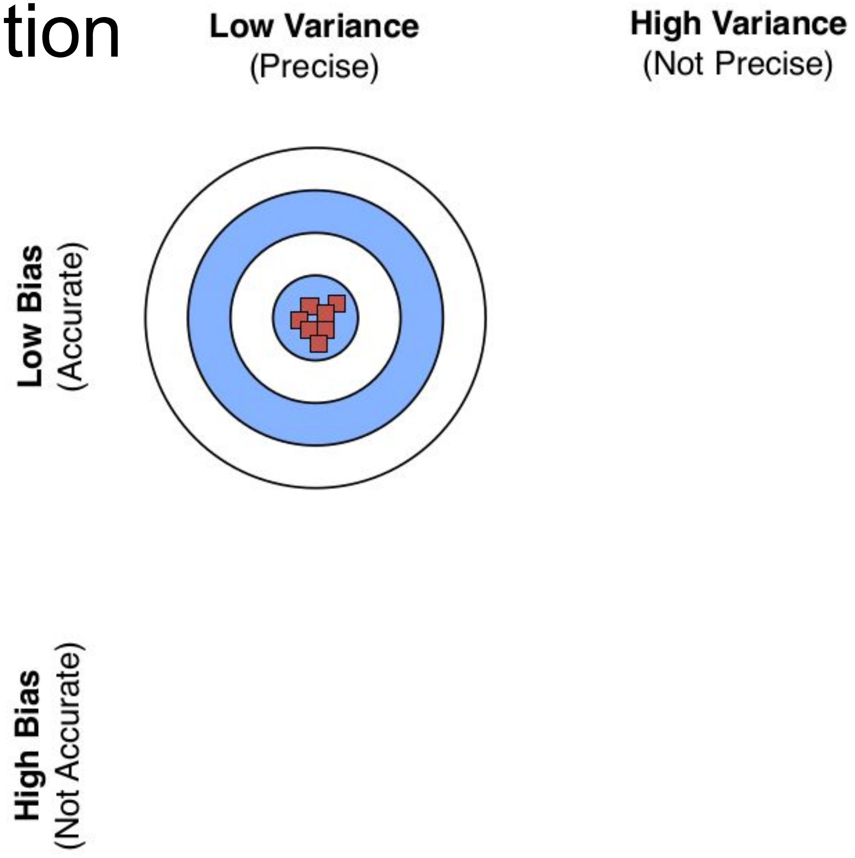
- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- **Error(loss) in classification and regression problems**
- Bias-Variance Decomposition of the Squared Error
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the 0/1 Loss
- Other Forms of Bias

Bias-Variance Decomposition

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

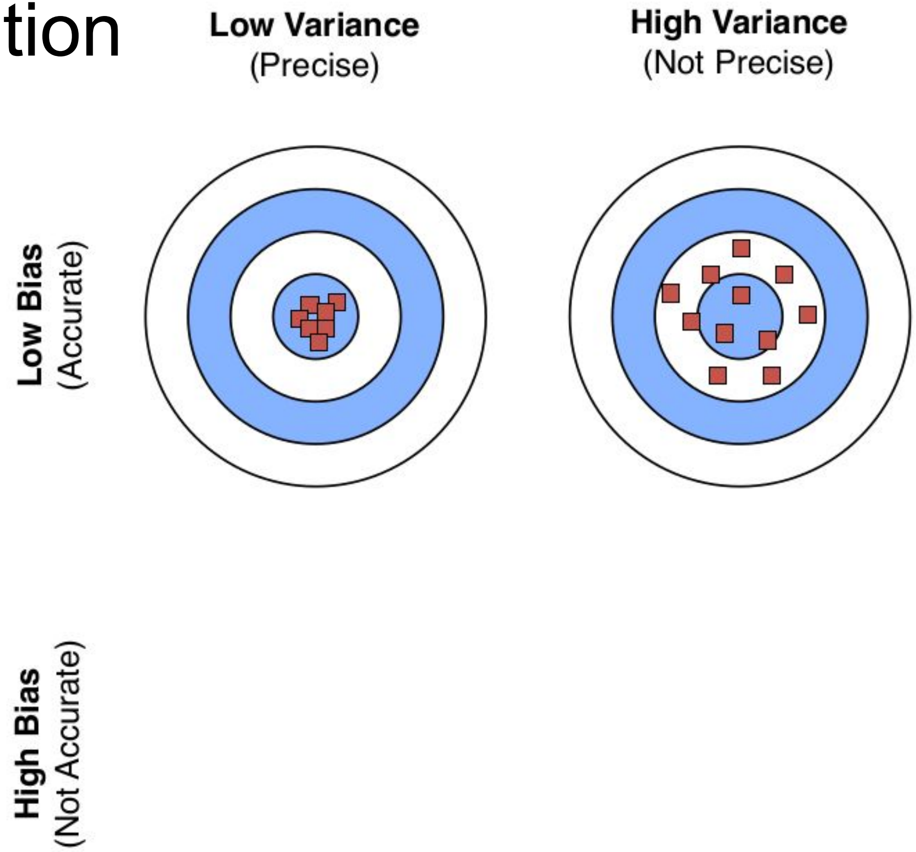
Bias-Variance Decomposition

Loss = Bias + Variance + Noise



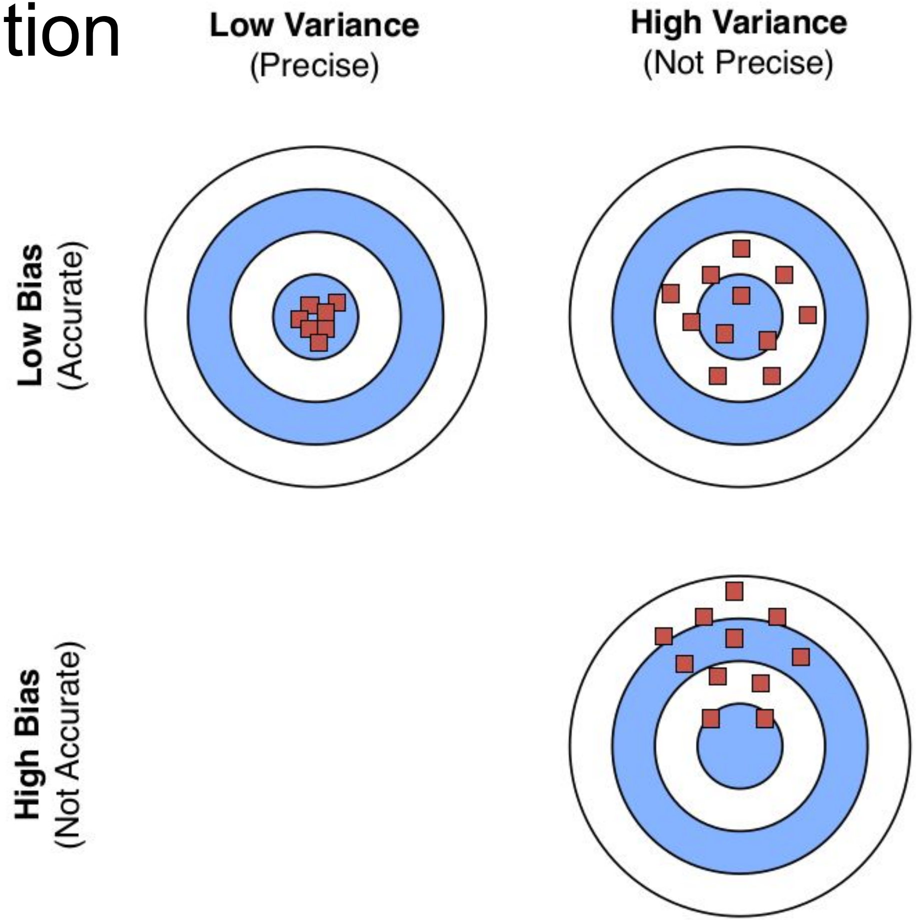
Bias-Variance Decomposition

Loss = Bias + Variance + Noise



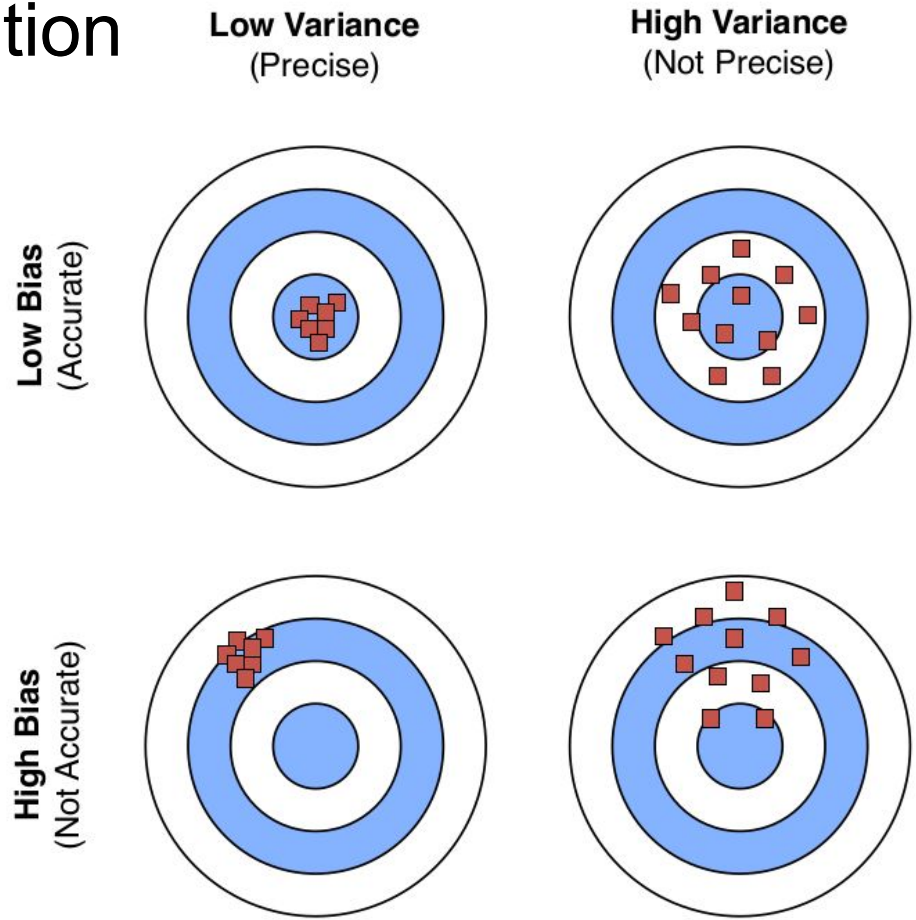
Bias-Variance Decomposition

Loss = Bias + Variance + Noise

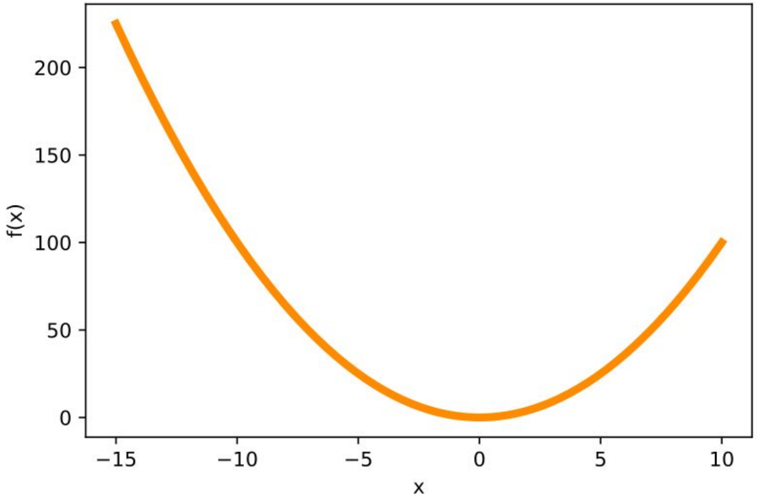


Bias-Variance Decomposition

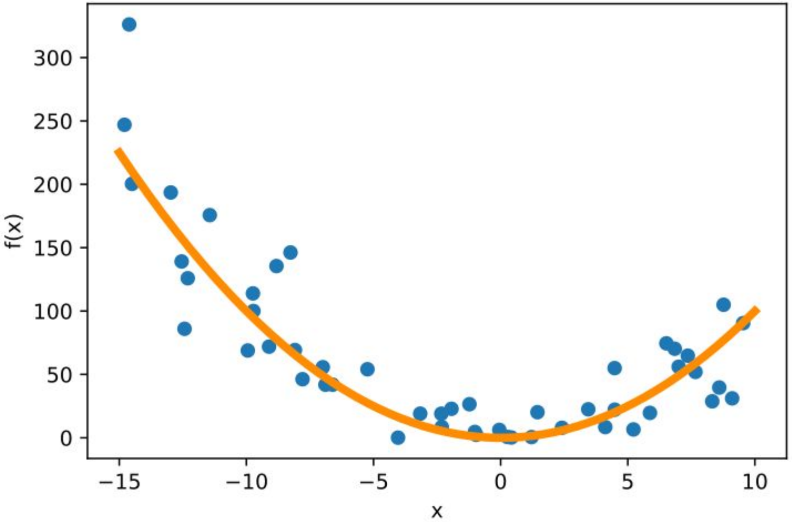
Loss = Bias + Variance + Noise



Bias and Variance Intuition



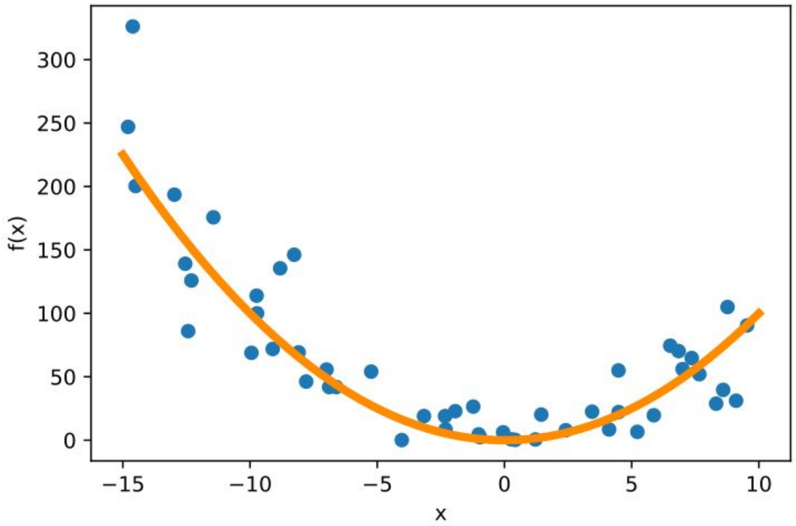
where f(x) is some true (target) function



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

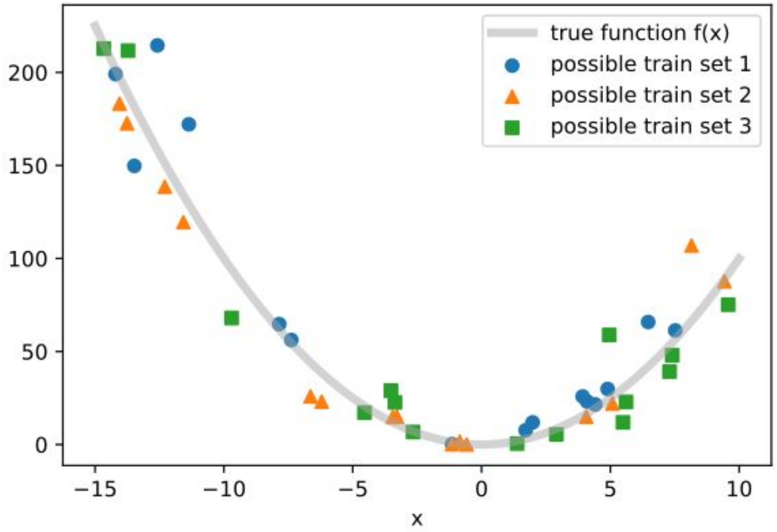
the blue dots are a training dataset;
here, I added some random Gaussian noise

Bias and Variance Intuition



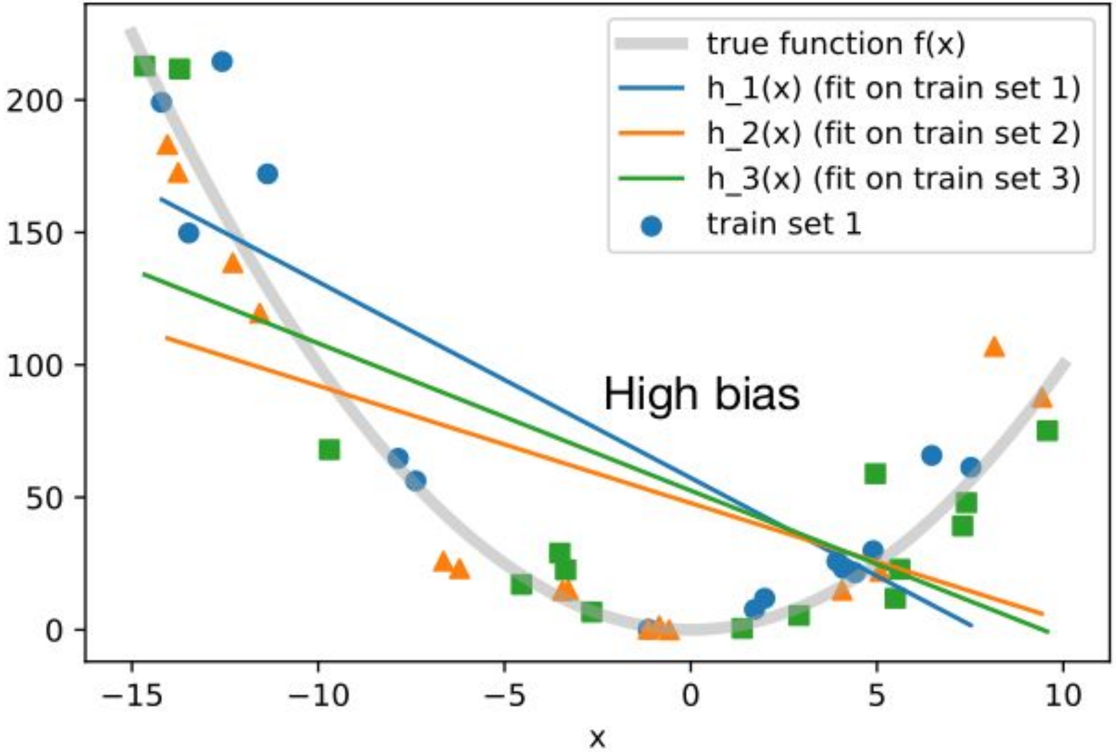
where f(x) is some true (target) function

the blue dots are a training dataset;
here, I added some random Gaussian noise



Bias and Variance Intuition

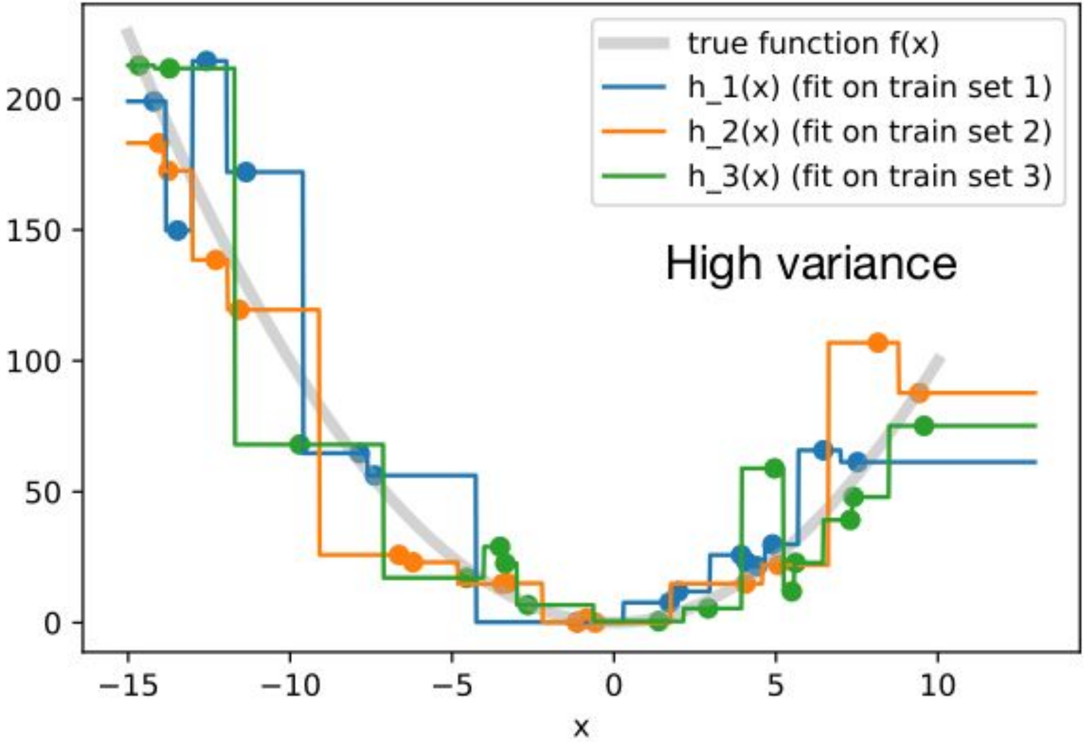
Suppose we have multiple training sets



linear regression models

Bias and Variance Intuition

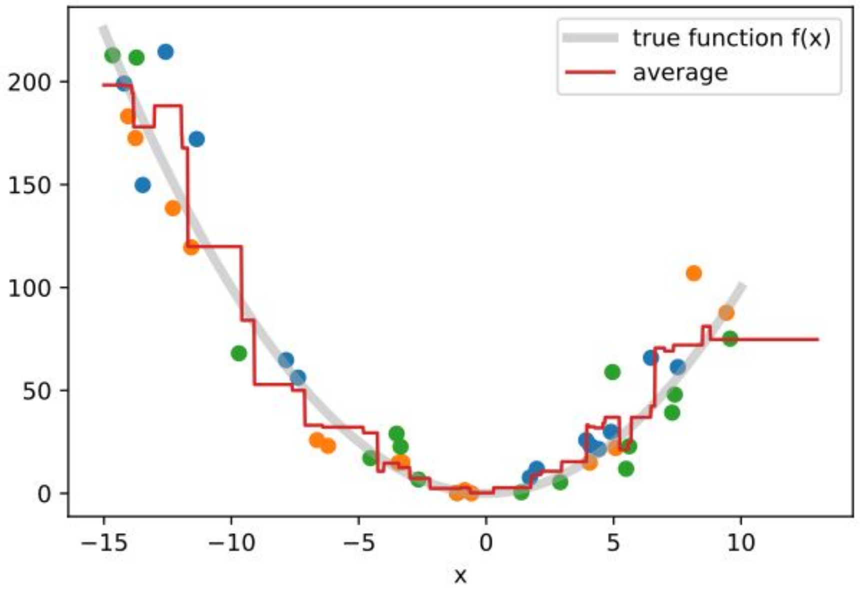
Suppose we have multiple training sets

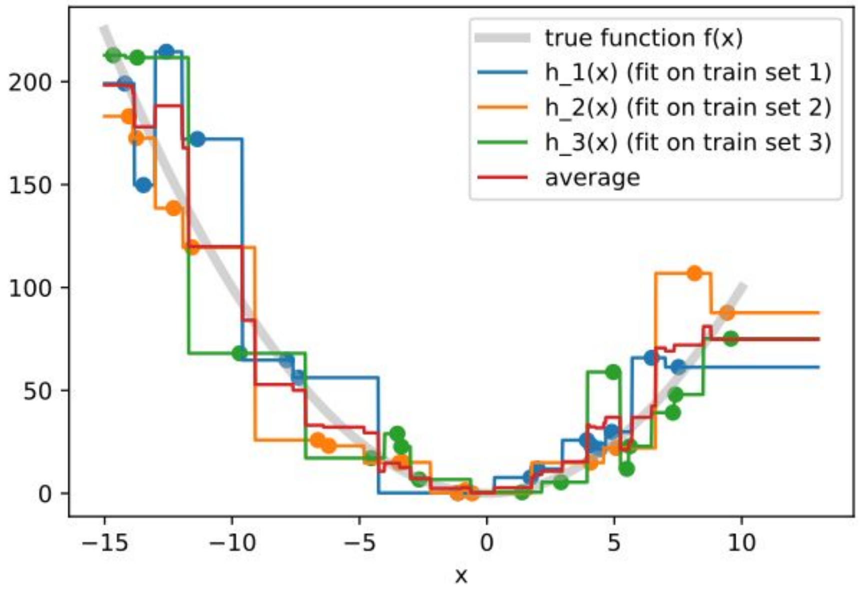


Bias and Variance Intuition

What happens if we take the average?

Does this remind you of something?





Terminology

Point estimator $\hat{\theta}$ of some parameter θ

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

$$\text{Bias} = E[\hat{\theta}] - \theta$$

General Definition

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

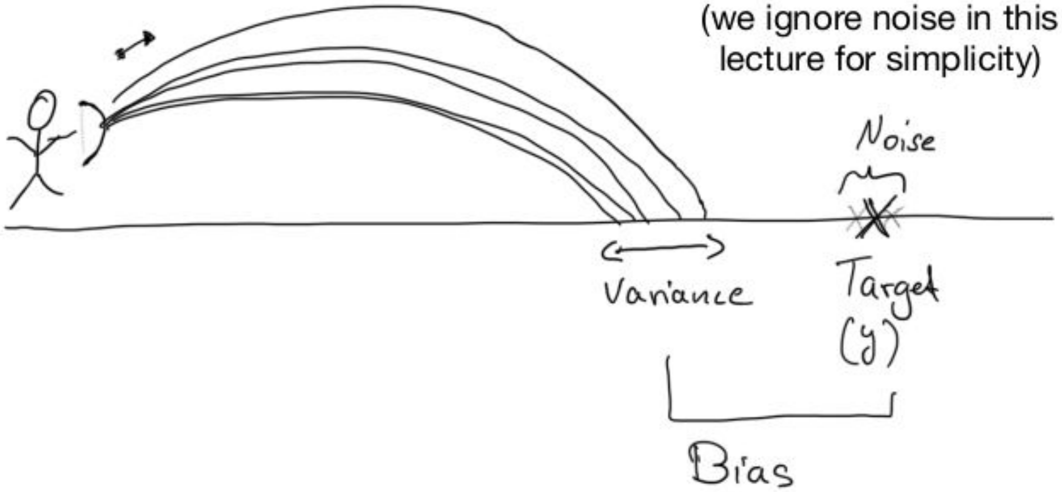
$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

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

Terminology

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

Intuition



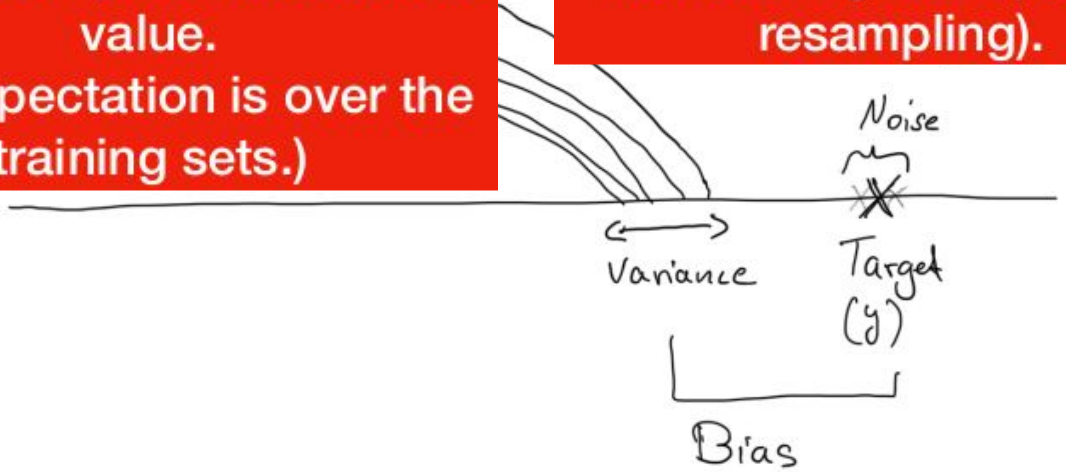
Terminology

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

Bias is the difference between the average estimator from different training samples and the true value.
(The expectation is over the training sets.)

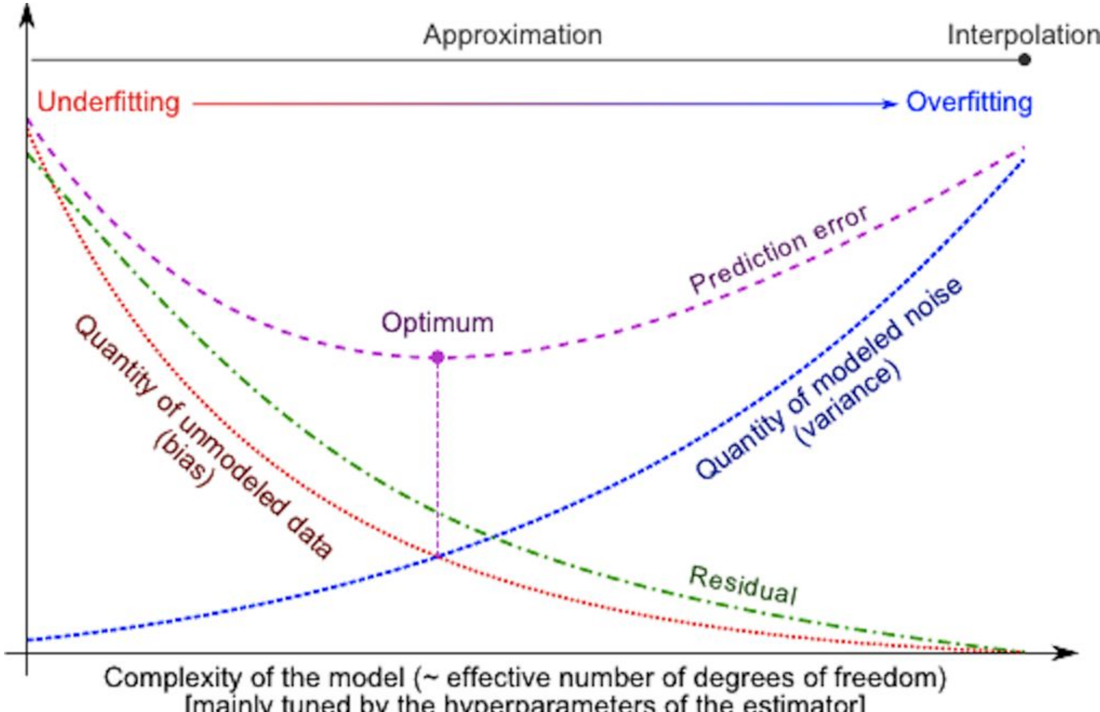
$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

The variance provides an estimate of how much the estimate varies as we vary the training data (e.g., by resampling).



Bias-Variance Decomposition

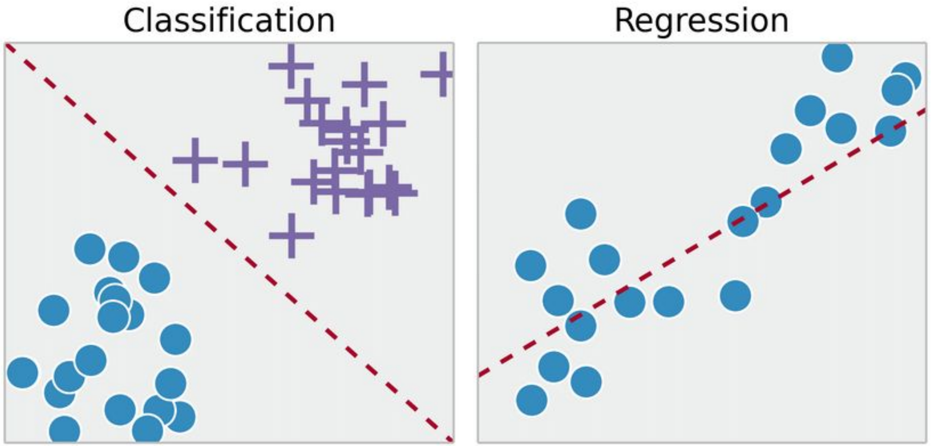
Loss = Bias + Variance + Noise

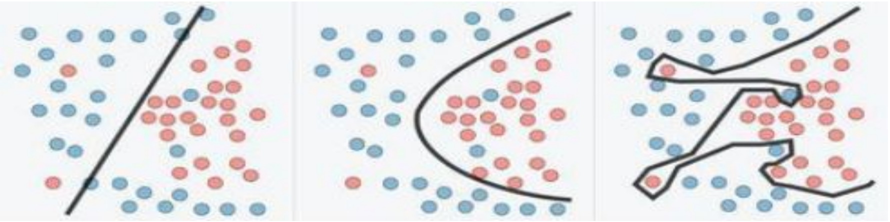


Lecturer Overview

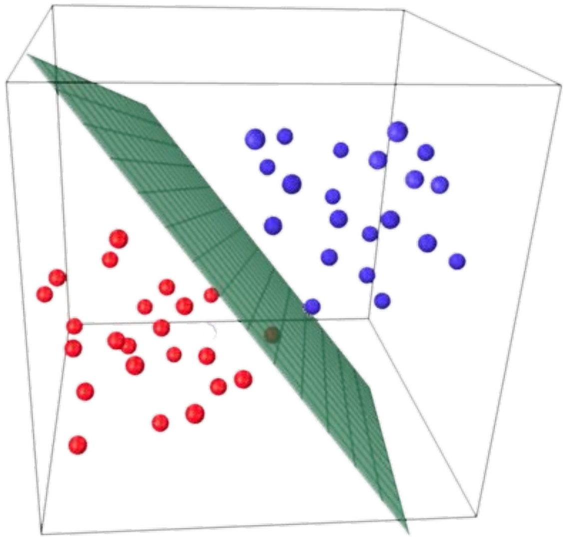
- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- **Error (loss) in classification and regression problems**
- Bias-Variance Decomposition of the Squared Error
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the 0/1 Loss
- Other Forms of Bias

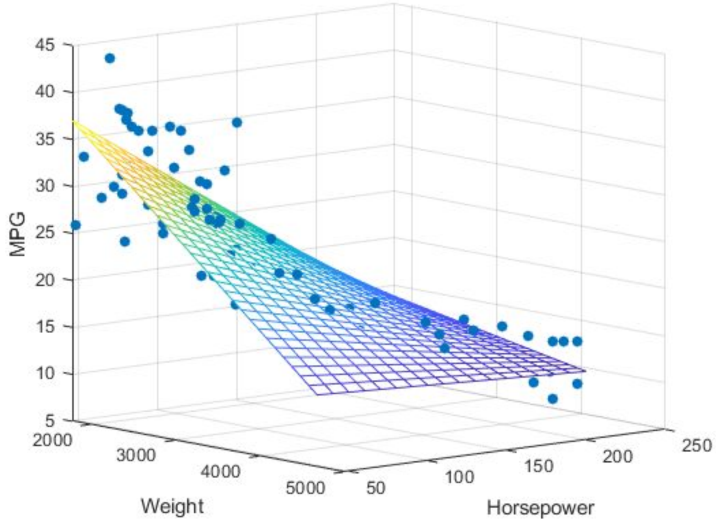
Classification x regression





Classification x regression





0-1 loss in classification

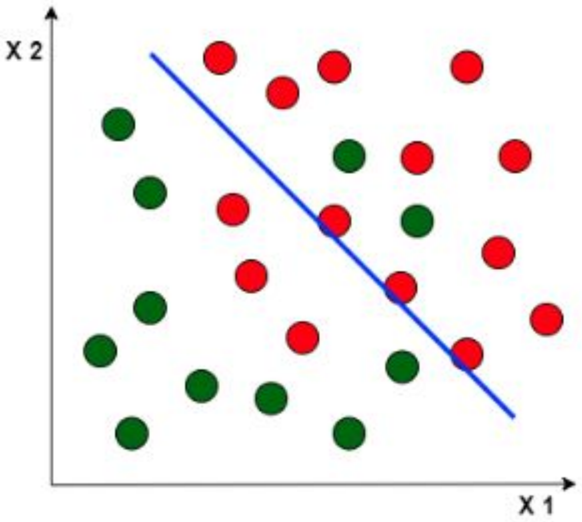
accuracy = 1-error rate

- 0.8 = 1 - 0.2

0-1 loss

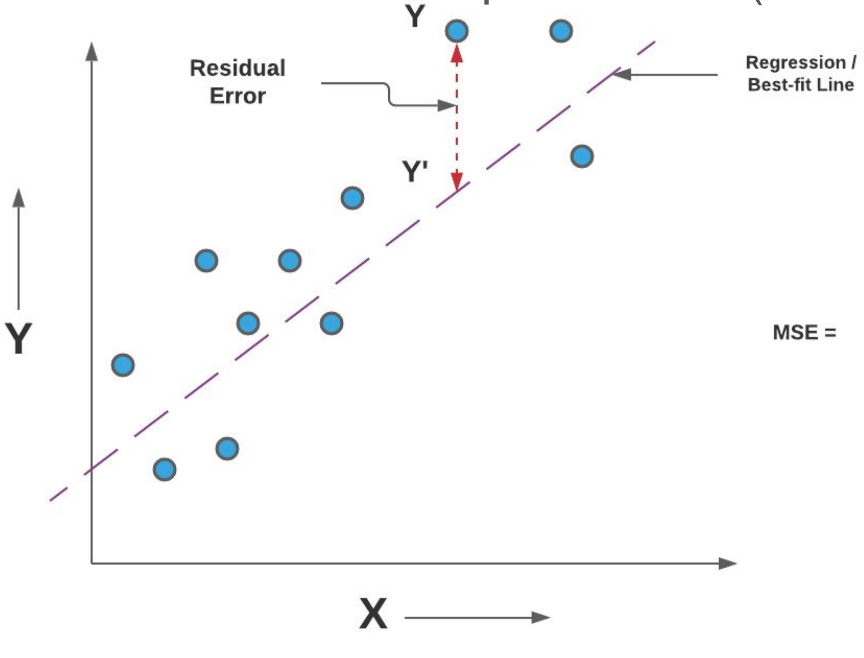
- $L_{0\text{-}1}$= 5

$$L_{0-1}(y_i, \hat{y}_i) = 1(\hat{y}_i \neq y_i)$$

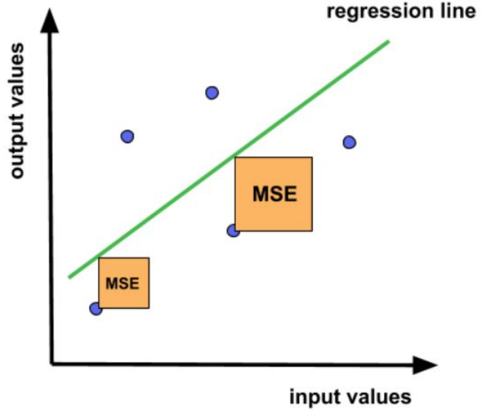


MSE loss in regression

Residuals and Mean Squared Error (MSE)

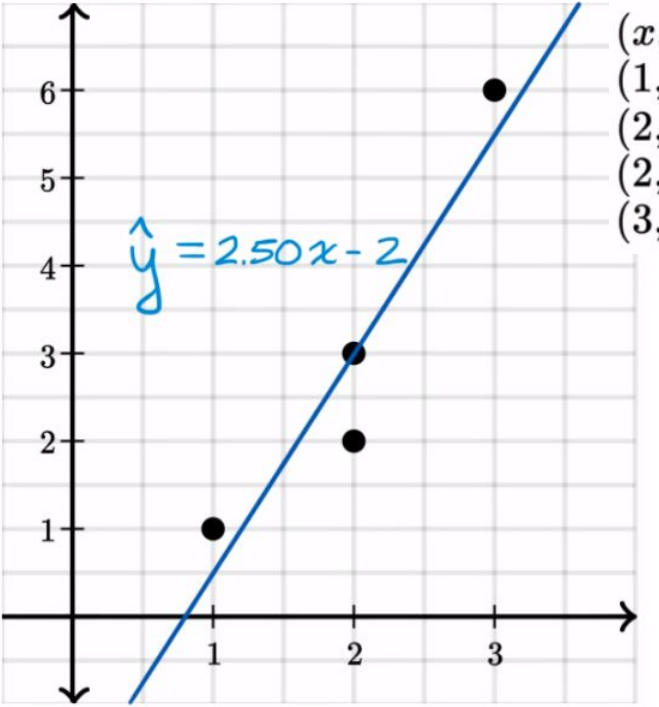


$$\text{MSE} = \frac{\text{Sum } (Y - Y')^2}{N}$$

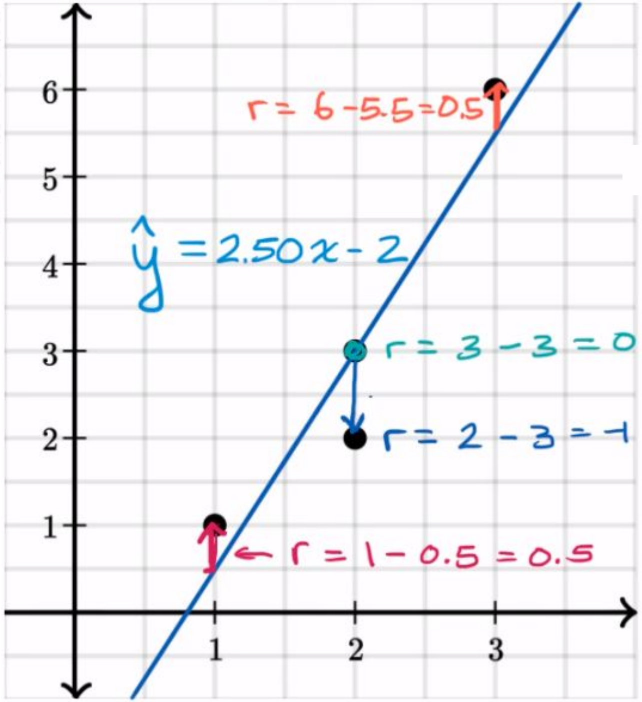


MSE loss in regression

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$



(x, y)
$(1, 1)$
$(2, 2)$
$(2, 3)$
$(3, 6)$



$$= (0.5)^2 + (0)^2 + (-1)^2 + (0.5)^2$$

$$= 1.5 / 4 = 0.375$$

Let's code!

7.2.1. Boston house prices dataset

Data Set Characteristics:

Number of Instances:	506
Number of Attributes:	13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.
Attribute Information (in order):	• CRIM per capita crime rate by town • ZN proportion of residential land zoned for lots over 25,000 sq.ft. • INDUS proportion of non-retail business acres per town • CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) • NOX nitric oxides concentration (parts per 10 million) • RM average number of rooms per dwelling • AGE proportion of owner-occupied units built prior to 1940 • DIS weighted distances to five Boston employment centres • RAD index of accessibility to radial highways • TAX full-value property-tax rate per $10,000 • PTRATIO pupil-teacher ratio by town • B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town • LSTAT % lower status of the population • MEDV Median value of owner-occupied homes in $1000's
Missing Attribute Values:	None
Creator:	Harrison, D. and Rubinfeld, D.L.

Lecturer Overview

- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- Error (loss) in classification and regression problems
- **Bias-Variance Decomposition of the Squared Error**
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the 0/1 Loss
- Other Forms of Bias

Terminology

average prediction over the training sets

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

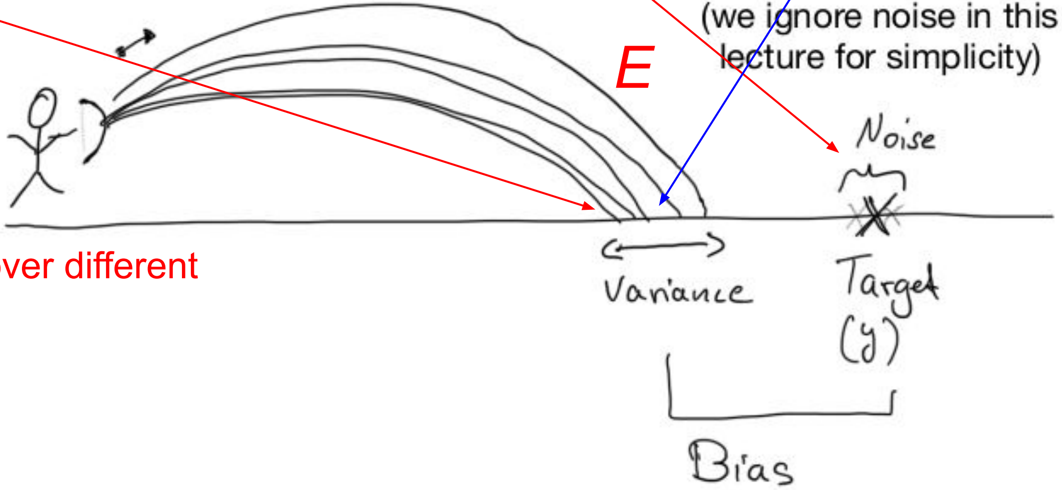
$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

spread out from the average prediction

Intuition

prediction

(we ignore noise in this lecture for simplicity)



models trained over different training sets

Bias-Variance of the Squared Error

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

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

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

Loss = Bias + Variance + Noise

for simplicity, we ignore the noise term

"ML Notation" for Squared Error Loss

$y = f(x)$ target

$\hat{y} = \hat{f}(x) = h(x)$ prediction

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

(Next slides: the expectation is over the training data, i.e, the average estimator from different training samples)

Bias-Variance of the Squared Error

$$y = f(x) \quad \text{target}$$

"ML Notation" for Squared Error Loss

$$\hat{y} = \hat{f}(x) = h(x) \quad \text{prediction}$$

$$S = (y - \hat{y})^2 \quad \text{squared error}$$

$$\begin{aligned}(a-b)^2 &= a^2 - 2ab + b^2 \\ &= a^2 + b^2 - 2ab\end{aligned}$$

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

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

Bias-Variance of the Squared Error

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

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

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

$$E\left[(y - \hat{y})^2\right] = (y - E[\hat{y}])^2 + E\left[(E[\hat{y}] - \hat{y})^2\right]$$

$$= \text{Bias}^2 + \text{Var}$$

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

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

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

Bias-Variance of the Squared Error

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

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

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