

## Lecture 14

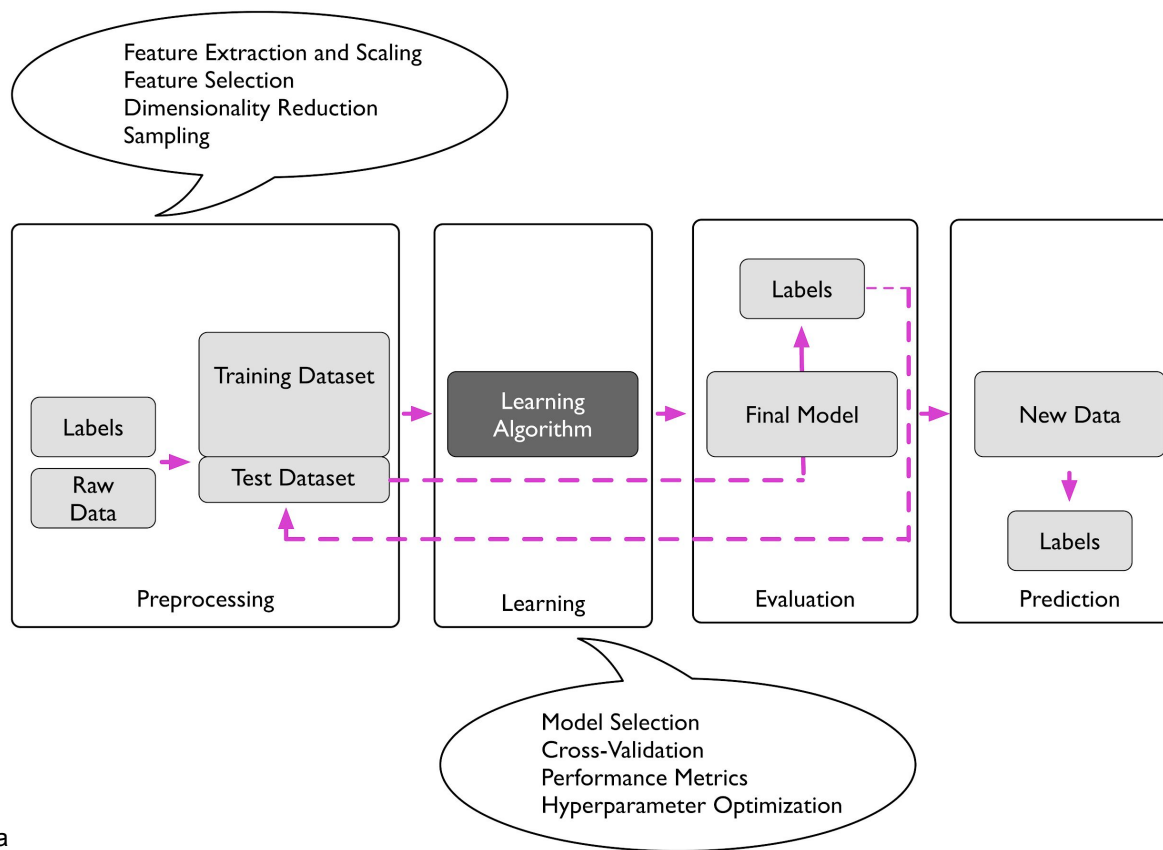
# Model evaluation

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

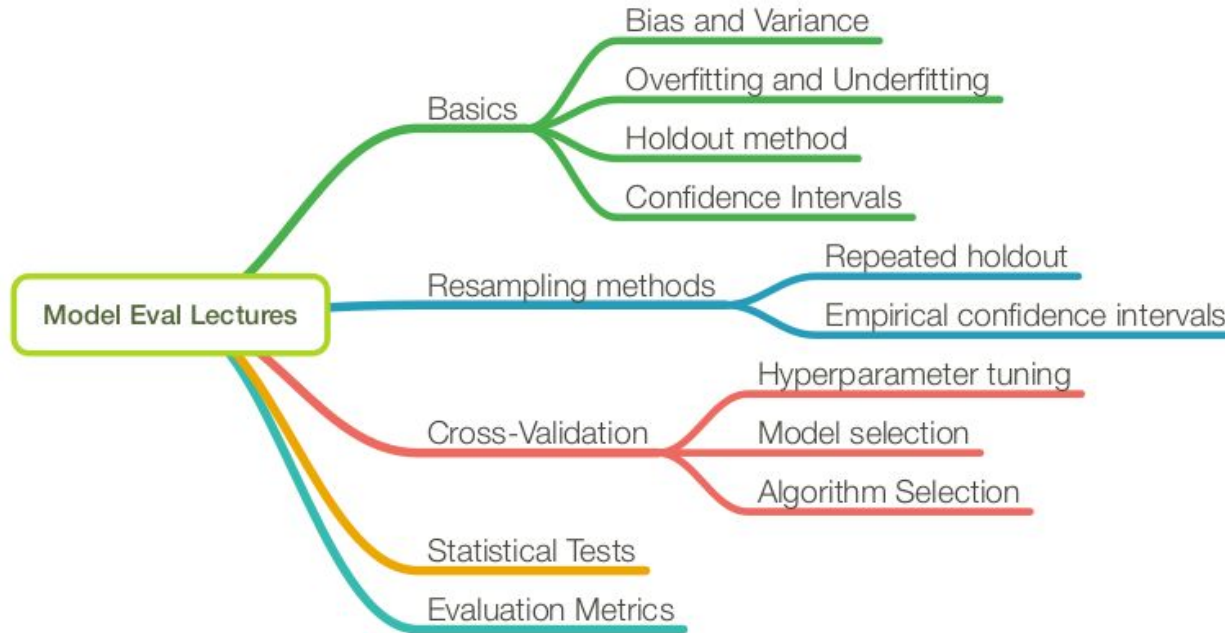
UTFPR - Federal University of Technology - Paraná

<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

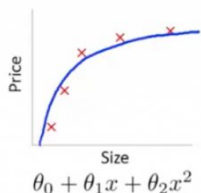
**Overfitting:** gap between training and test error (where test error is larger)

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

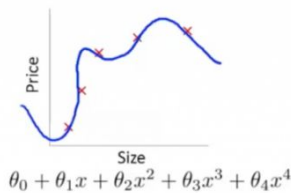
**Bias/variance**



High bias  
(underfit)  
 $d=1$



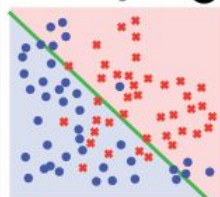
"Just right"  
 $d=2$



High variance  
(overfit)  
 $d=4$

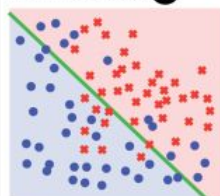
Model 1...

...on Training data. ①



■ 30 ■ 10 error: 22.5%  
● 32 ● 8 acc.: 77.5%

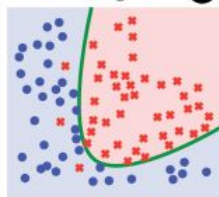
...on Test data. ④



■ 32 ■ 8 error: 23.8%  
● 29 ● 11 acc.: 76.2%

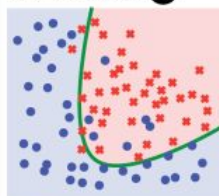
Model 2...

...on Training data. ②



■ 37 ■ 3 error: 7.5%  
● 37 ● 3 acc.: 92.5%

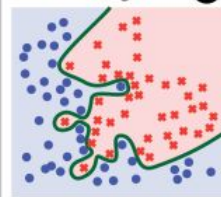
...on Test data. ⑤



■ 37 ■ 3 error: 11.3%  
● 34 ● 6 acc.: 88.7%

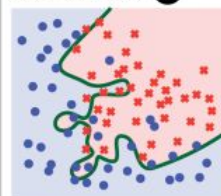
Model 3...

...on Training data. ③

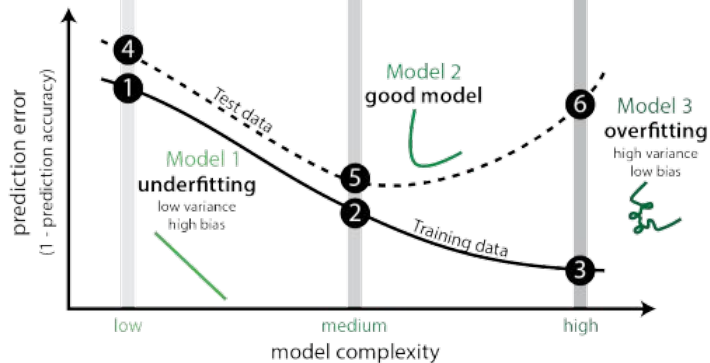


■ 37 ■ 0 error: 0%  
● 37 ● 0 acc.: 100%

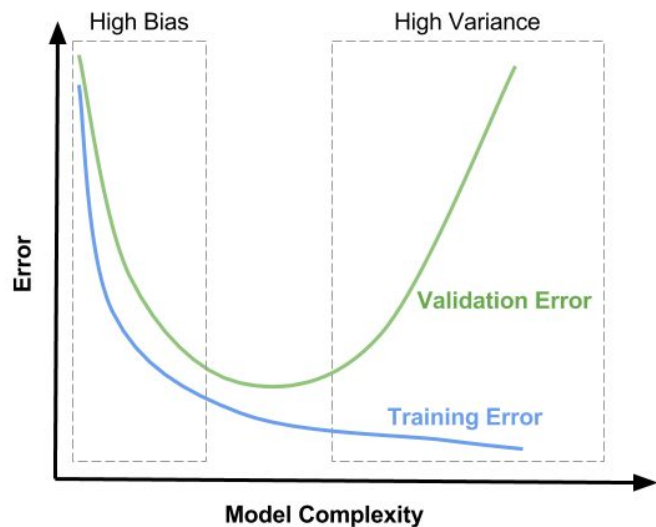
...on Test data. ⑥



■ 34 ■ 6 error: 21.3%  
● 29 ● 11 acc.: 78.7%



# Overfitting and Underfit



	Underfitting	Just right	Overfitting
<b>Symptoms</b>	<ul style="list-style-type: none"> <li>• High training error</li> <li>• Training error close to test error</li> <li>• High bias</li> </ul>	<ul style="list-style-type: none"> <li>• Training error slightly lower than test error</li> </ul>	<ul style="list-style-type: none"> <li>• Very low training error</li> <li>• Training error much lower than test error</li> <li>• High variance</li> </ul>
<b>Regression illustration</b>			
<b>Classification illustration</b>			
<b>Deep learning illustration</b>			
<b>Possible remedies</b>	<ul style="list-style-type: none"> <li>• Complexify model</li> <li>• Add more features</li> <li>• Train longer</li> </ul>		<ul style="list-style-type: none"> <li>• Perform regularization</li> <li>• Get more data</li> </ul>



# 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

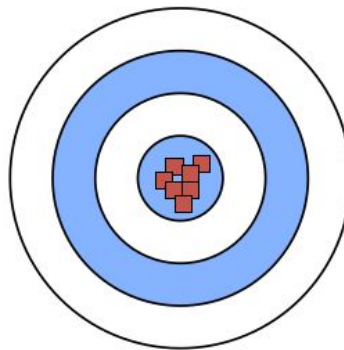
# Bias-Variance Decomposition

$$\text{Loss} = \text{Bias} + \text{Variance} + \text{Noise}$$

**Low Variance**  
(Precise)

**High Variance**  
(Not Precise)

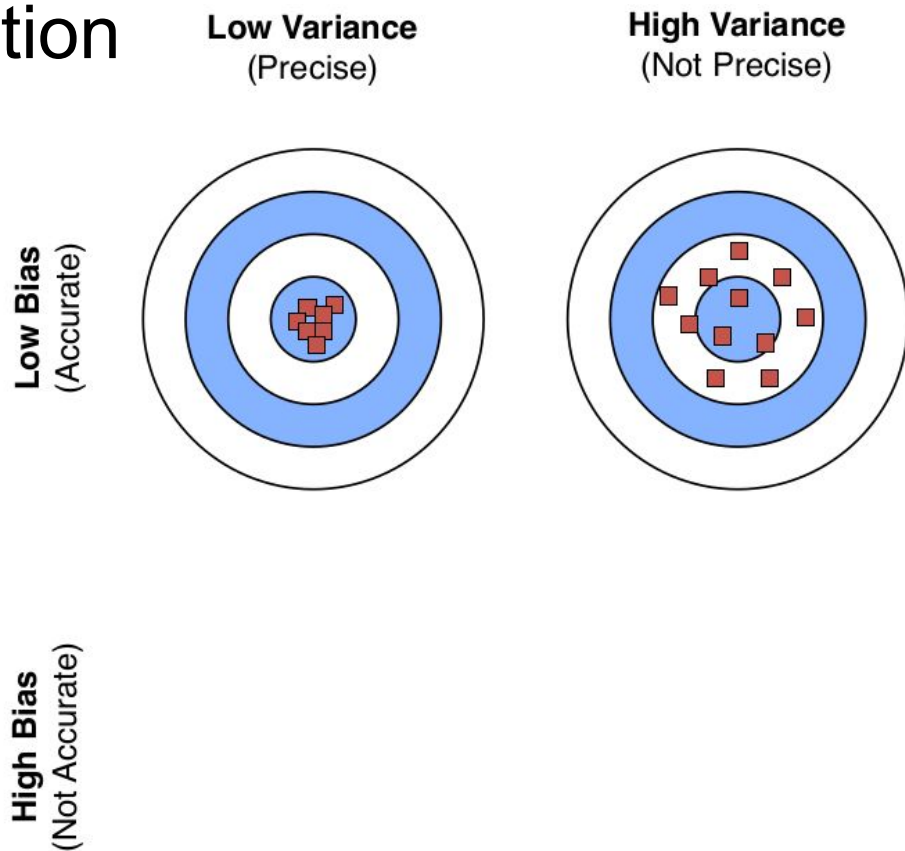
**Low Bias**  
(Accurate)



**High Bias**  
(Not Accurate)

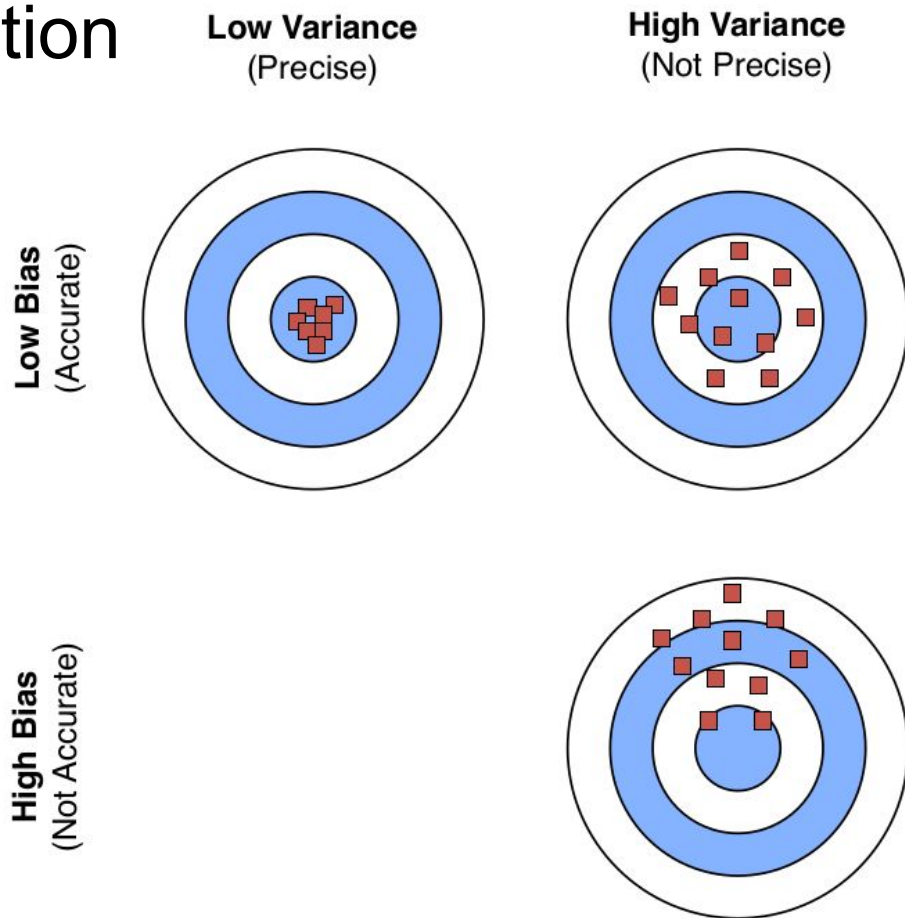
# Bias-Variance Decomposition

$$\text{Loss} = \text{Bias} + \text{Variance} + \text{Noise}$$



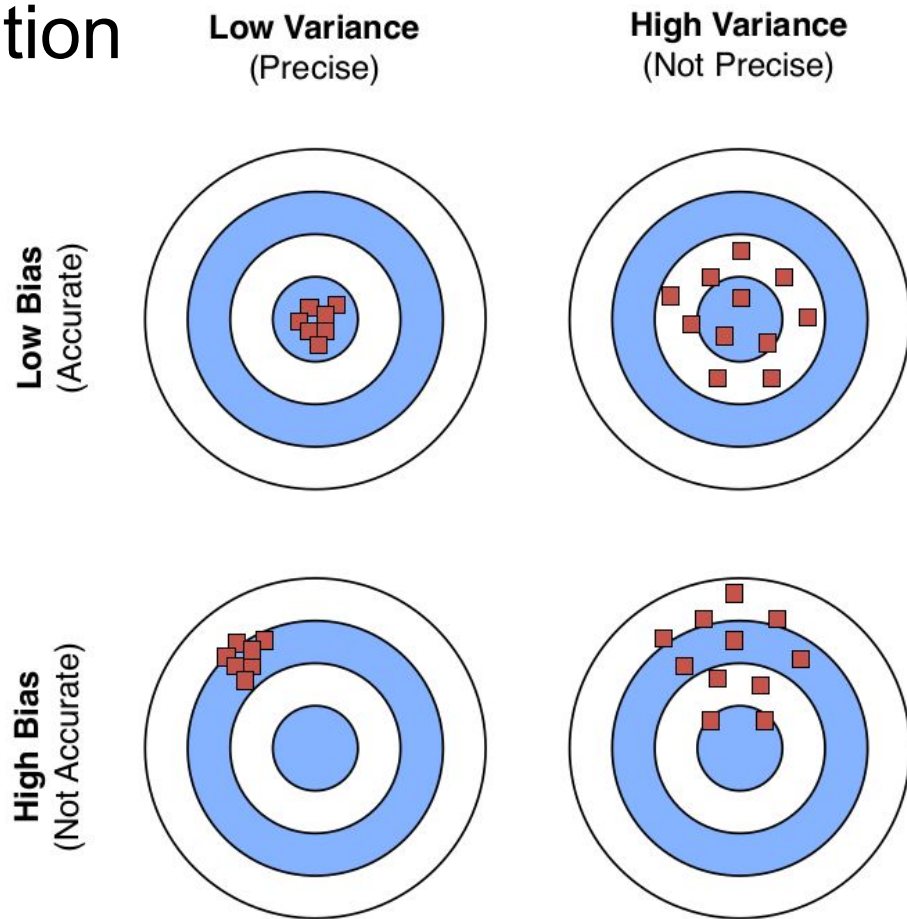
# Bias-Variance Decomposition

$$\text{Loss} = \text{Bias} + \text{Variance} + \text{Noise}$$

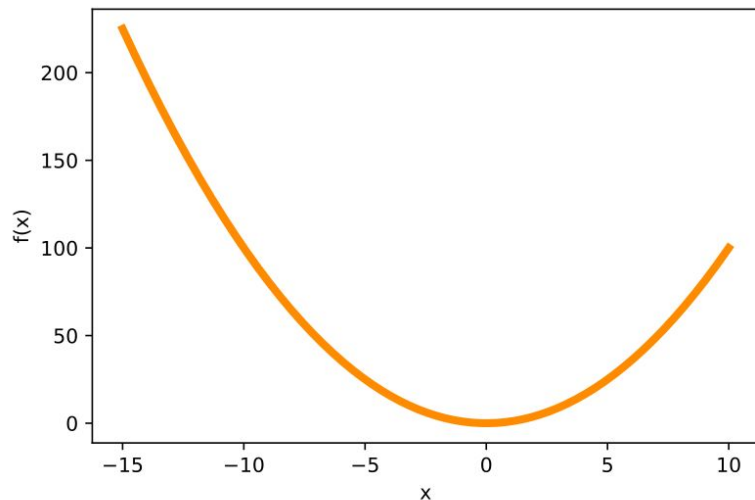


# Bias-Variance Decomposition

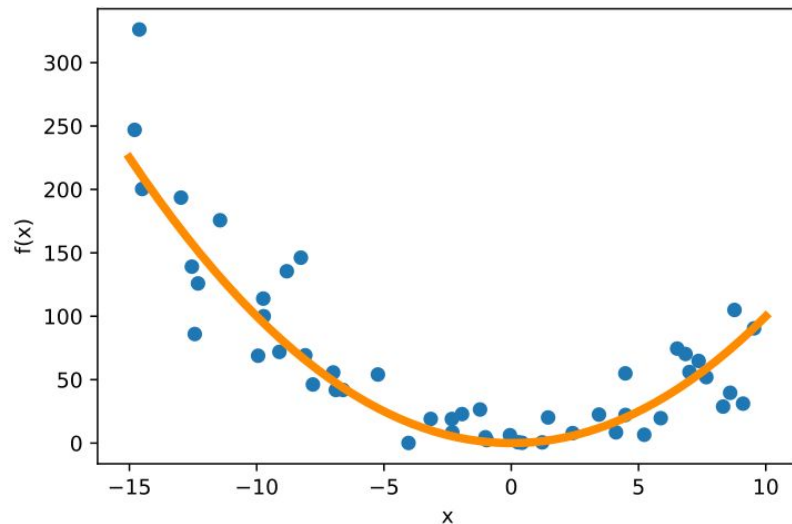
$$\text{Loss} = \text{Bias} + \text{Variance} + \text{Noise}$$



# Bias and Variance Intuition



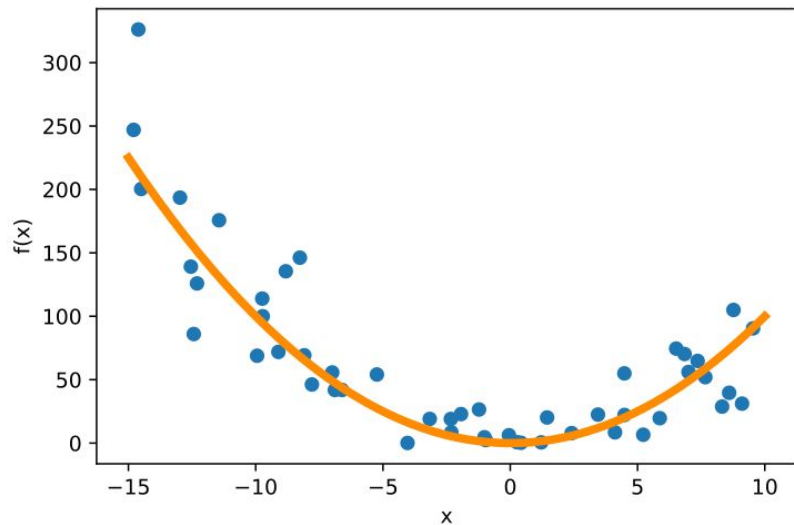
where  $f(x)$  is some true (target) function



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

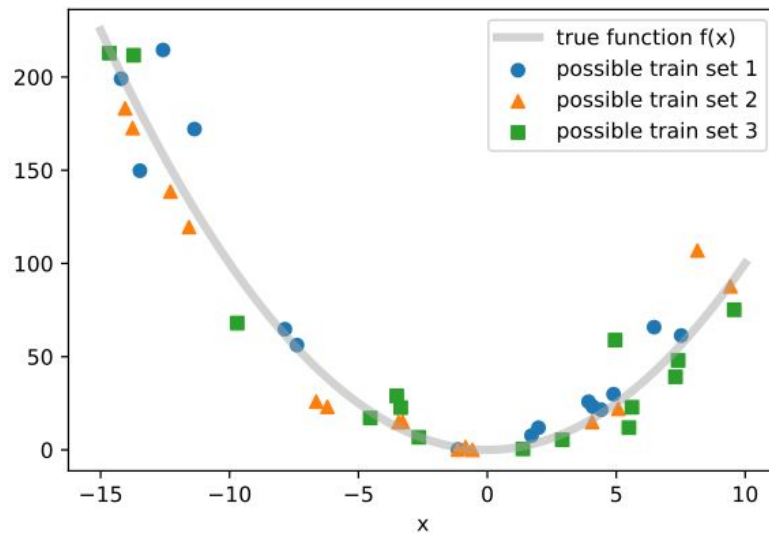
the blue dots are a training dataset;  
here, I added some random Gaussian noise

# Bias and Variance Intuition



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

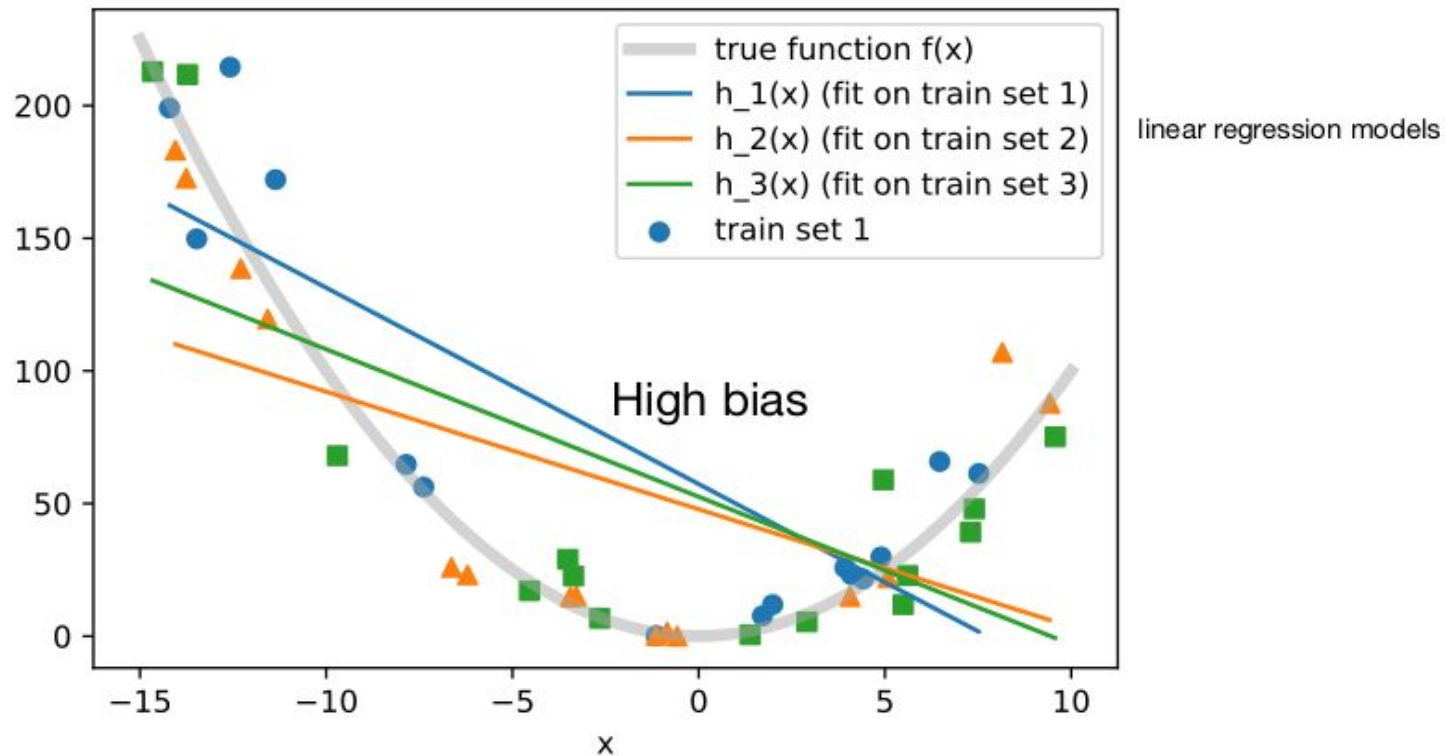
the blue dots are a training dataset;  
here, I added some random Gaussian noise





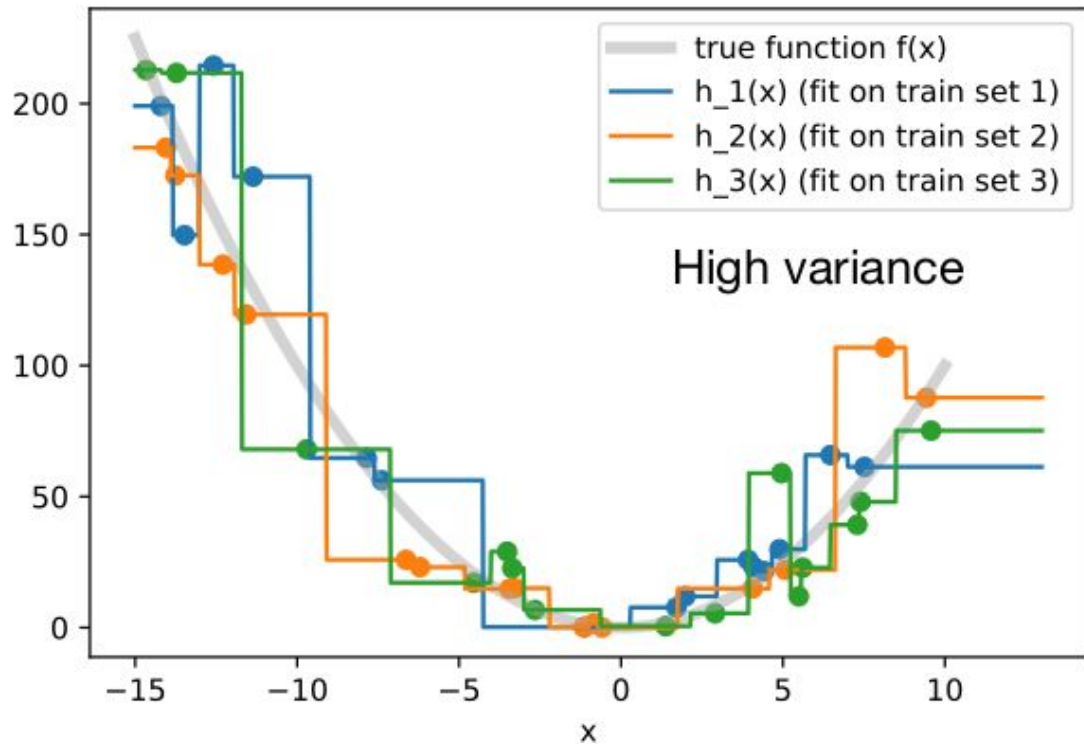
# Bias and Variance Intuition

Suppose we have multiple training sets



# Bias and Variance Intuition

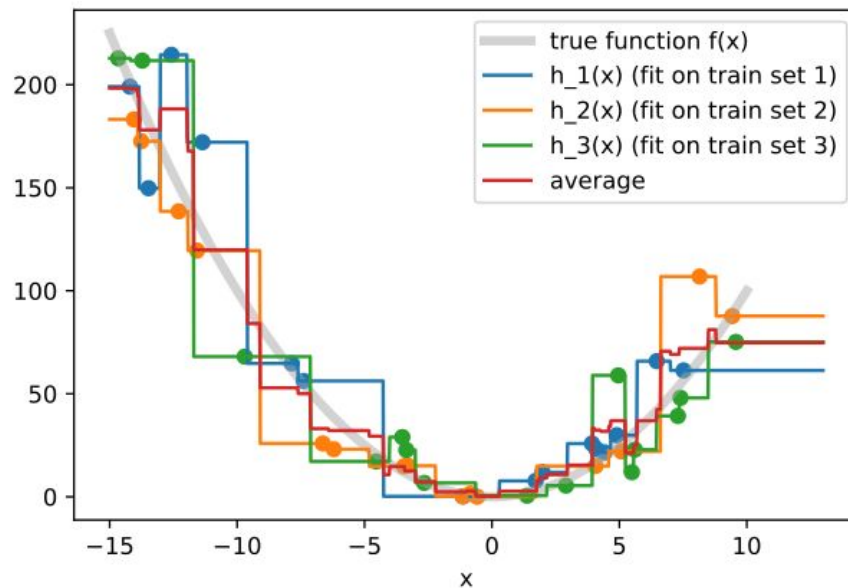
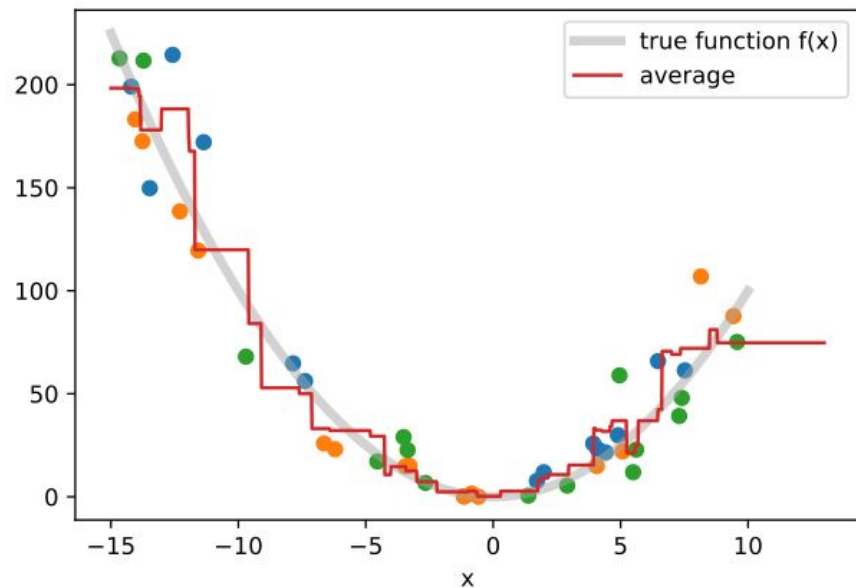
Suppose we have multiple training sets



# 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  $\theta$

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

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

General Definition

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

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

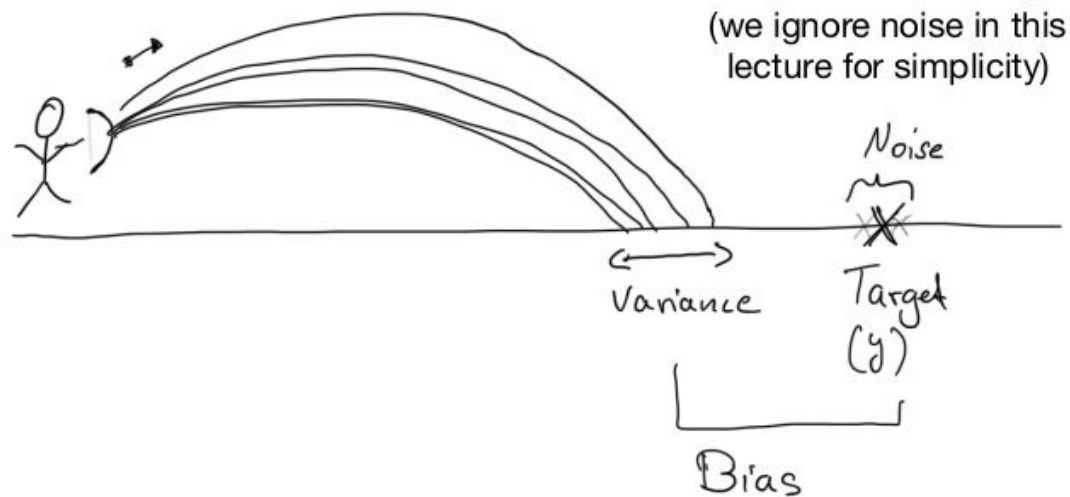
$$\text{Var}[\hat{\theta}] = E \left[ (E[\hat{\theta}] - \hat{\theta})^2 \right]$$

# Terminology

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

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

## Intuition



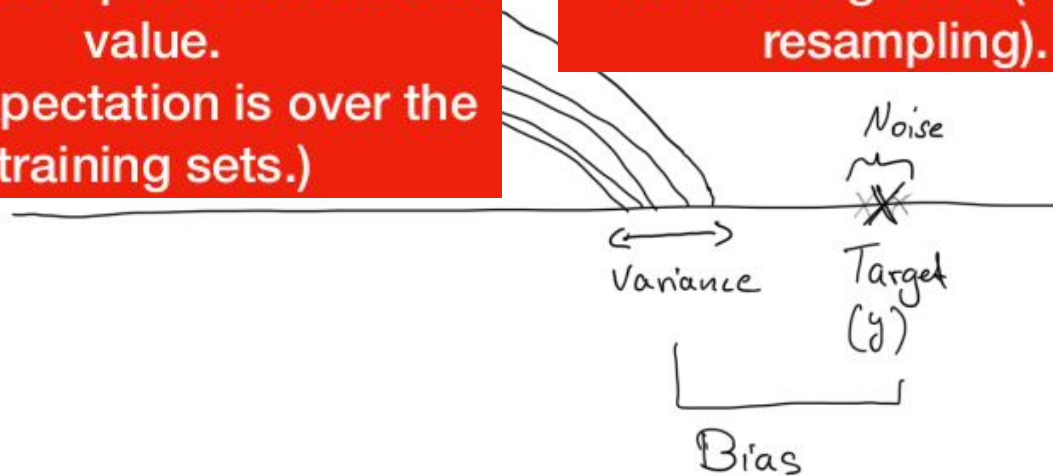
# Terminology

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

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

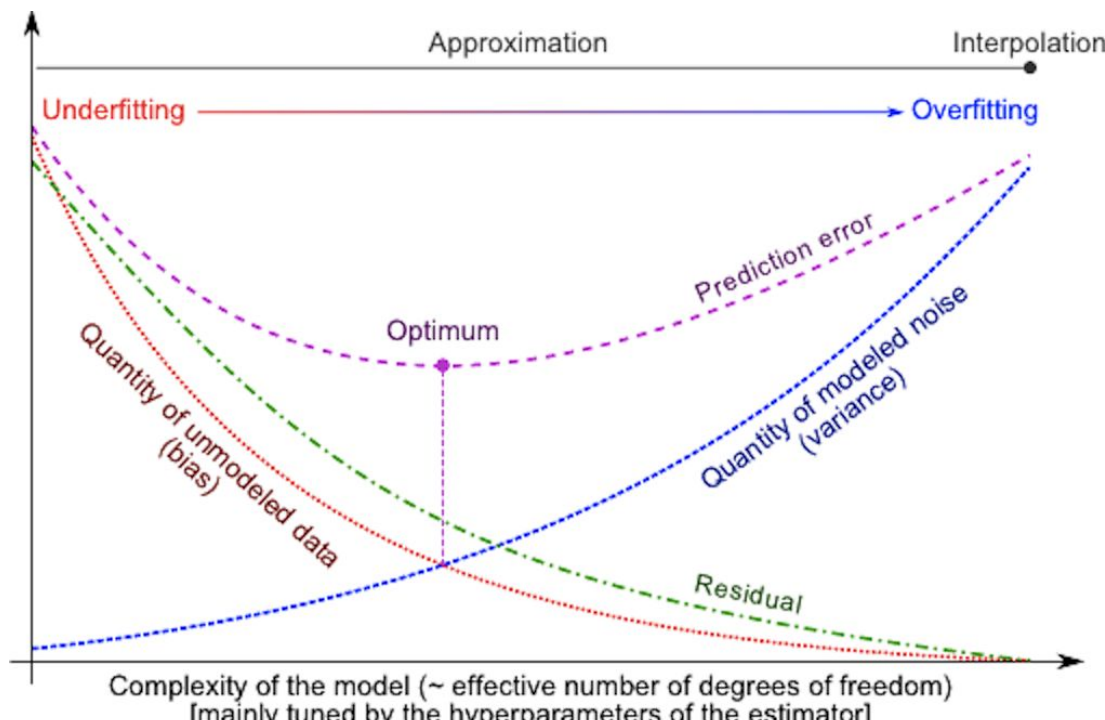
$$\text{Var}[\hat{\theta}] = E[(E[\hat{\theta}] - \hat{\theta})^2]$$

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



# Bias-Variance Decomposition

$$\text{Loss} = \text{Bias} + \text{Variance} + \text{Noise}$$

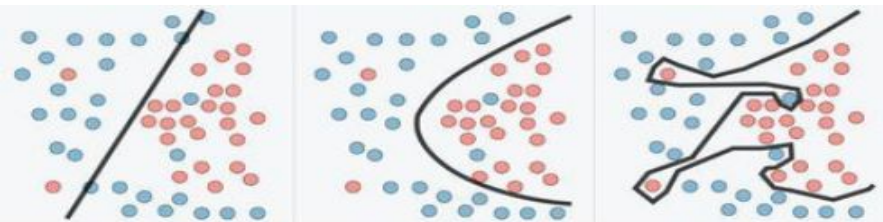
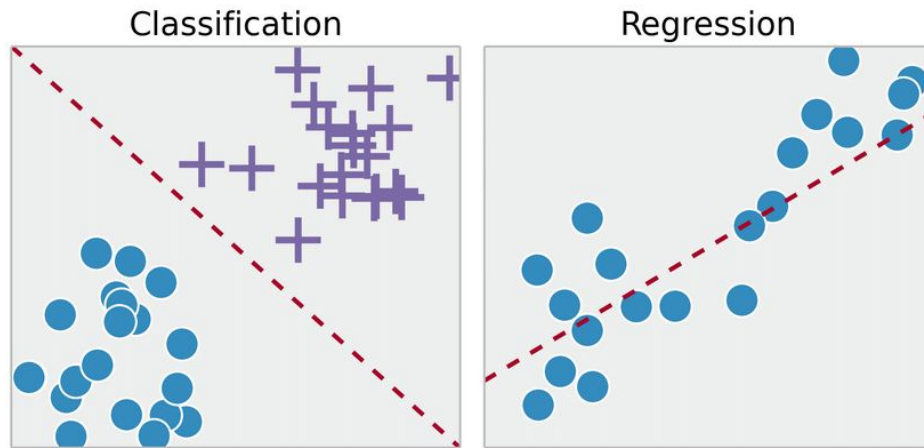


# Lecturer Overview

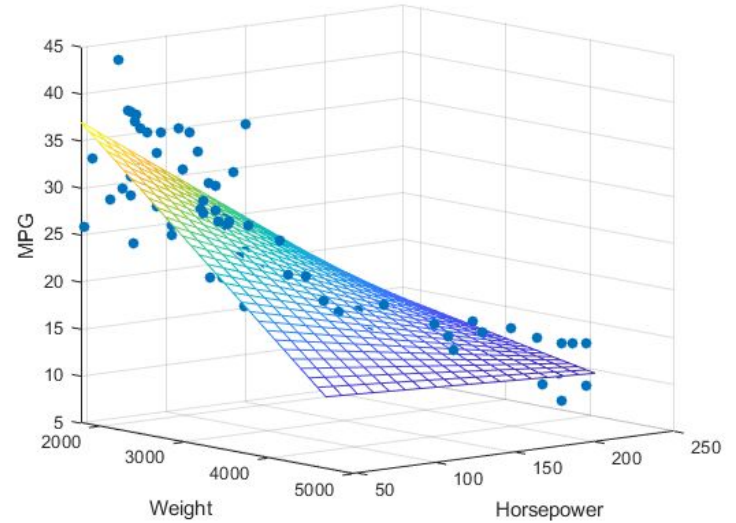
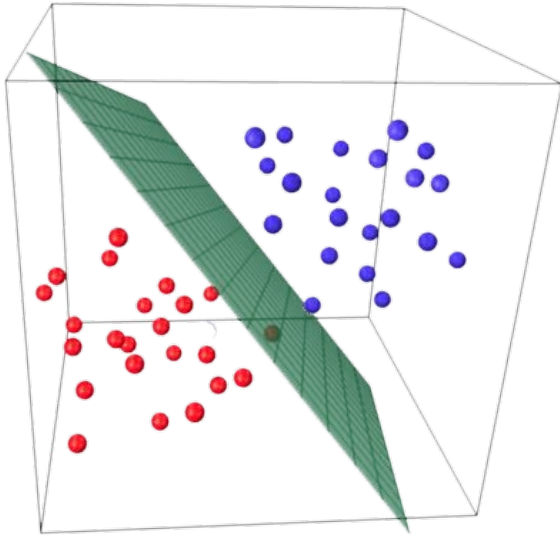
- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- **Error (loss) in classification and regression problems**
- Bias-Variance Decomposition of the Squared Error
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the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



# Classification x regression



# Classification x regression



# 0-1 loss in classification

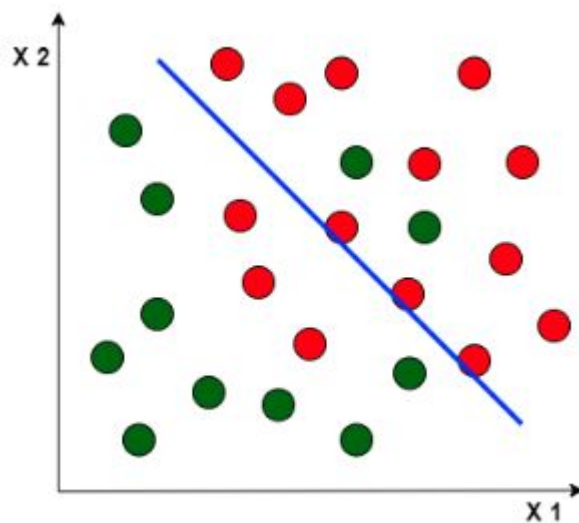
accuracy = 1-error rate

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

- $0.8 = 1 - 0.2$

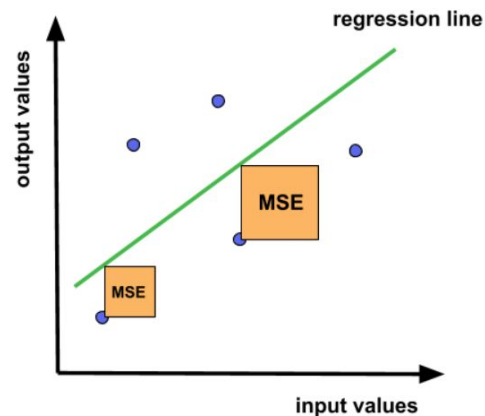
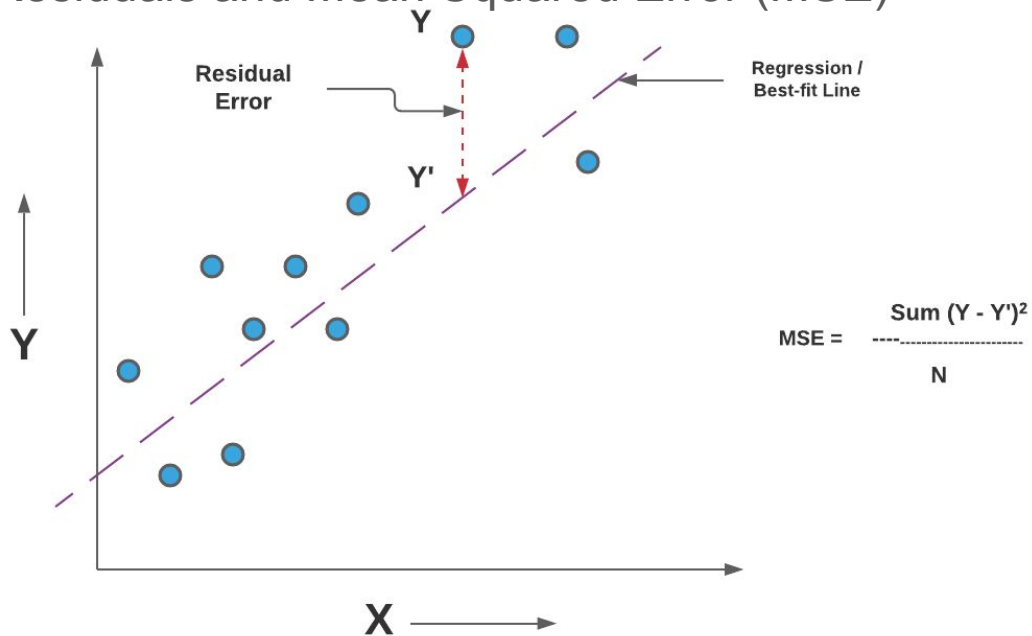
0-1 loss

- $L_{0-1} = 5$

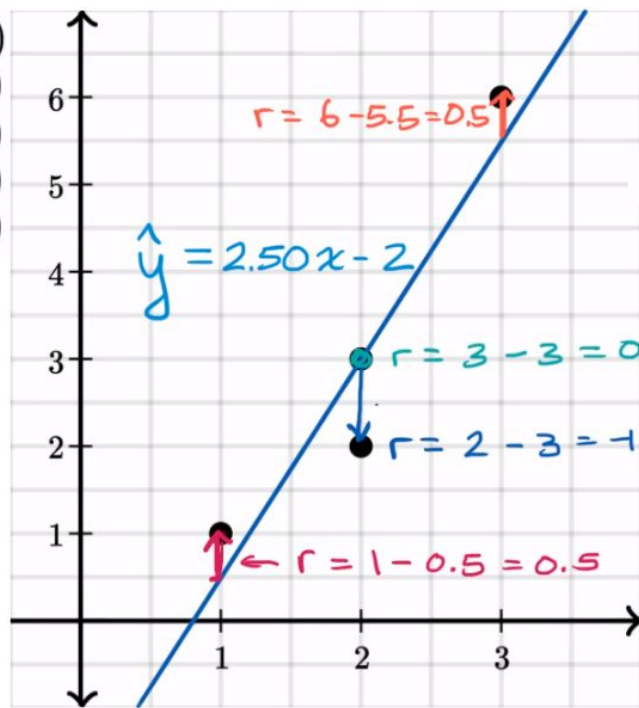
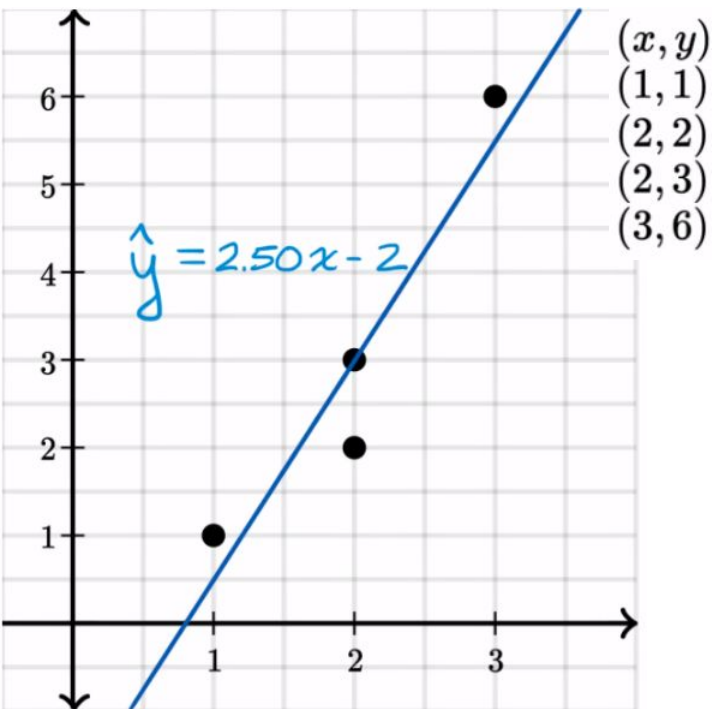


# MSE loss in regression

## Residuals and Mean Squared Error (MSE)



# MSE loss in regression



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

$$= (0.5)^2 + (-1)^2 + (0)^2 + (0.5)^2$$
$$= 1.5 / 4 = 0.375$$

# Let's code!



## 7.2.1. Boston house prices dataset

### Data Set Characteristics:

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

average prediction over the training sets

# Terminology

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

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

spread out from the average prediction

Intuition

prediction

(we ignore noise in this lecture for simplicity)

$E$

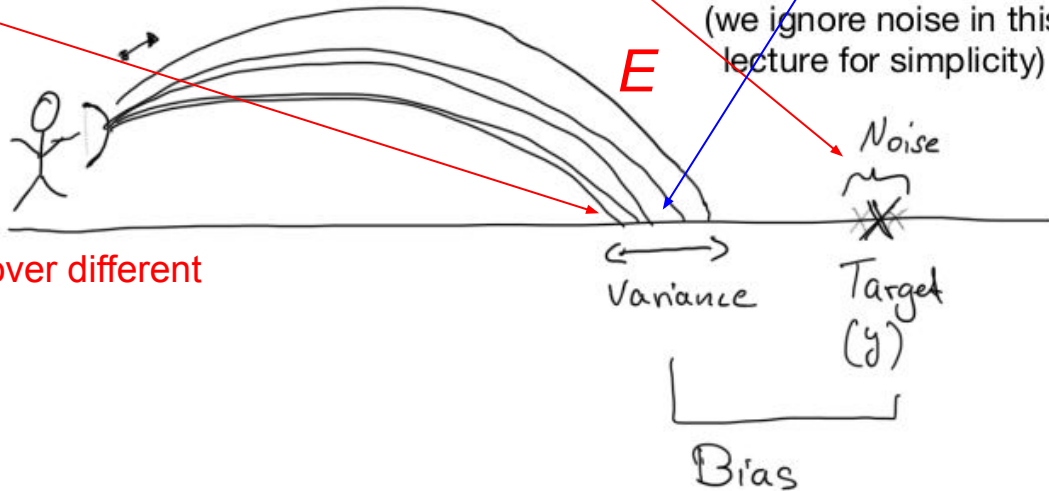
Noise

Target  
( $y$ )

Variance

Bias

models trained over different training sets





# Bias-Variance of the Squared Error

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

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

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

**"ML Notation" for  
Squared Error Loss**

Loss = Bias + Variance + Noise

$y = f(x)$  target

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

for simplicity, we ignore  
the noise term

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

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

## "ML Notation" for Squared Error Loss

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

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

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

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

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

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

# Bias-Variance of the Squared Error

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

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

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

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

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

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

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

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

# Bias-Variance of the Squared Error

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

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

???

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

Let's code!

