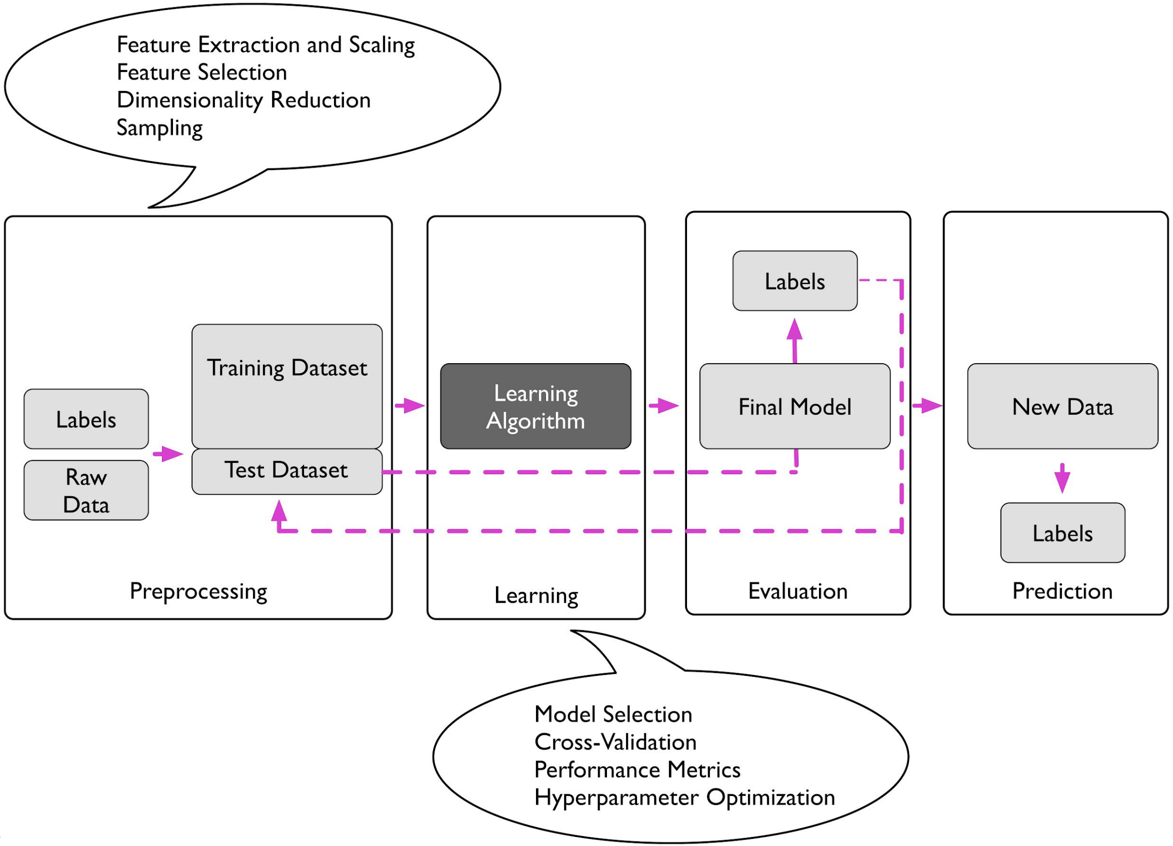
Lecture 14

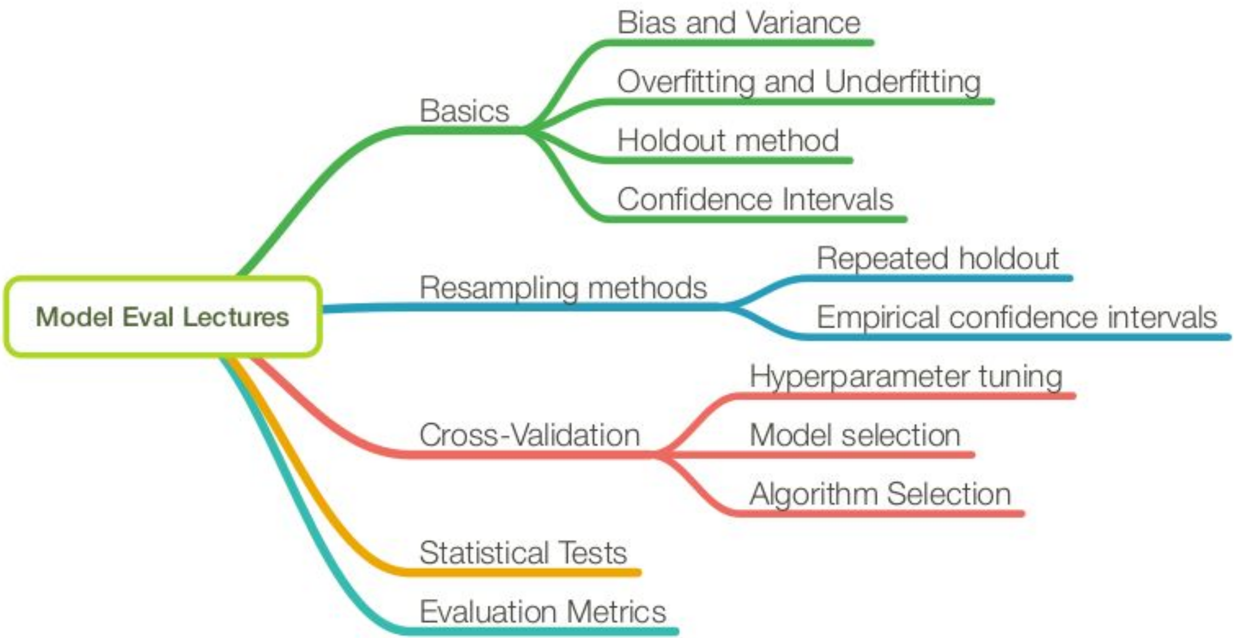
Model evaluation

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

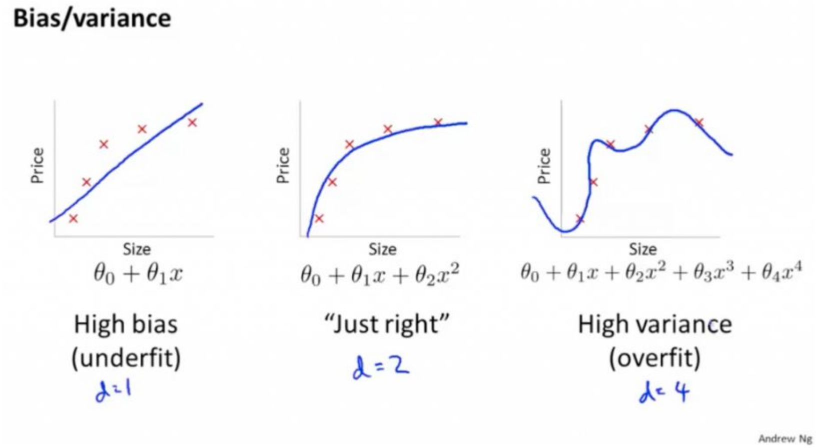
The training error or accuracy provides an **optimistically biased estimate** of the generalization performance

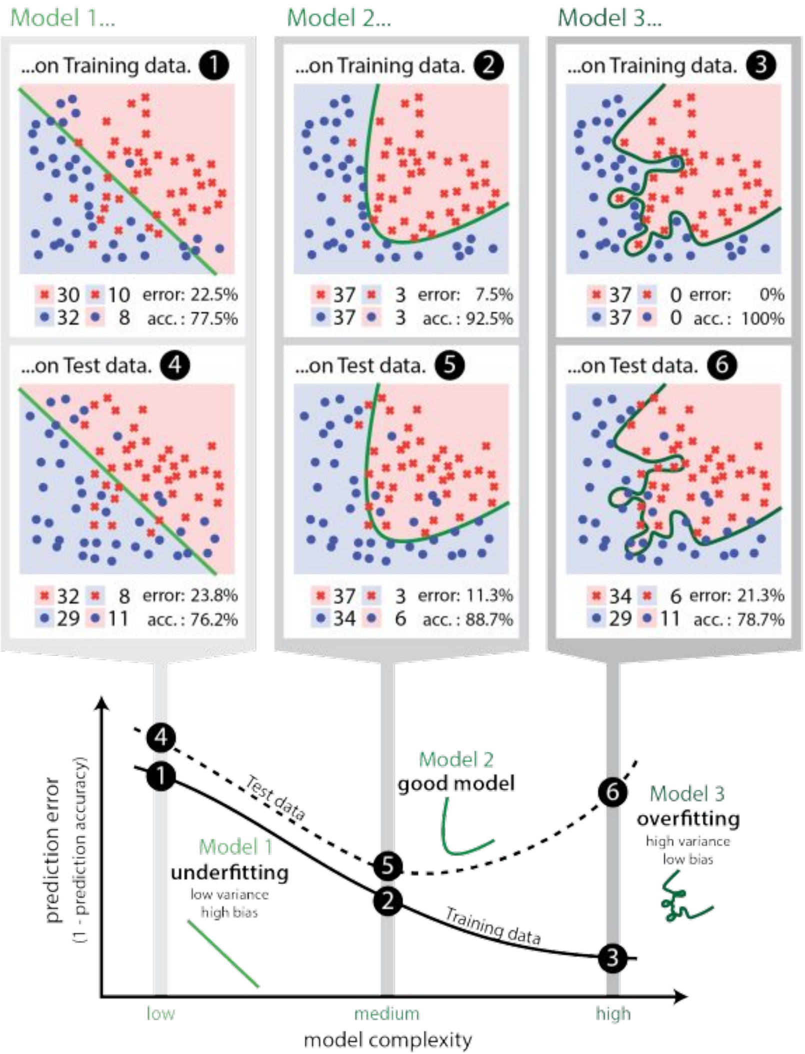
Model Capacity

Underfitting: both the training and test error are high

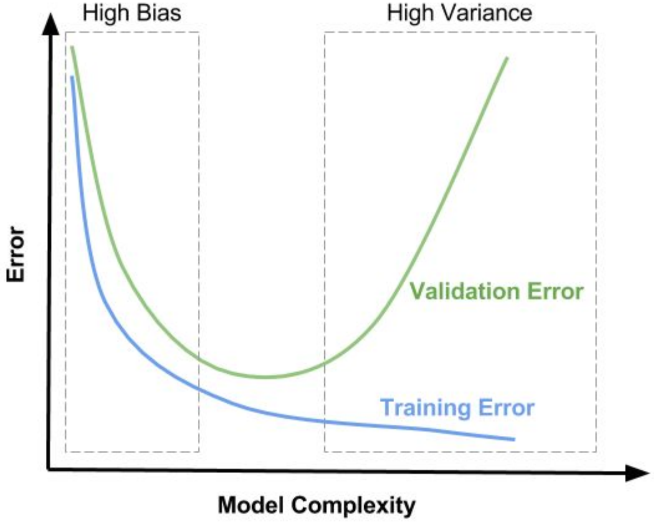
Overfitting: gap between training and test error (where test error is larger)

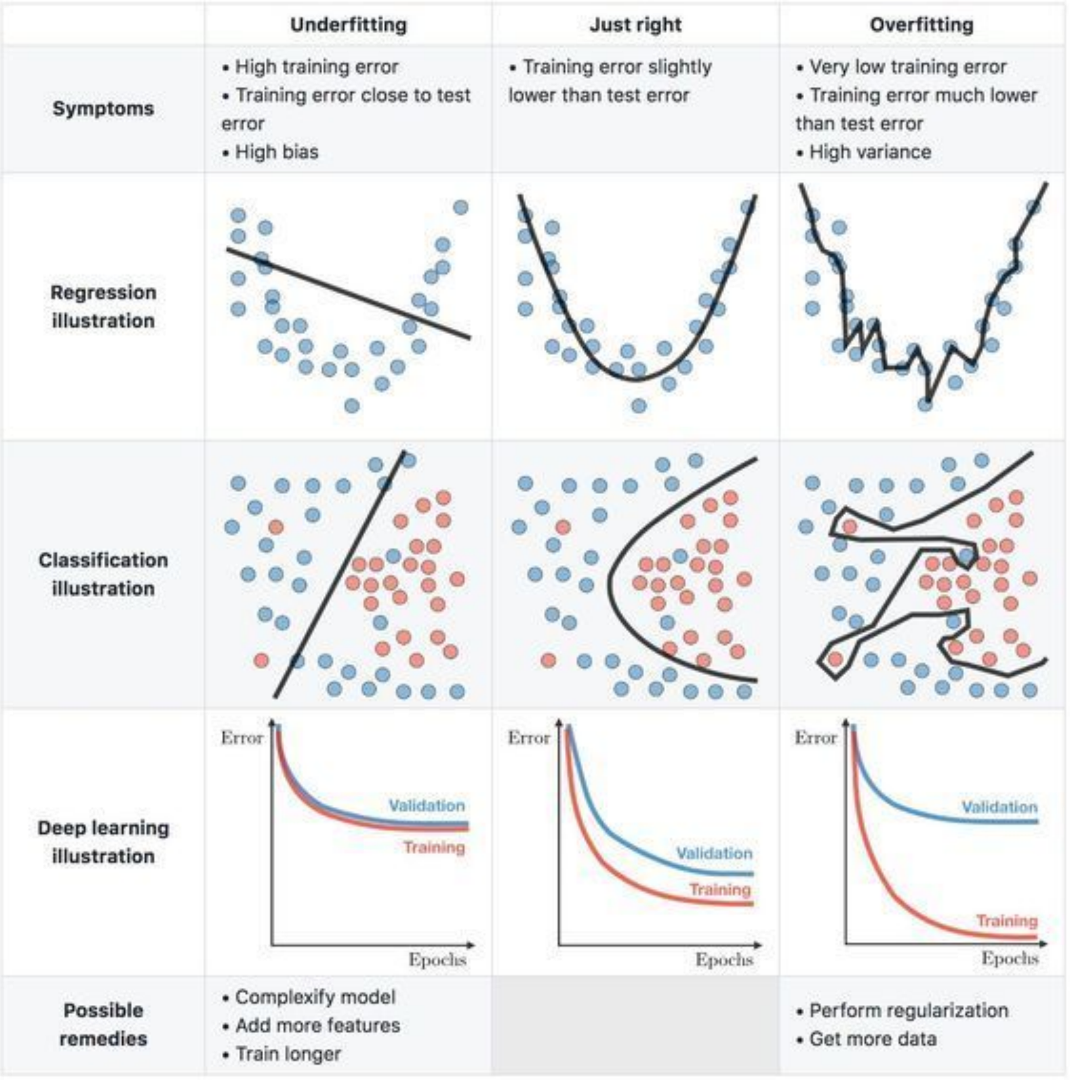
- Large hypothesis space being searched by a learning algorithm -> high tendency to overfit





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

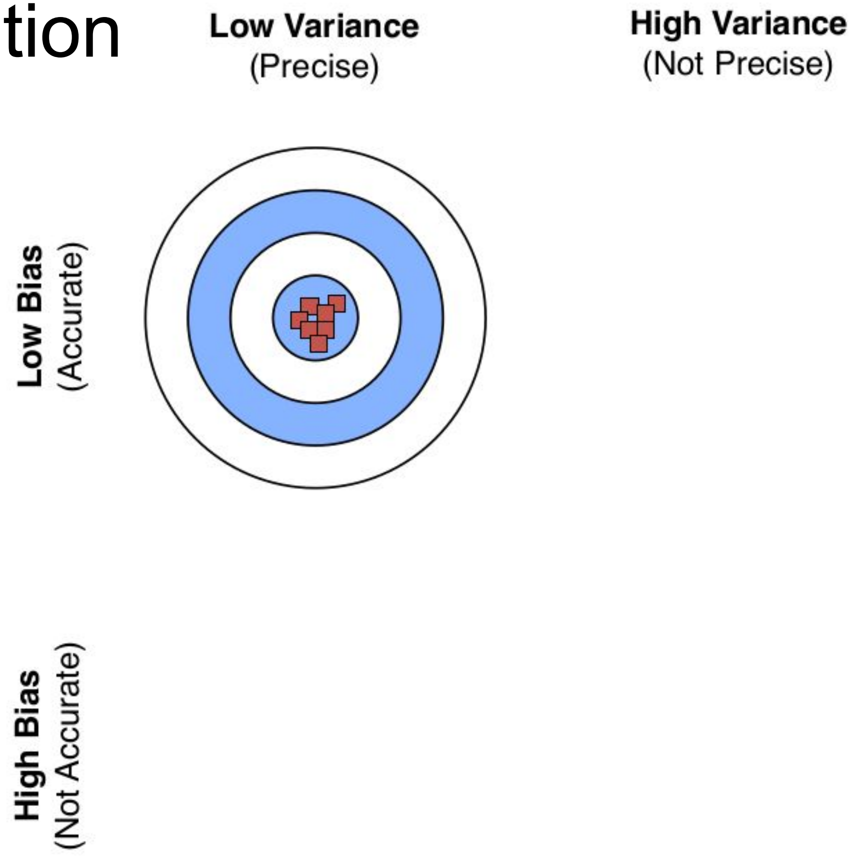
- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- **Error(loss) in classification and regression problems**
- Bias-Variance Decomposition of the Squared Error
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the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

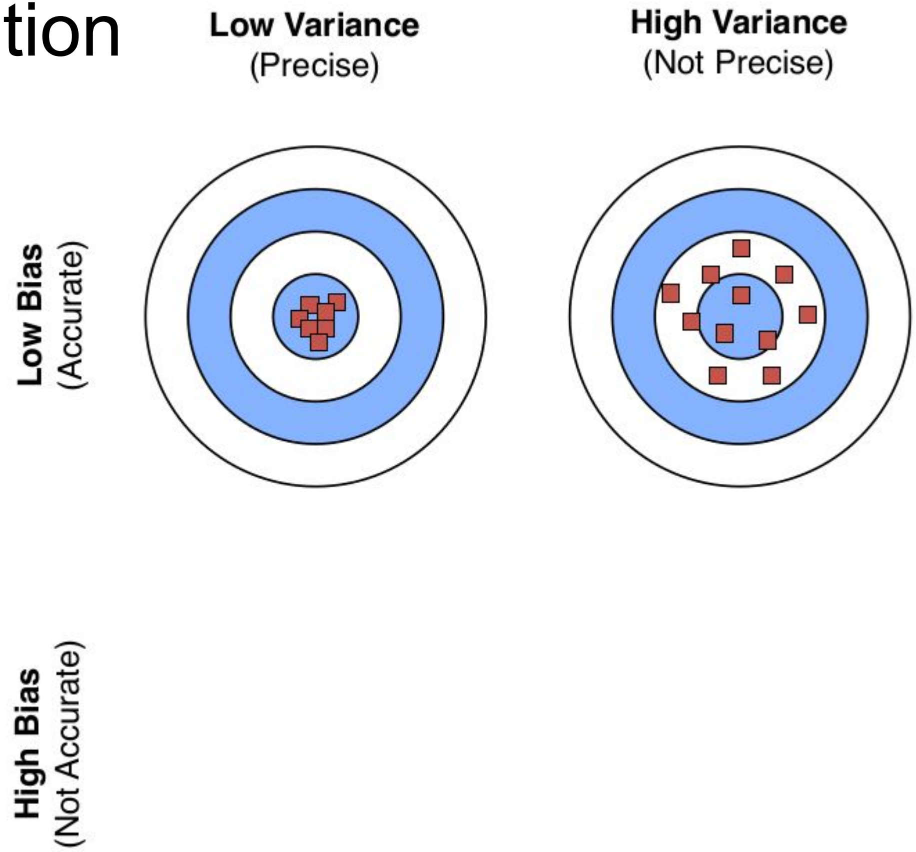
Bias-Variance Decomposition

Loss = Bias + Variance + Noise



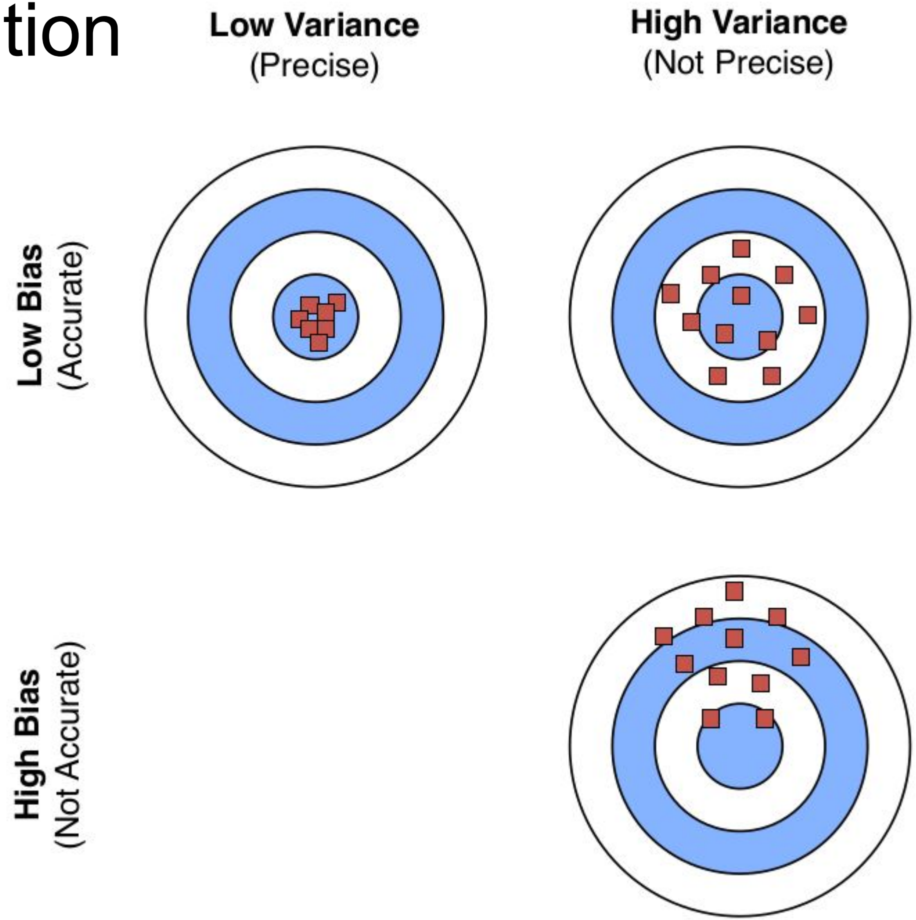
Bias-Variance Decomposition

Loss = Bias + Variance + Noise



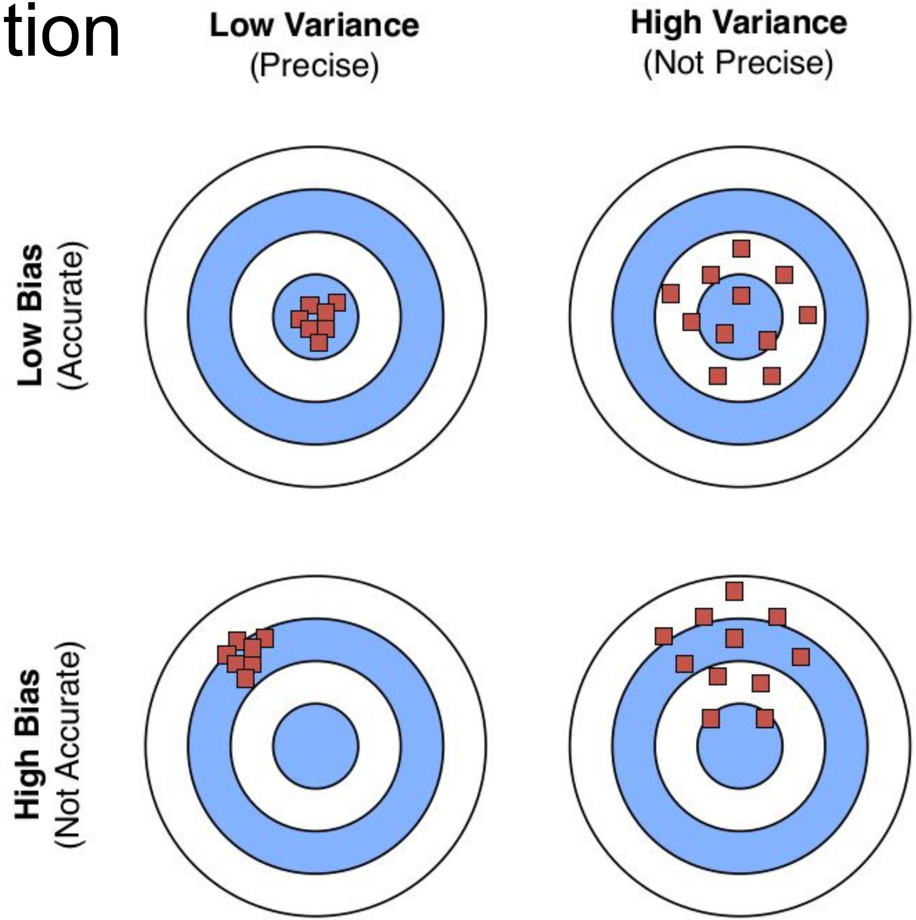
Bias-Variance Decomposition

Loss = Bias + Variance + Noise

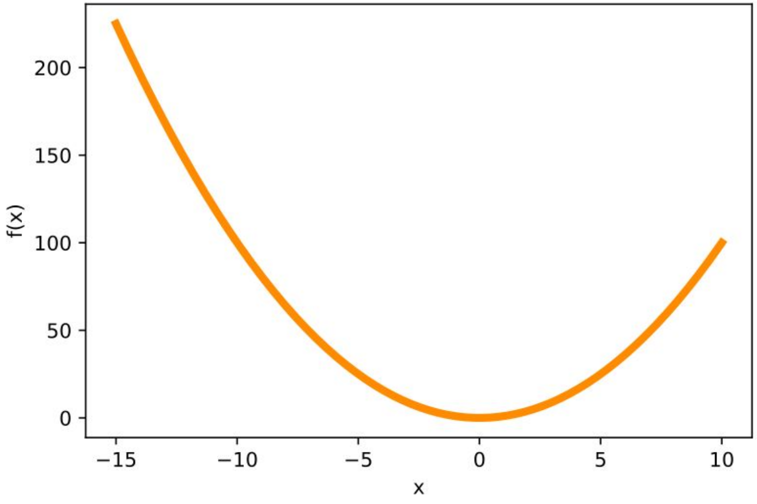


Bias-Variance Decomposition

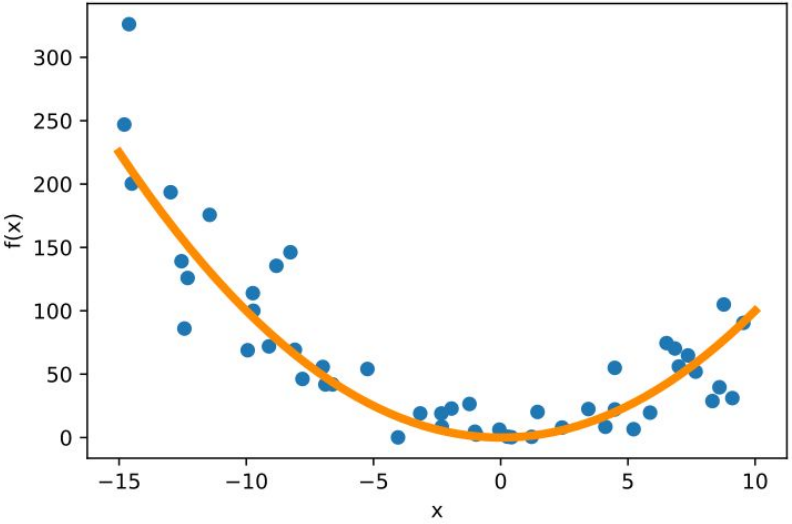
Loss = Bias + Variance + Noise



Bias and Variance Intuition



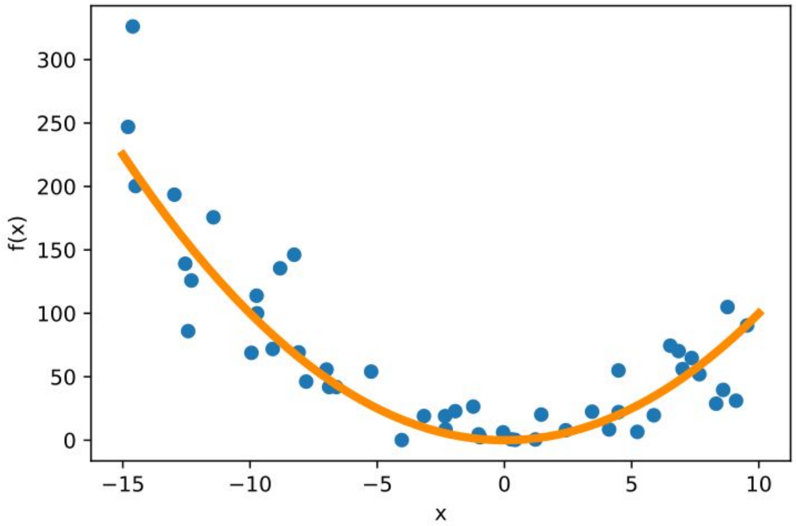
where f(x) is some true (target) function



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

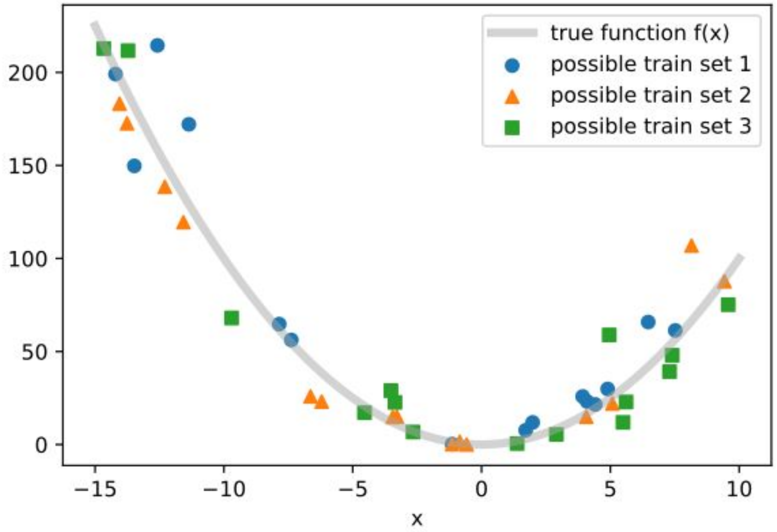
the blue dots are a training dataset;
here, I added some random Gaussian noise

Bias and Variance Intuition



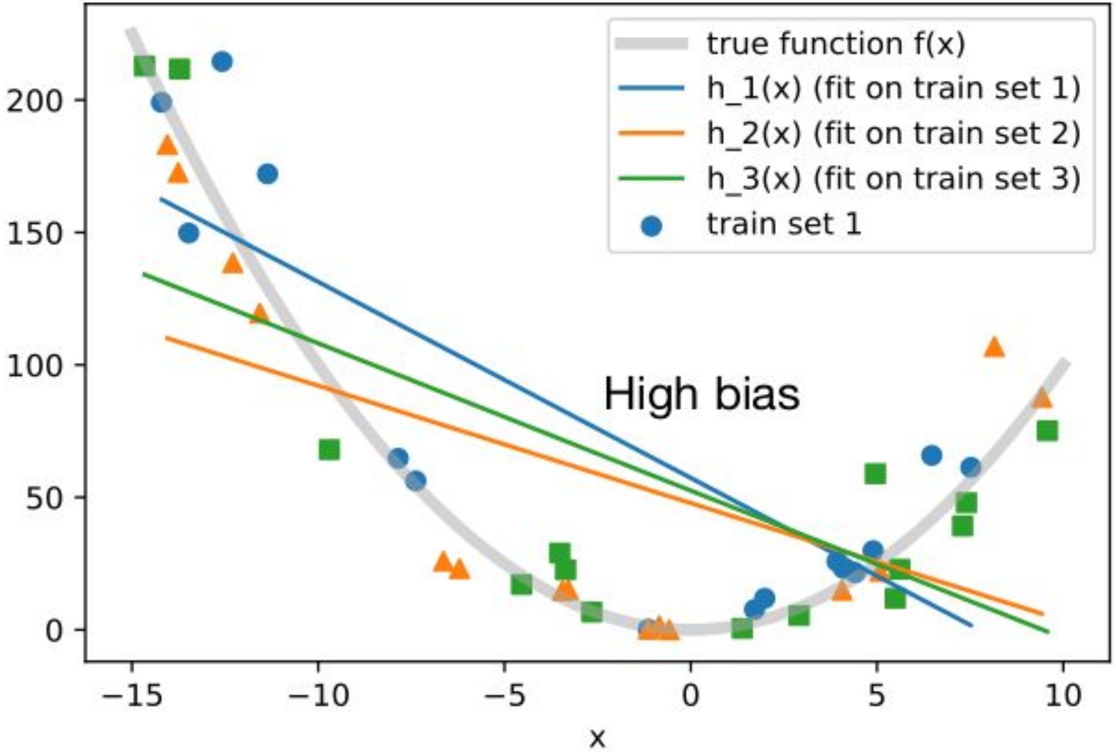
where f(x) is some true (target) function

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



Bias and Variance Intuition

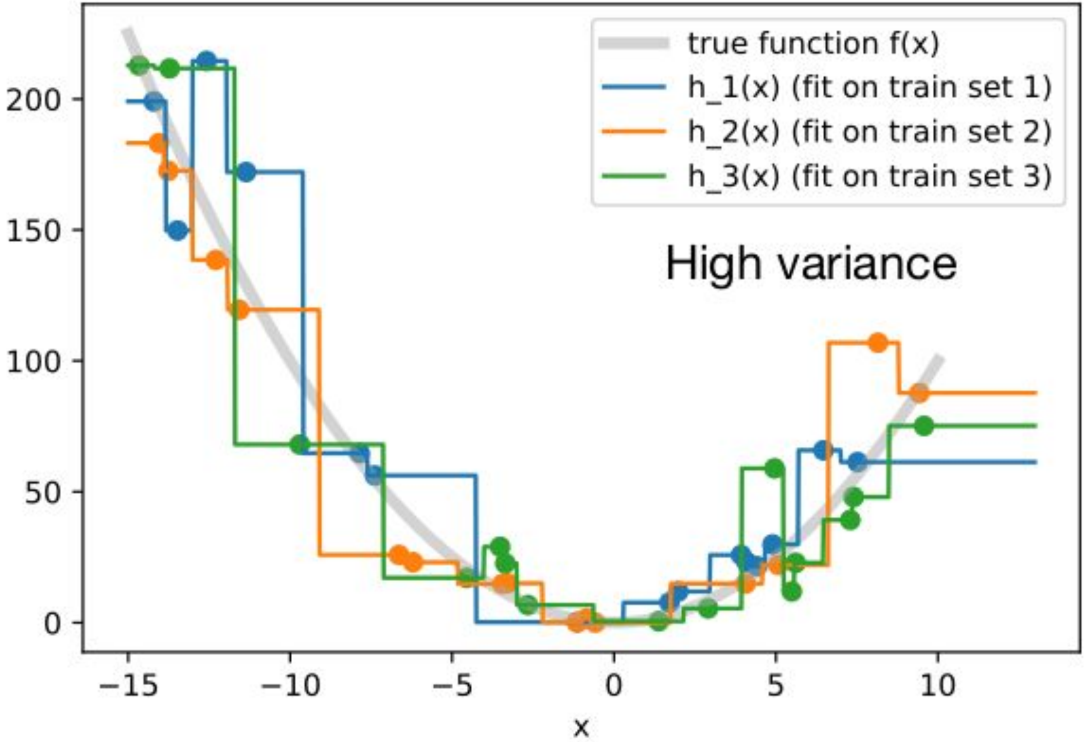
Suppose we have multiple training sets



linear regression models

Bias and Variance Intuition

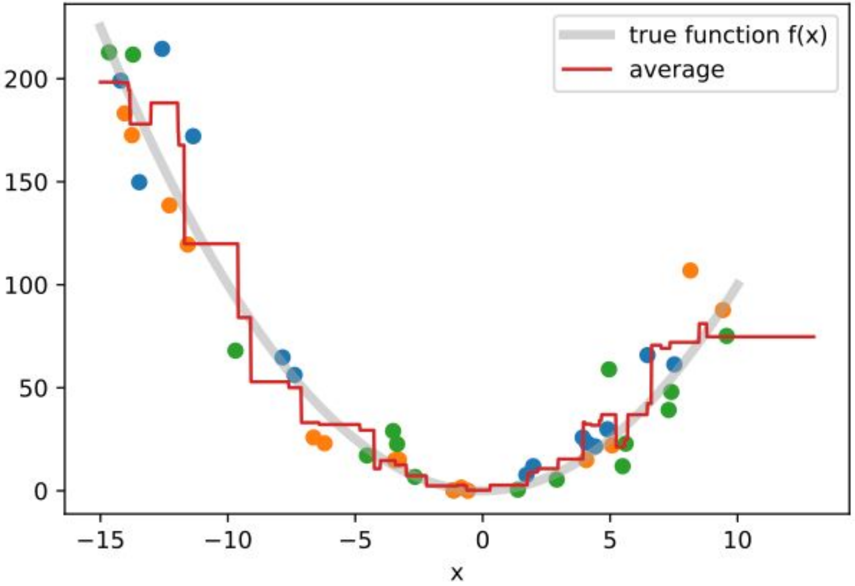
Suppose we have multiple training sets

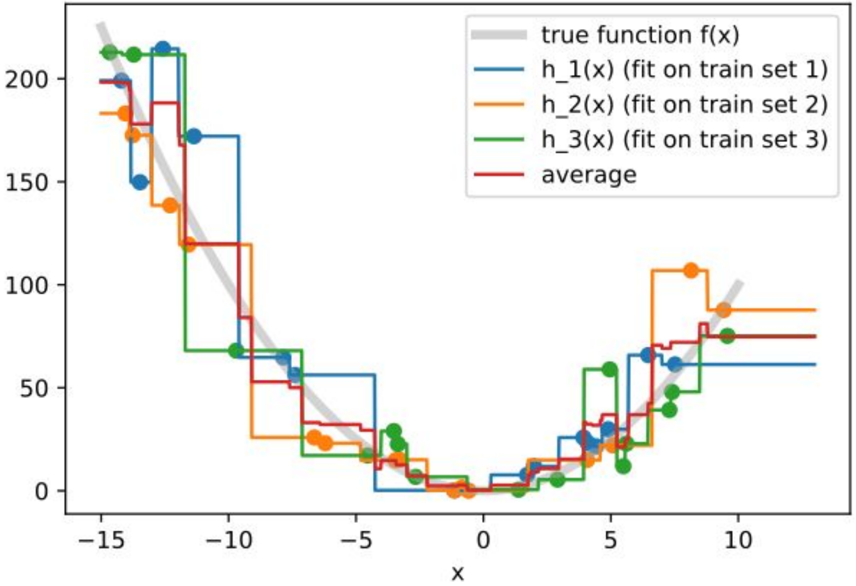


Bias and Variance Intuition

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

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

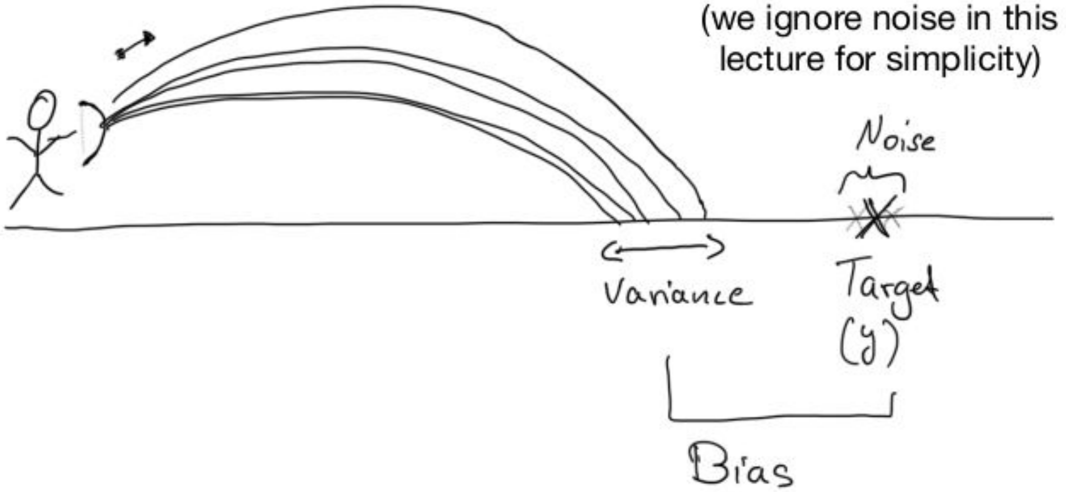
$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

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

Terminology

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

Intuition



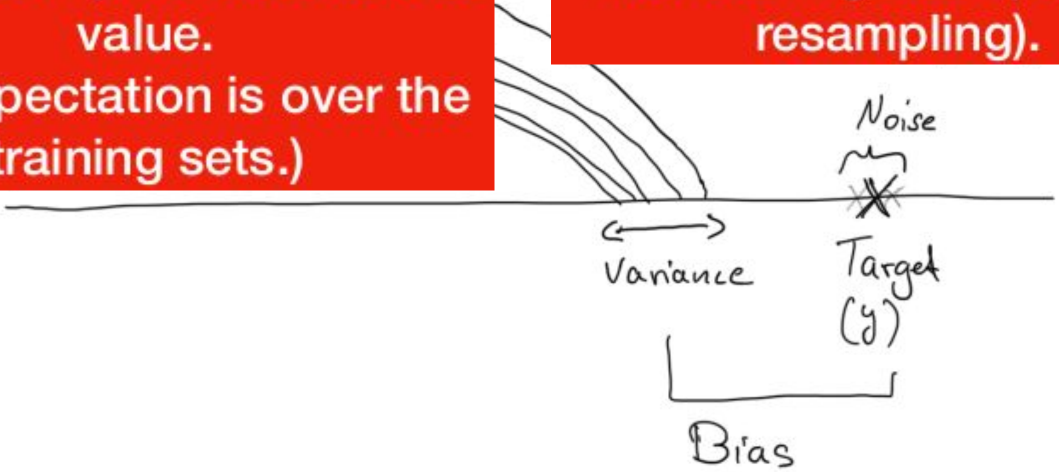
Terminology

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

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

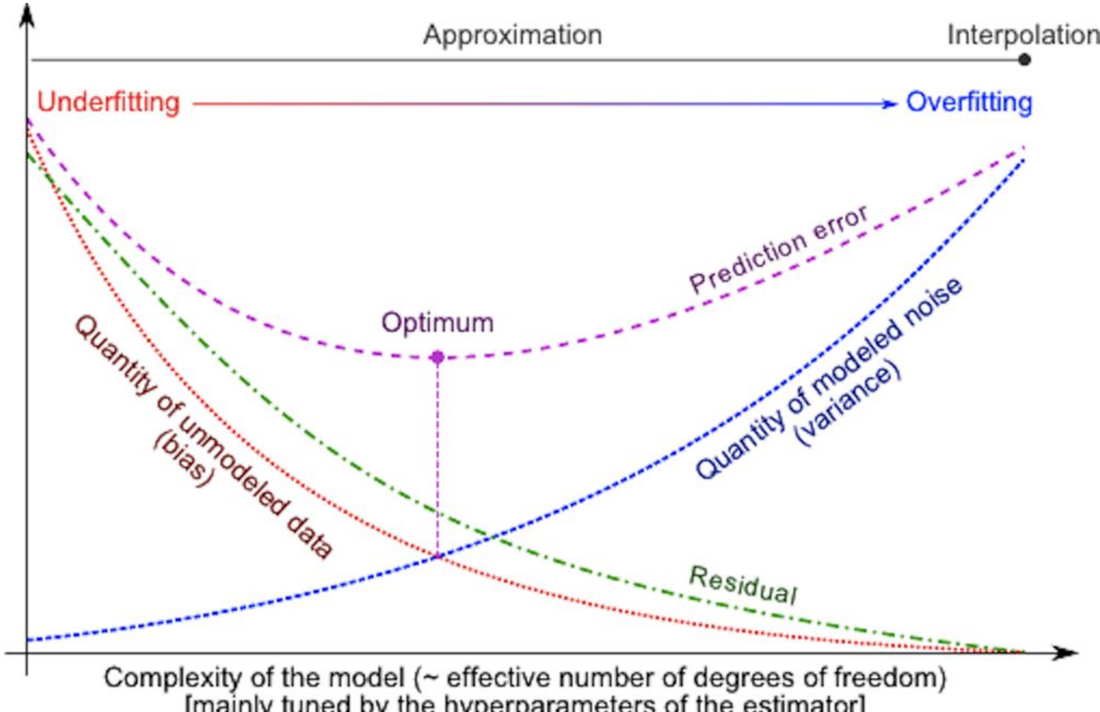
$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

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



Bias-Variance Decomposition

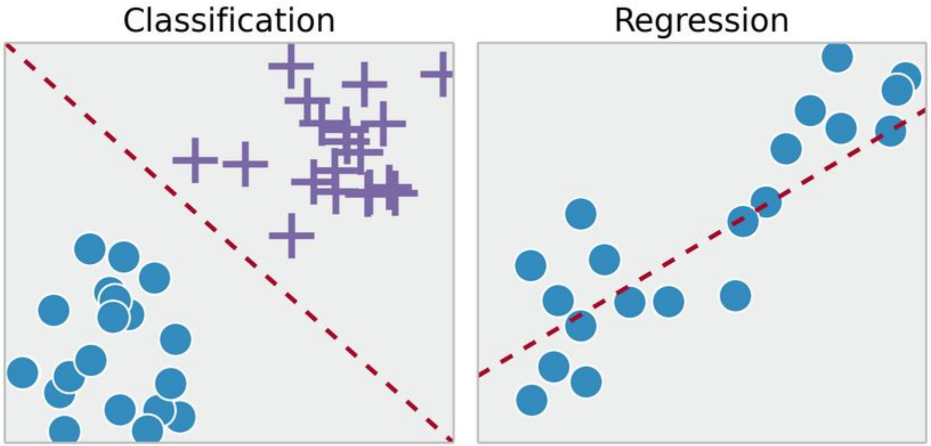
Loss = Bias + Variance + Noise

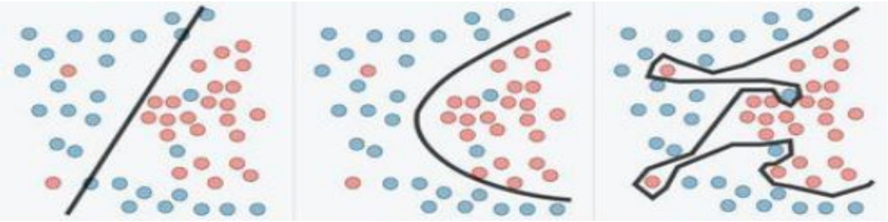


Lecturer Overview

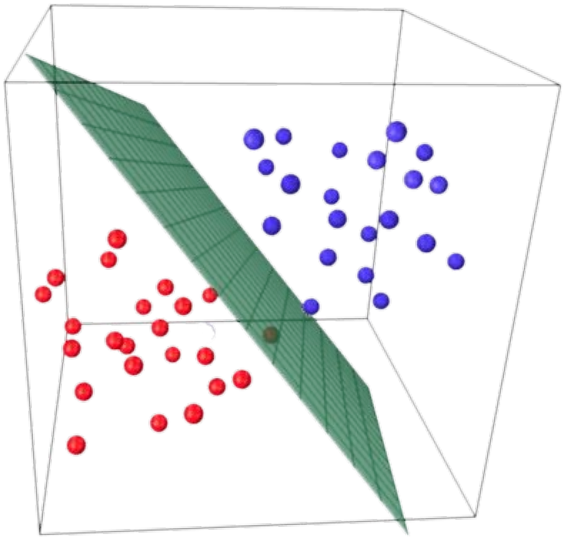
- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- **Error (loss) in classification and regression problems**
- Bias-Variance Decomposition of the Squared Error
- Relationship between Bias-Variance Decomposition and overfitting and underfitting
- Bias-Variance Decomposition of the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

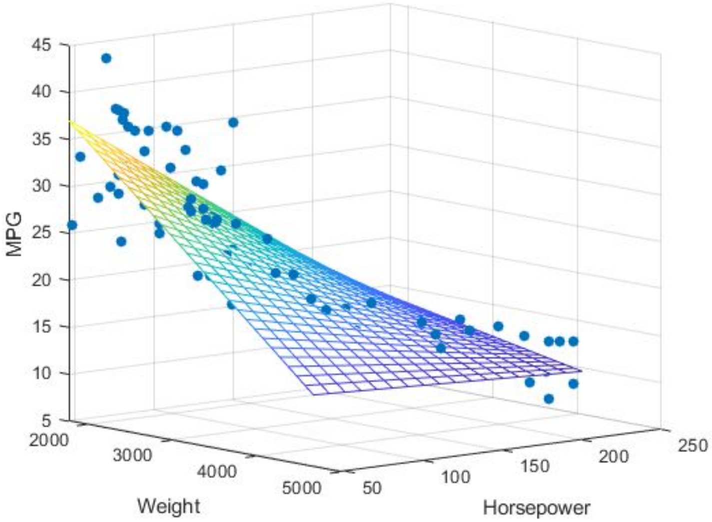
Classification x regression





Classification x regression





0-1 loss in classification

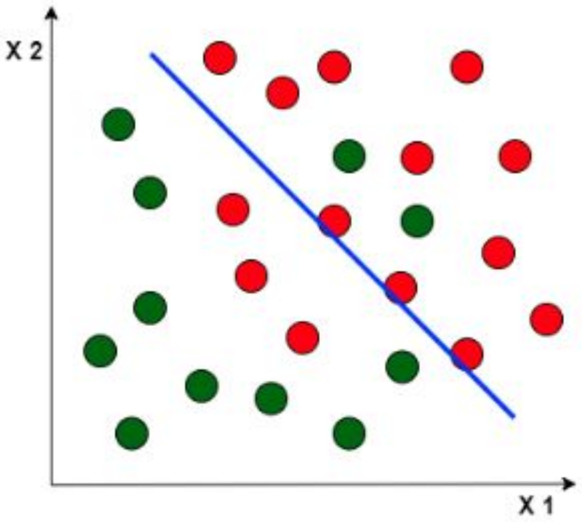
accuracy = 1-error rate

- 0.8 = 1 - 0.2

0-1 loss

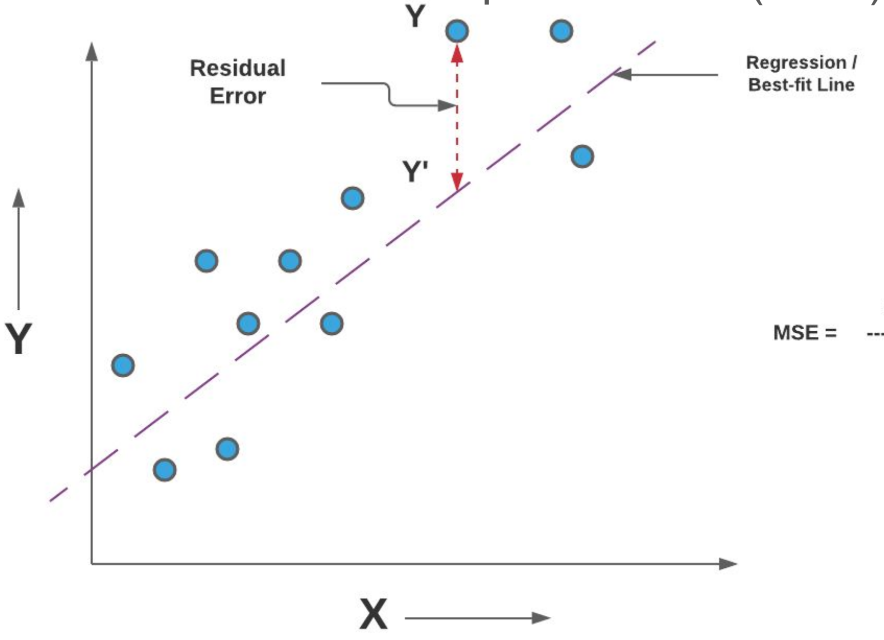
- $L_{0\text{-}1} = 5$

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

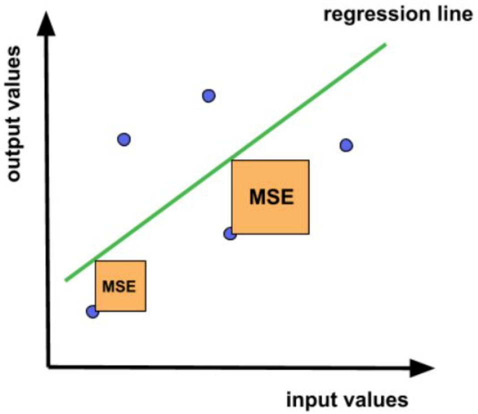


MSE loss in regression

Residuals and Mean Squared Error (MSE)

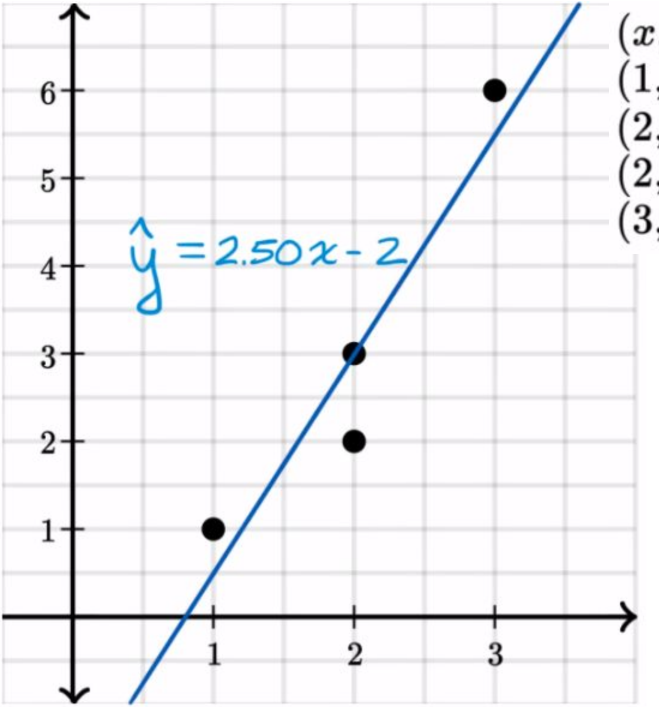


$$\text{MSE} = \frac{\text{Sum } (Y - Y')^2}{N}$$

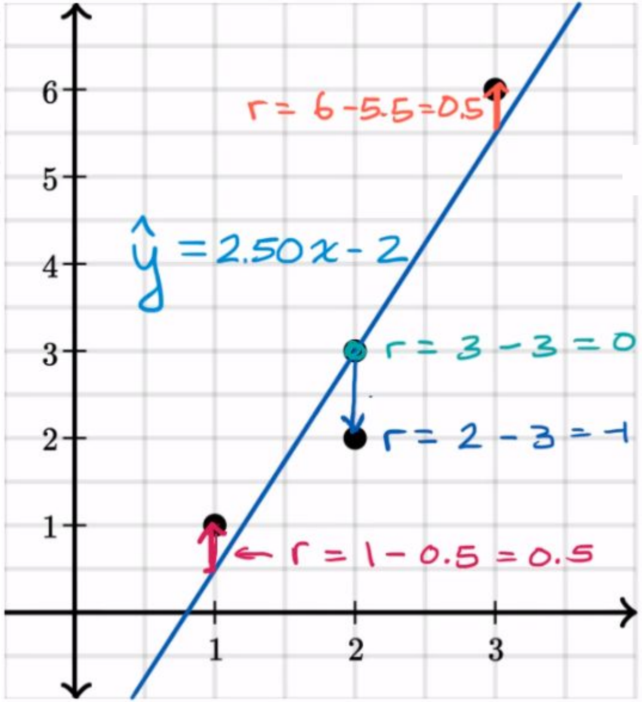


MSE loss in regression

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



(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

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

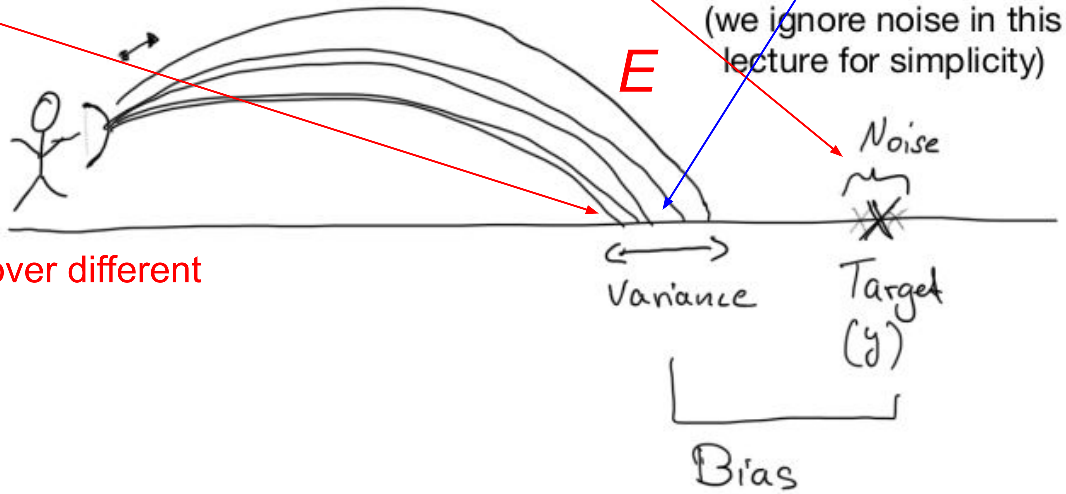
$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

spread out from the average prediction

prediction

Intuition

(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

"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

$(a-b)^2 = a^2 - 2ab + b^2$
$= a^2 + b^2 - 2ab$

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

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

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

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

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

???

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

Let’s code!

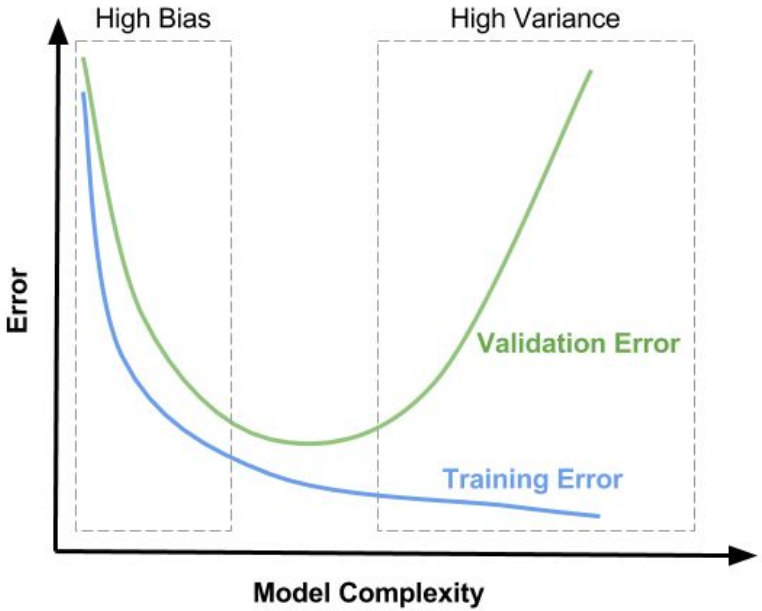
Lecturer Overview

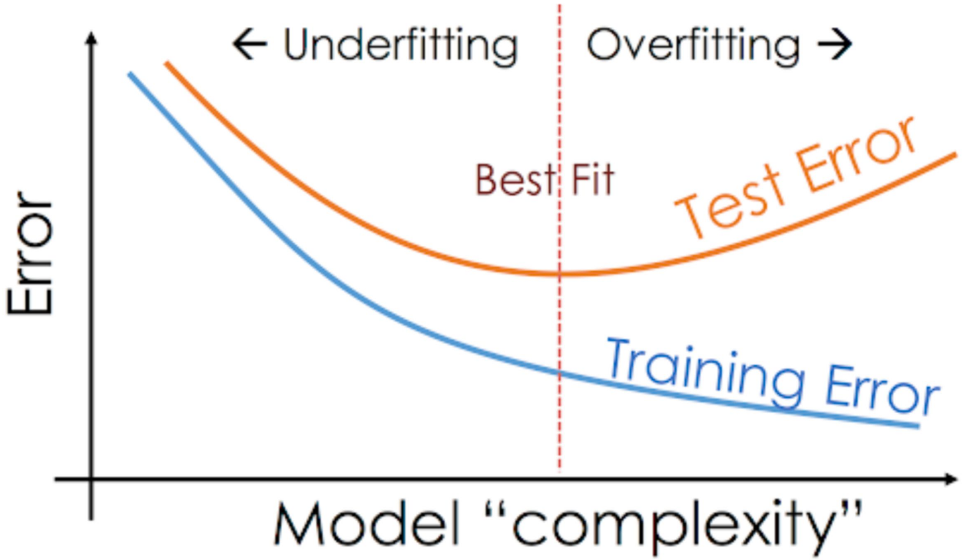
- Overfitting and Underfitting
- Intro to Bias-Variance Decomposition
- Error (loss) in classification and regression problems
- Bias-Variance Decomposition of the Squared Error
- **Relationship between Bias-Variance Decomposition and overfitting and underfitting**
- Bias-Variance Decomposition of the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

How is this related to overfitting and underfitting?

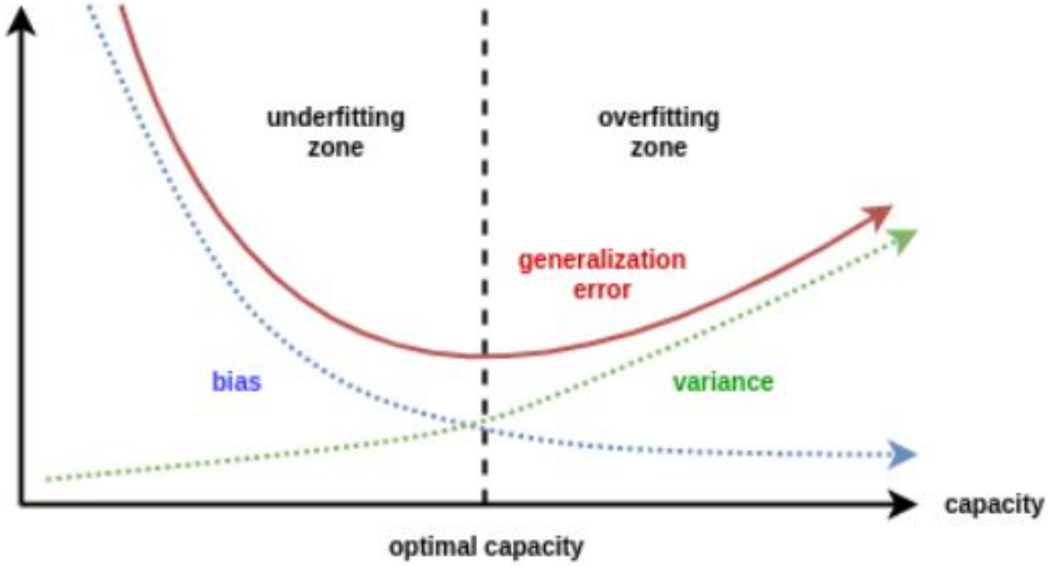
$$E[(y-\hat{y})^2] = (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2]$$

Bias² + Variance



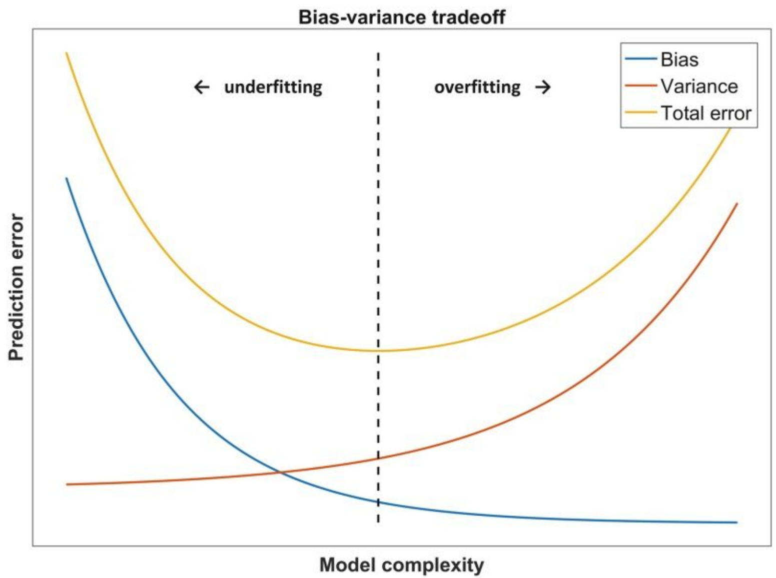


How is this related to overfitting and underfitting?



Minimize 2 error sources!

bias-variance tradeoff...



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 of 0-1 Loss

Domingos, P. (2000). *A unified bias-variance decomposition.*
In Proceedings of 17th International Conference on Machine Learning (pp. 231-238).

"several authors have proposed bias-variance decompositions related to zero-one loss (Kong & Dietterich, 1995; Breiman, 1996b; Kohavi & Wolpert, 1996; Tibshirani, 1996; Friedman, 1997). However, each of these decompositions has significant shortcomings."

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238).

Bias-Variance Decomposition of 0-1 Loss

Squared Loss

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

Expectation over trainings sets to a particular sample

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

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

Bias^2 + Variance

Main prediction -> Mean

Bias^2: $(y - \boxed{E[\hat{y}]})^2$

Variance: $E[(E[\hat{y}] - \hat{y})^2]$

0-1 Loss

$$L(y, \hat{y})$$

$$E[L(y, \hat{y})]$$

Main prediction -> Mode

$$L(y, \boxed{E[\hat{y}]})$$

$$E[L(\hat{y}, E[\hat{y}])]$$

Bias-Variance Decomposition of 0-1 Loss

0-1 Loss

Loss = $P(\hat{y} \neq y)$

$$Bias = \begin{cases} 1 \textbf{ if } y \neq \bar{\hat{y}} \\ 0 \textbf{ otherwise} \end{cases}$$

Variance can improve loss!!
Why is that so?

Loss = $P(\hat{y} \neq y) = 1 - P(\hat{y} = y) = 1 - P(\hat{y} \neq \bar{\hat{y}})$

Loss = Bias - Variance

Domingos, P. (2000). A unified bias-variance decomposition
17th International Conference on Machine Learning (pp. 23[illegible]).

includes noise
and more general: Loss = Bias + *c* Variance

or more precisely $c_1 N(x) + B(x) + c_2 V(x)$

where, e.g., $c_1 = c_2 = 1$ for squared loss

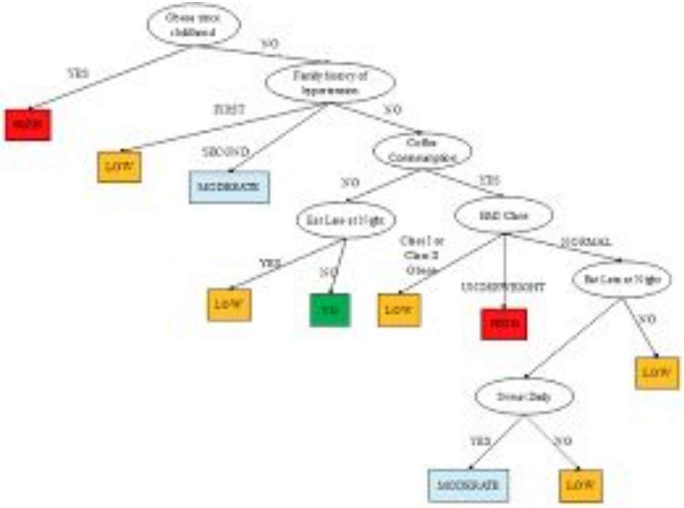
Lecturer Overview

- 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**

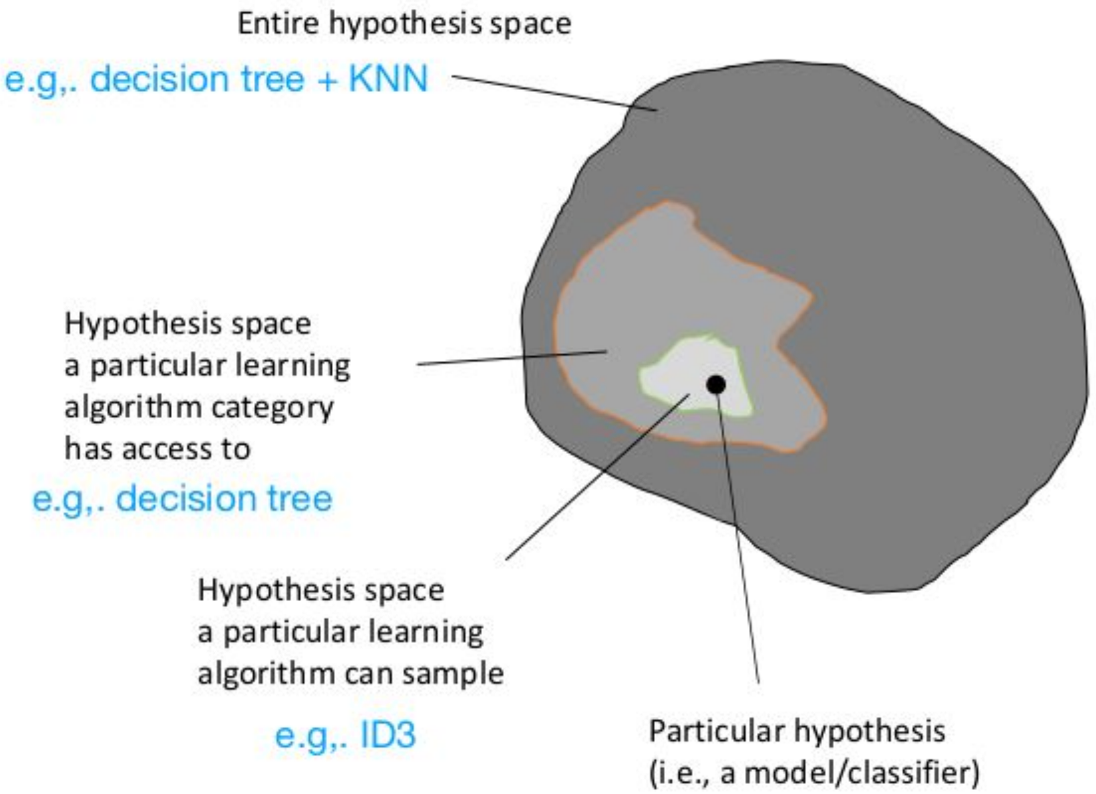
Statistical Bias vs "Machine Learning Bias"

"Machine learning bias" sometimes also called "inductive bias"

- e.g., decision tree algorithms consider small trees before they consider large trees
  - (if training data can be classified by small tree, large trees are not considered)



Hypothesis Space



Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

Table 1: Relationship between ML bias and statistical bias and variance

ML Bias		Statistical	
Absolute	Relative	Bias	Variance
appropriate	too strong	high	low
appropriate	ok	low	low
appropriate	too weak	low	high
inappropriate	too strong	high	low
inappropriate	ok	high	moderate
inappropriate	too weak	high	high

e.g. classify time series with knn
e.g. use a bernoulli NB on gaussian data

e.g. DT stump
e.g. DT unpruned

bias can be characterized as appropriate or inappropriate. The hypothesis space of an inappropriate absolute bias does not contain any good approximations to the target function. An appropriate bias does contain good approximations.

A relative bias can be described as being too strong or too weak. A bias that is too strong is one that, though it may not rule out good approximations to the target function, prefers other, poorer hypotheses instead. A bias that is too weak does not focus the learning algorithm on the appropriate hypotheses but instead allows it to consider too many hypotheses.

Bias-Variance Simulation of C 4.5

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University.

- simulation on 200 training sets with 200 examples each (0-1 labels)
  - 200 hypotheses

- test set: 22,801 examples (1 data point for each grid point)

- mean error rate is 536 errors (out of the 22,801 test examples)
  - 297 as a result of bias
  - 239 as a result of variance

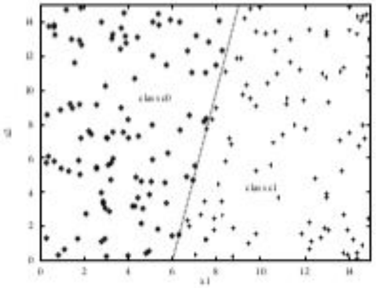


Figure 1: A two-class problem with 200 training examples.

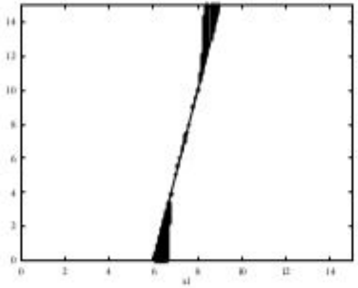
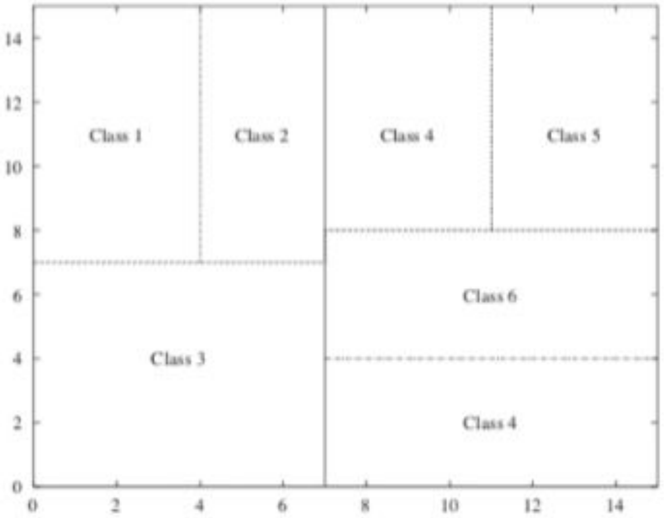


Figure 2: Bias errors of C4.5 on the problem from Figure 1.

(remember that trees use a "staircase" to approximate diagonal boundaries)

Bias-Variance Simulation of C 4.5

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.* Technical report, Department of Computer Science, Oregon State University.



errors due to bias: 0
errors due to variance: 17

ML Bias		Statistical	
Absolute	Relative	Bias	Variance
appropriate	too strong	high	low
appropriate	ok	low	low
appropriate	too weak	low	high
inappropriate	too strong	high	low
inappropriate	ok	high	moderate
inappropriate	too weak	high	high

"Fairness" Bias

"The term bias is often used to refer to demographic disparities in algorithmic systems that are objectionable for societal reasons. "

Barocas, S., Hardt, M., & Narayanan, A. Fairness and Machine Learning.
https://fairmlbook.org/introduction.html

Microsoft’s AI Chatbot Tay

With chatbots becoming popular across social networks, Microsoft launched its version for Twitter users in March 2016. Monikered ‘Tay’, it was programmed to have casual conversations in the language of a typical millennial.

According to the company, Tay leveraged AI to learn from these interactions to hold better conversations in the future. However, the Twitter chatbot had to be taken down less than 24 hours post its launch.

Targeting its vulnerabilities, trolls on the microblogging website manipulated Tay into making deeply sexist and racist statements.

Following this debacle, Peter Lee, Microsoft’s corporate VP for AI and research issued a public apology, which stated that the company took “full responsibility for not seeing this possibility ahead of time.”

Amazon’s AI-Powered Recruiting Tool

According to a Reuters report, Amazon had been building machine learning (ML) programs since 2014 to review job applicants’ resumes. But it is well known that AI has a big bias problem and the company demonstrated this with example in 2015 when it realised that its new system was not rating candidates in a gender-neutral way. That is, its ML specialists had taught their own AI to prefer male candidates over female ones.

This happened because these models were trained to verify applicants by tracking patterns in resumes submitted to the company over a 10-year period.

The Seattle company reportedly disbanded the team a few years later after failing to develop or work to resolve that problem.

Facial Recognition Failure In China

Back in November 2018, Chinese police admitted to wrongly shaming a billionaire businesswoman after a facial recognition system designed to catch jaywalkers 'caught' her on an advert on a passing bus.

Traffic police in major Chinese cities deploy smart cameras that use facial recognition techniques to detect jaywalkers, whose names and faces then show up on a public display screen. After this went viral on Chinese social media, a CloudWalk researcher stated that the algorithm's lack of live detection could have been the problem.

Amazon’s Rekognition

In 2018, Members of US Congress rained down on Amazon after its facial recognition software falsely matched 28 congresspeople with mugshots of criminals. In fact, according to the American Civil Liberties Union (ACLU), nearly 40% of the matches were of people of colour, indicating that the technology is racially biased.

Gender prediction

Facial Recognition Is Accurate, if You're a White Guy

Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

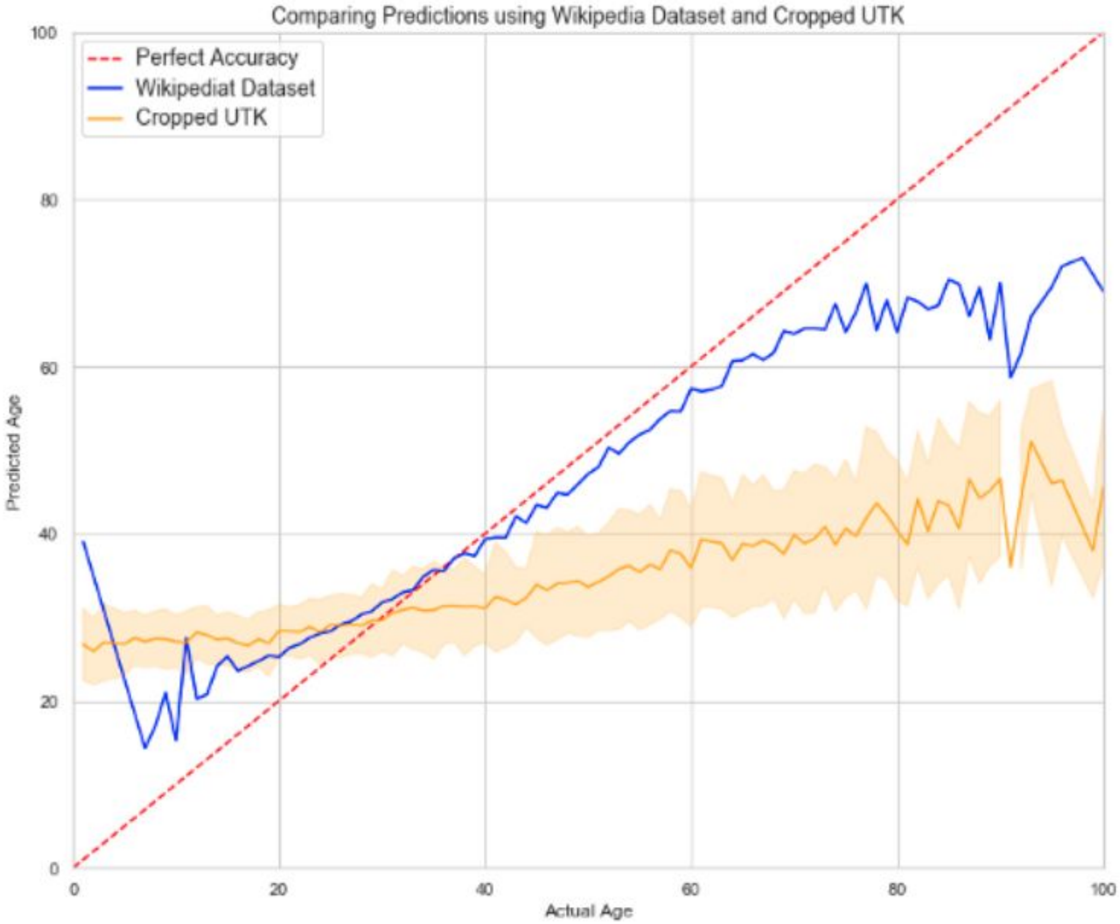
Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

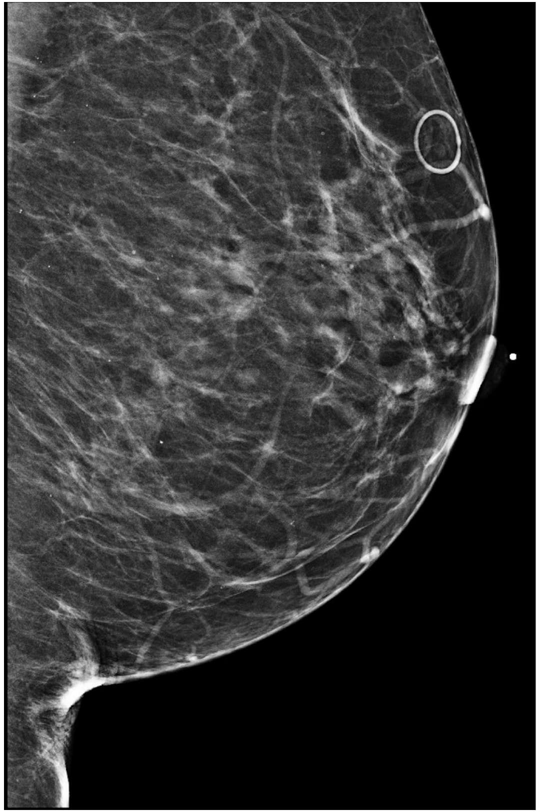
Error Rates in Commercial Gender Classification Products

		Face++ 旷视	IBM
Dark Skinned Female	20.8%	34.5%	34.7%
Light Skinned Female	1.7%	6.0%	7.1%
Dark Skinned Male	6.0%	0.7%	12.0%
Light Skinned Male	0.0%	0.8%	0.3%

Age prediction



Mamography



Mammography databases have a lot of images in them, but they suffer from one problem that has caused significant issues in recent years — almost all of the x-rays are from white women. This may not sound like a big deal, but actually, black women have been shown to be 42 percent more likely to die from breast cancer due to a wide range of factors that may include differences in detection and access to health care. Thus, training an algorithm primarily on white women adversely impacts black women in this case.