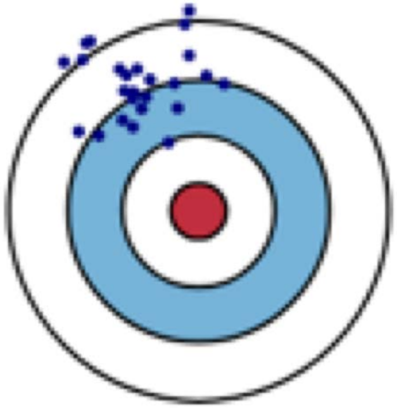
Tufts COMP 135: Introduction to Machine Learning
https://www.cs.tufts.edu/comp/135/2019s/

Probability and Statistical Decision Theory



Many slides attributable to:
Erik Sudderth (UCI)
Finale Doshi-Velez (Harvard)
James, Witten, Hastie, Tibshirani (ISL/ESL books)

Prof. Mike Hughes

Logistics

- Recitation tonight: 730-830pm, Halligan 111B
  - More on pipelines and feature transforms
  - Cross validation

Unit Objectives

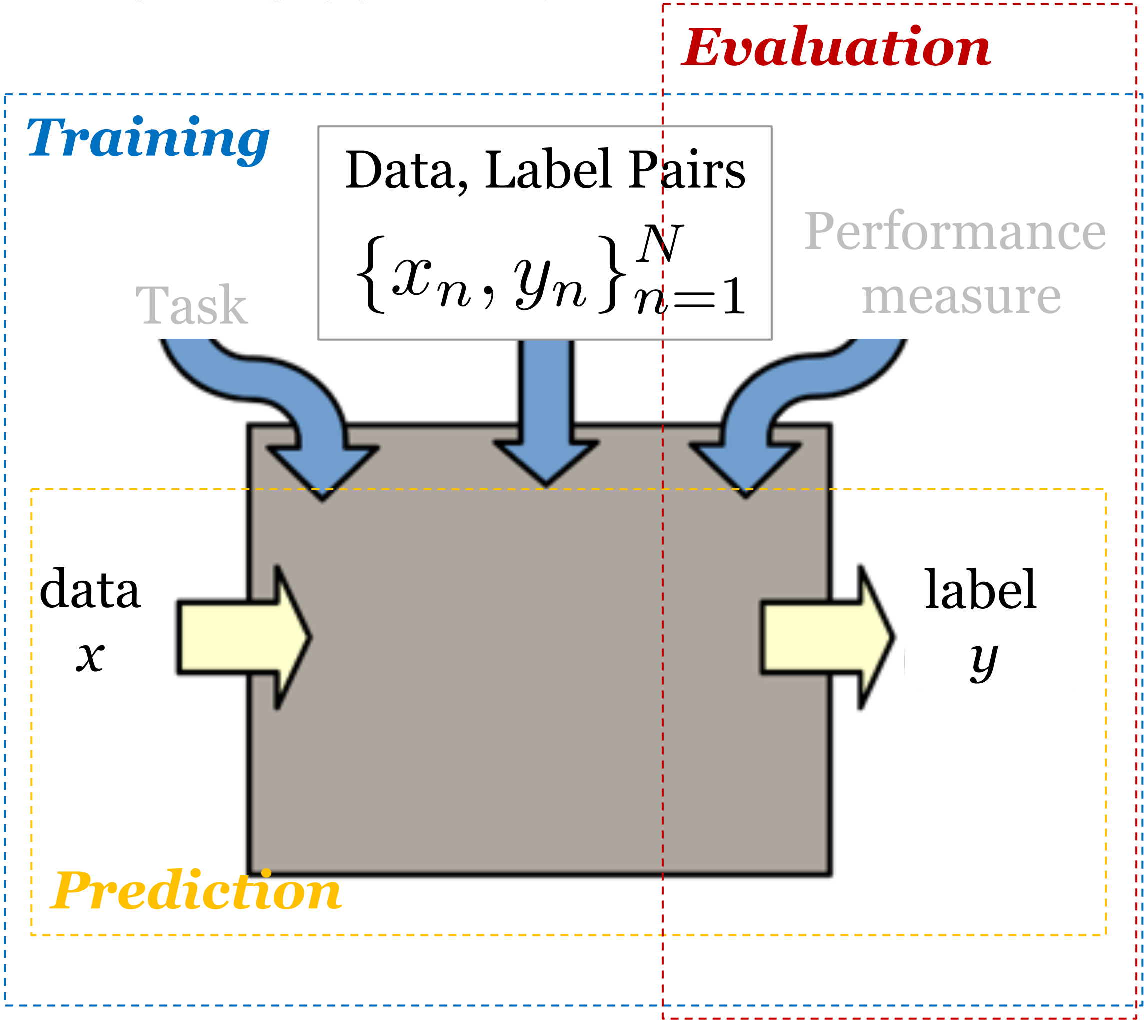
- Probability Basics
  - Discrete random variables
  - Continuous random variables
- Decision Theory: Making optimal predictions
- Limits of learning
  - The curse of dimensionality
  - The bias-variance tradeoff

What will we learn?

Supervised Learning

Unsupervised Learning

Reinforcement Learning



Task: Regression

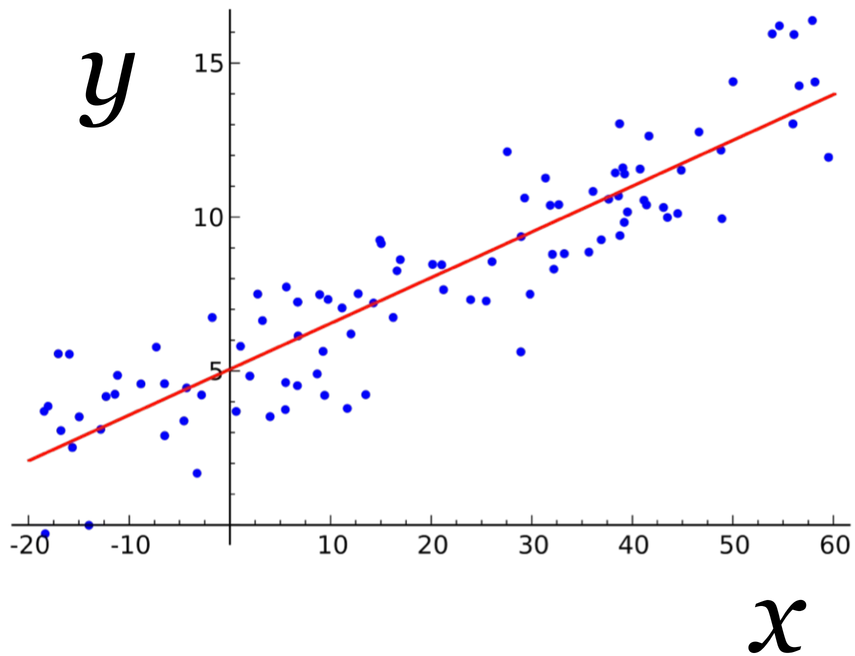
Supervised Learning

regression

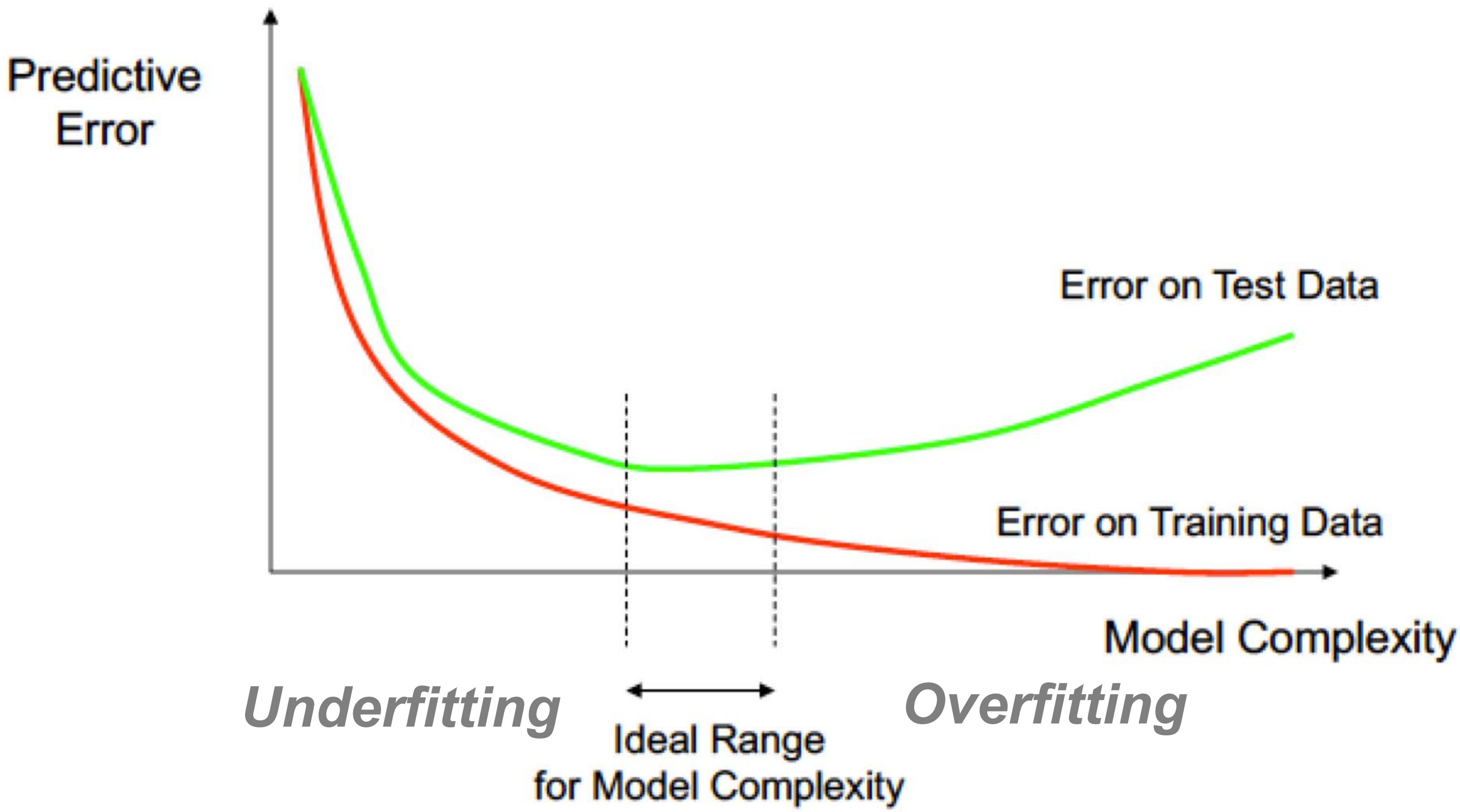
Unsupervised Learning

Reinforcement Learning

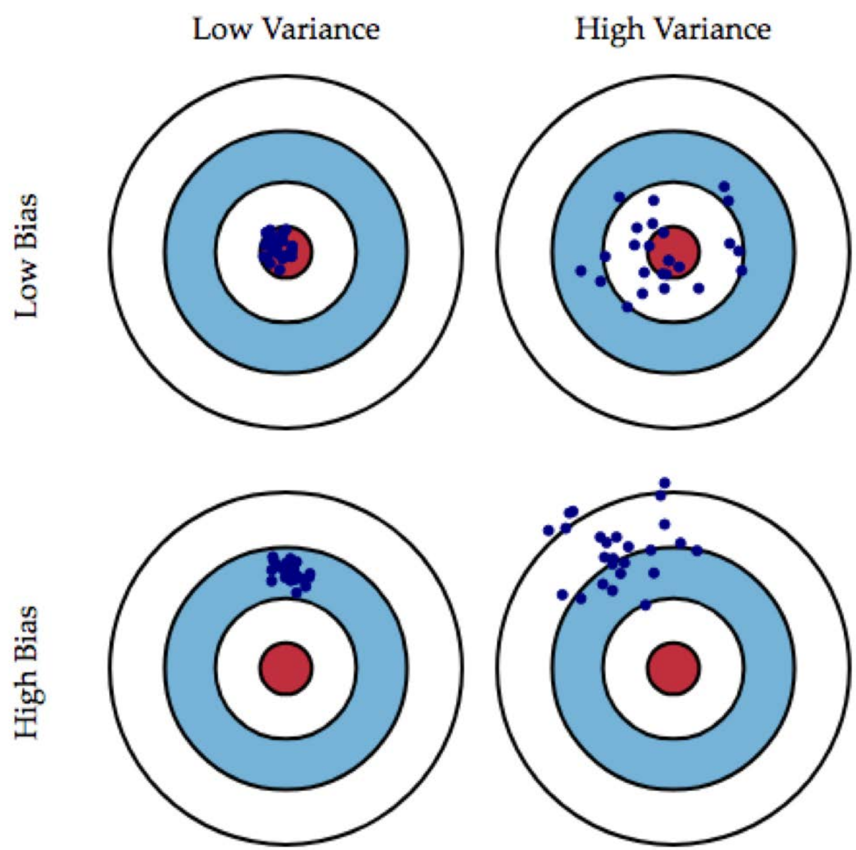
y is a numeric variable
e.g. sales in $$



Model Complexity vs Error

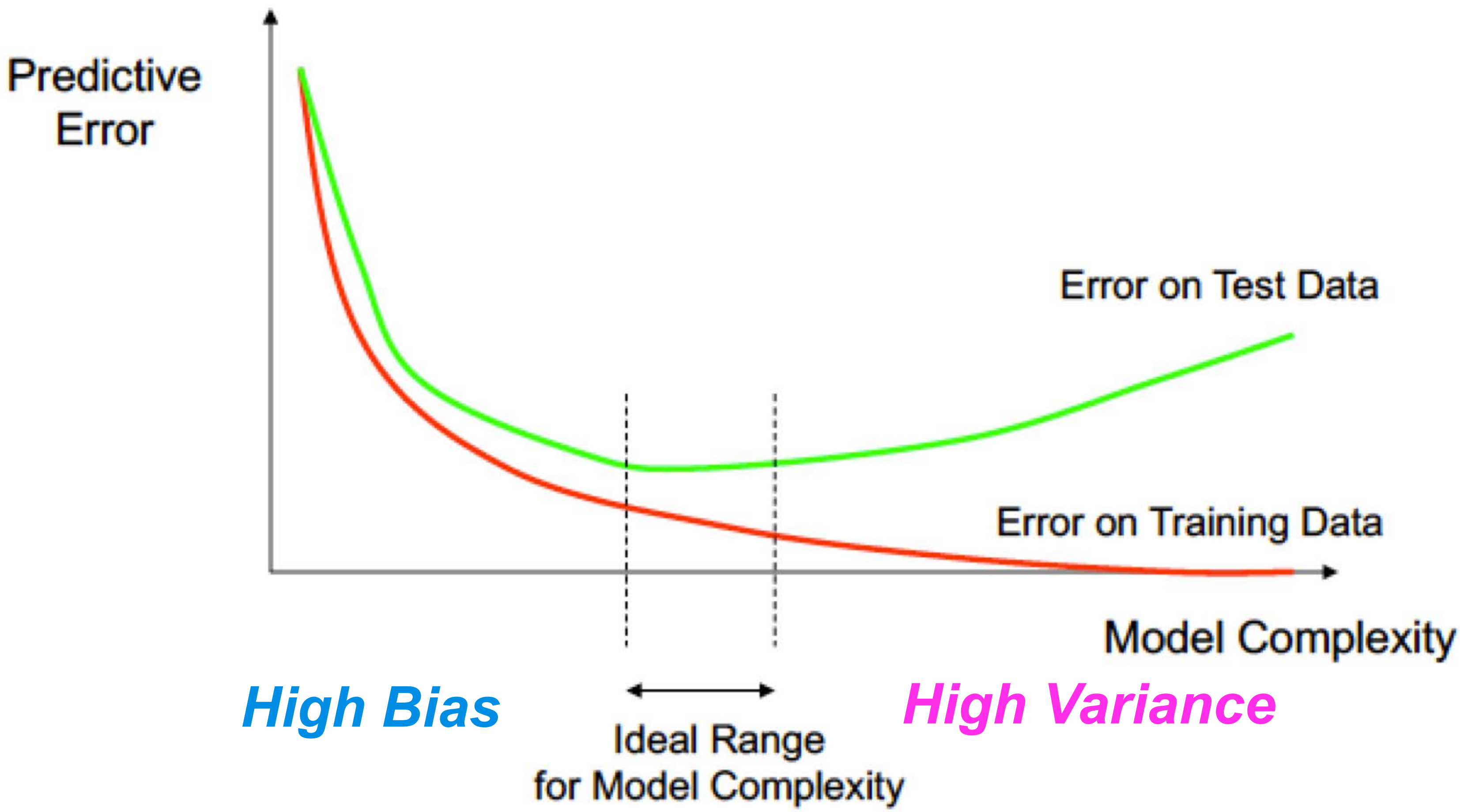


Today: Bias and Variance



Credit: Scott Fortmann-Roe
http://scott.fortmann-roe.com/docs/BiasVariance.html

Model Complexity vs Error



Discrete Random Variable

Examples:

- Coin flip! Heads or tails?
- Dice roll! 1 or 2 or … 6?

In general, random variable is defined by:

- Countable set of **all possible** outcomes
- **Probability value** for each outcome

Probability Mass Function

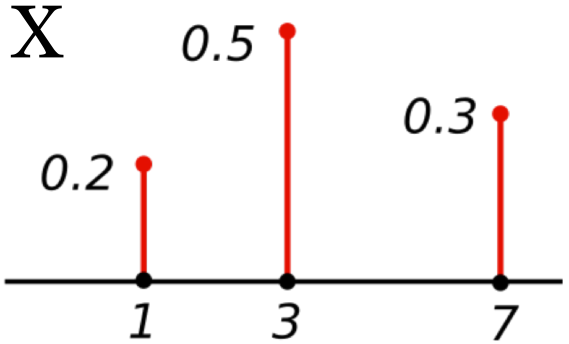
Notation:

- X is random variable
- x is a particular observed value
- Probability of observation: $p(X = x)$

Function p is a probability mass function (pmf)

Maps possible values to probabilities in [0, 1]

Must sum to one over domain of X



Pair exercise

- Draw the pmf for a normal 6-sided dice roll

- Draw pmf if there are:
    - 2 sides with 1 pip
    - 0 sides with 2 pips

Expected Values

What is the *expected* value of a dice roll?

Expected means probability-weighted average

$$\mathbb{E}[X] = \sum_x p(X = x)x$$

Joint Probability

X

Y

	Candidate A	Candidate B
Young voters	0.28	0.42
Senior voters	0.24	0.06

$$p(X = \text{candidate A} \quad \text{AND} \quad Y = \text{young})$$

Marginal Probability

X

Y		Candidate A	Candidate B
	Young voters	0.28	0.42
	Senior voters	0.24	0.06

Marginal p(Y):

0.7
0.3

Marginal p(X):	0.52	0.48

Conditional Probability

What is the probability of support for candidate A, *if we assume* that the voter is young?

Goal: $p(X = \text{candidate A} | Y = \text{young})$

X

Y		Candidate A	Candidate B
	Young voters	0.28	0.42
	Senior voters	0.24	0.06

Marginal p(Y):

0.7
0.3

Try it with your partner!

Conditional Probability

What is the probability of support for candidate A, *if we assume* that the voter is young?

Goal: $p(X = \text{candidate A} | Y = \text{young})$

X

Y

	Candidate A	Candidate B
Young voters	0.28	0.42
Senior voters	0.24	0.06

Marginal p(Y):

0.7
0.3

Answer:

	Candidate A	Candidate B
Young voters	0.4	0.6

The Rules of Probability

sum rule $$p(X) = \sum_{Y} p(X, Y)$$

product rule $$\begin{aligned} p(X, Y) &= p(Y|X)p(X) \\ &= p(X|Y)p(Y) \end{aligned}$$

Continuous Random Variables

Any r.v. whose possible outcomes are not a discrete set, but take values on a number line

Examples:

uniform draw between 0 and 1

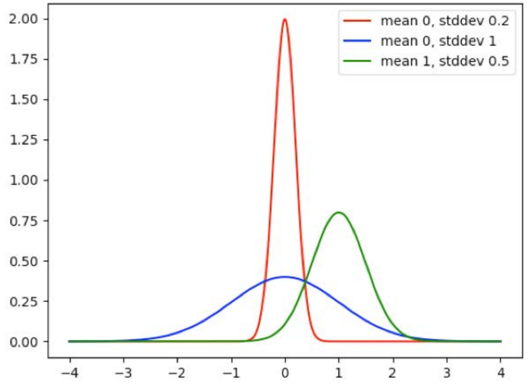
draw from Gaussian “bell curve” distribution

Probability Density Function

- Generalizes pmf for discrete r.v. to continuous
- Any pdf p(x) must satisfy two properties:

$$\forall x : p(x) \geq 0$$

$$\int_x p(x) dx = 1$$



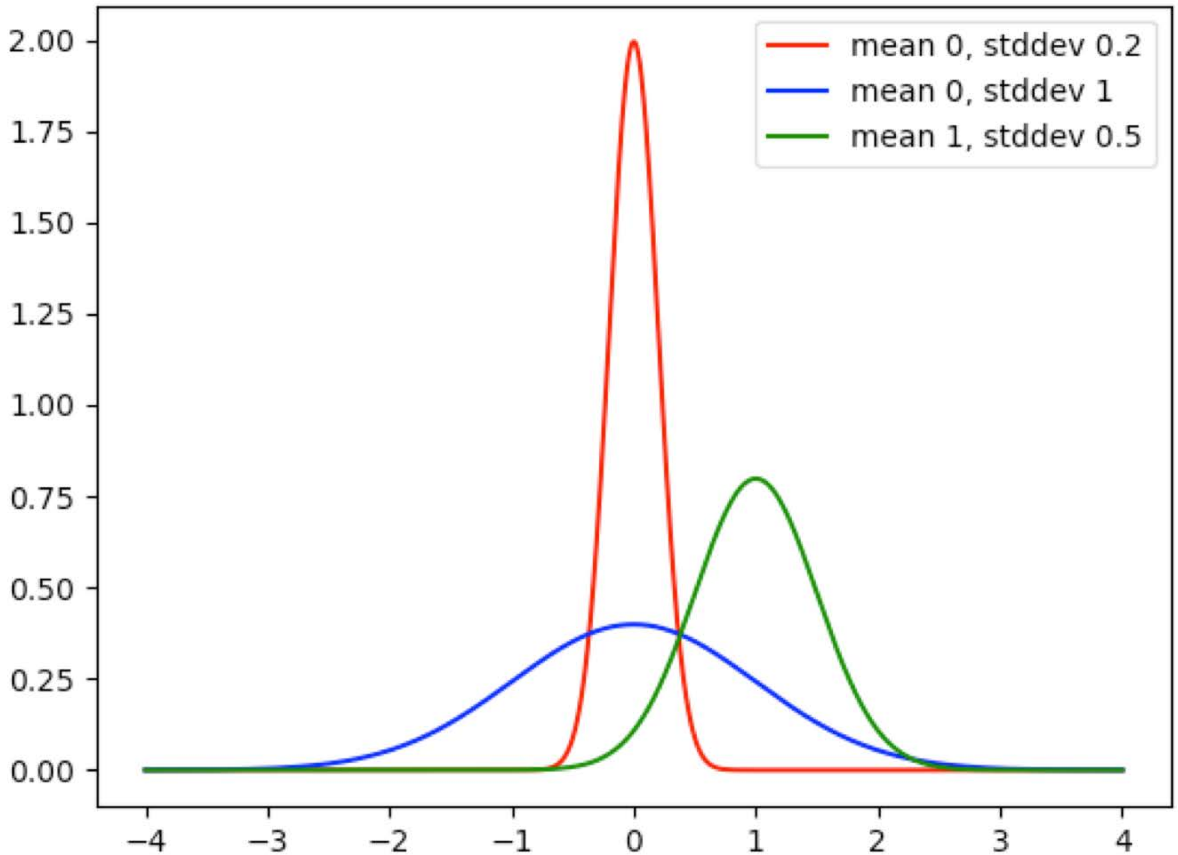
Example

Consider a uniform distribution over entire real line
(from -inf to + inf)

Draw the pdf, verify that it can meet the required conditions (nonnegative, integrates to one).

Is there a problem here?

Plots of Gaussian pdf



What do you notice about y-axis values…

Is there a problem here?

Probability Density Function

- Generalizes pmf for discrete r.v. to continuous
- Any pdf p(x) must satisfy two properties:

$$\forall x : p(x) \geq 0$$

$$\int_x p(x) dx = 1$$

Value of p(x) can take ANY value > 0, even sometimes larger than 1

Should NOT interpret as "probability of drawing exactly x"

Should interpret as "density at vanishingly small interval around x"

Remember: density = mass / volume

Continuous Expectations

$$\mathbb{E}[X] = \int_{x \in \mathrm{domain}(X)} x p(x) dx$$

$$\mathbb{E}[h(X)] = \int_{x \in \mathrm{domain}(X)} h(x) p(x) dx$$

Approximating Expectations

Use “Monte Carlo”: average of a sample!

- 1) Draw S i.i.d. samples from distribution

$$x^1, x^2, \dots x^S \sim p(x)$$

- 2) Compute mean of these sampled values

$$\mathbb{E}[h(X)] \approx \frac{1}{S} \sum_{s=1}^{S} h(x^s)$$

For any function h, the mean of this random estimator is unbiased.
As number of samples S increases, variance of estimator decreases.

Statistical Decision Theory

- See ESL textbook in Ch. 2 and Ch. 7

How to predict best if we know conditional probability?

Assume we have: a specific x input of interest

a known "true" conditional p(Y | X)

error metric we care about

How should we set our predictor $\hat{y}$? *Minimize the expected error!*

$$\min_{\hat{y}} \quad \mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x]$$

Key ideas:

prediction will be a scalar

conditional distribution p(Y|X) tells us everything we need to know

Expected y at a given fixed x

$$\mathbb{E}[Y|X = x] = \int_y y \ p(y|X = x) dy$$

Recall from HW1

- Two constant value estimators
  - Mean of training set
  - Median of training set

- Two possible error metrics
  - Squared error
  - Absolute error

Which estimator did best under which error metric?

Minimize expected squared error

Assume we have: a specific x input of interest
a known “true” conditional p(y | x)

$$\mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x] = \int_y (y - \hat{y})^2 \; p(y | X = x) dy$$

What is your intuition from HW1? Express in terms of p(Y|X=x)…

How should we set our predictor $\hat{y}$ to minimize the expected error?

$$\min_{\hat{y}} \quad \mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x]$$

Minimize expected squared error

Assume we have: a specific x input of interest
a known "true" conditional p(y | x)

$$\mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x] = \int_y (y - \hat{y})^2 \; p(y | X = x) dy$$

How should we set our predictor $\hat{y}$ to minimize the expected error?

$$\min_{\hat{y}} \quad \mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x]$$

Optimal predictor for squared error: mean y value under p(Y|X=x)

$$\hat{y} = \mathbb{E}[Y | X = x]$$

In practice, mean of sampled y values at/around x

Minimize expected absolute error

Assume we have: a specific x input of interest
a known "true" conditional p(y | x)

$$\mathbb{E}[\text{err}(Y, \hat{y}) | X = x] = \int_y \boxed{|y - \hat{y}|}\, p(y | X = x) dy$$

How should we set our predictor $\hat{y}$ to minimize the expected error?

$$\min_{\hat{y}} \quad \mathbb{E}[\text{err}(Y, \hat{y}) | X = x]$$

What is your intuition from HW 1?

Minimize expected absolute error

Assume we have: a specific x input of interest

a known "true" conditional p(y | x)

$$\mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x] = \int_y \boxed{|y - \hat{y}|}\, p(y | X = x) dy$$

How should we set our predictor $\hat{y}$ to minimize the expected error?

$$\min_{\hat{y}} \quad \mathbb{E}[\mathrm{err}(Y, \hat{y}) | X = x]$$

Optimal predictor for squared error: **median** y value under p(Y|X=x)

$$\hat{y}^* = \mathrm{median}(p(Y | X = x))$$

In practice, median of sampled y values at/around x

Minimizing error with K-NN

Ideal

- know “true” conditional p(y | x)

Approximation

- Use neighborhood around x
- Take average of y values in neighborhood

If we have enough training data, K-NN is **good approximation**

Some theorems say KNN estimate ideal as # examples (N) gets infinitely large

Problem in practice: we **never** have enough data, esp. if feature dimensions are large

Curse of Dimensionality

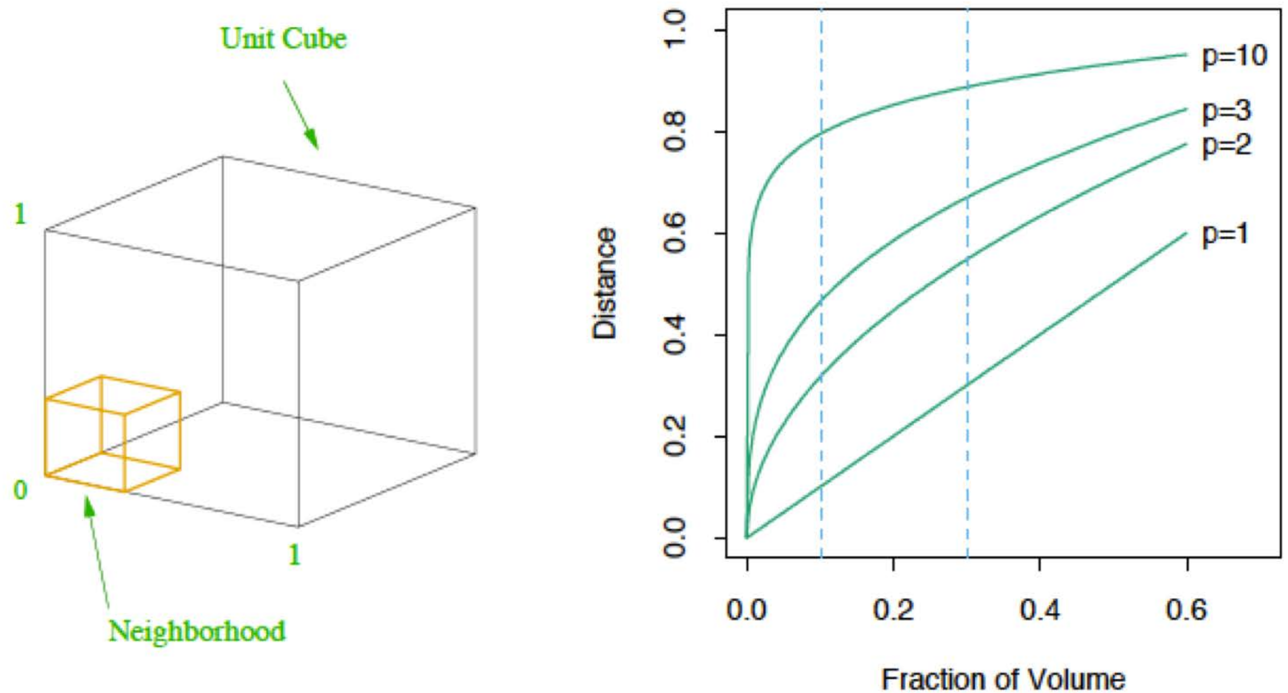
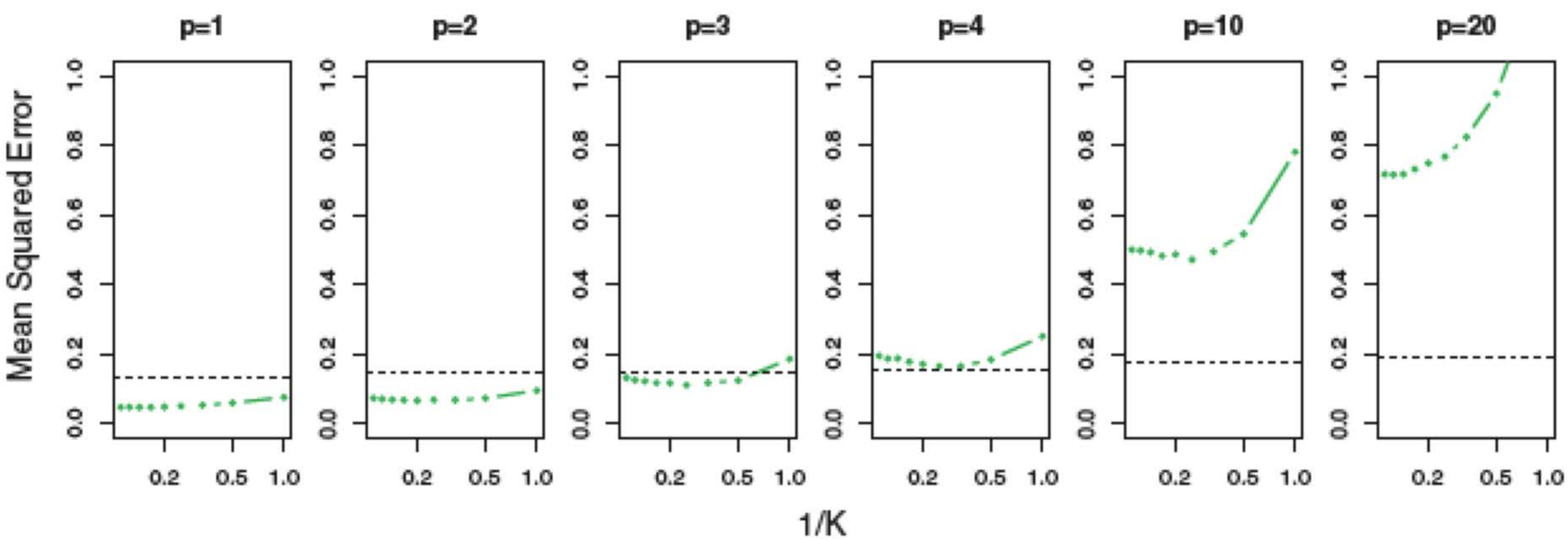


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

MSE as dimension increases



-- Linear Regression

• K Neighbors Regression

Credit: ISL textbook, Fig 3.20

Write MSE via Bias & Variance

y is known "true" response value at given fixed input x

$\hat{y}$ is a Random Variable obtained by fitting estimator to random sample of N training data examples, then predicting at fixed x

$$\mathbb{E}\Big[\big(\hat{y}(x^{tr}, y^{tr}) - y\big)^2 \Big] = \mathbb{E}\Big[(\hat{y} - y)^2 \Big]$$

$$= \mathbb{E}\Big[\hat{y}^2 - 2\hat{y}y + y^2\Big]$$

$$= \mathbb{E}\Big[\hat{y}^2\Big] - 2\bar{y}y + y^2$$

$$\boxed{\bar{y} \triangleq \mathbb{E}[\hat{y}]}$$

Write MSE via Bias & Variance

$$\mathbb{E}\Big[\big(\hat{y}(x^{tr}, y^{tr}) - y\big)^2\Big] = \mathbb{E}\Big[(\hat{y} - y)^2\Big]$$

$$= \mathbb{E}\Big[\hat{y}^2 - 2\hat{y}y + y^2\Big]$$

$$= \mathbb{E}\Big[\hat{y}^2\Big] - 2\bar{y}y + y^2$$

Add net value of zero
Pick 0 = -a + a

$$= \mathbb{E}\Big[\hat{y}^2\Big] \boxed{- \bar{y}^2 + \bar{y}^2} - 2\bar{y}y + y^2$$

Write MSE via Bias & Variance

$$\mathbb{E}\Big[\big(\hat{y}(x^{tr}, y^{tr}) - y\big)^2 \Big] = \mathbb{E}\Big[(\hat{y} - y)^2 \Big]$$

$$= \mathbb{E}\Big[\hat{y}^2 - 2\hat{y}y + y^2\Big]$$

$$= \mathbb{E}\Big[\hat{y}^2\Big] - 2\bar{y}y + y^2$$

$$= \mathbb{E}\Big[\hat{y}^2\Big] - \bar{y}^2 + \boxed{\bar{y}^2 - 2\bar{y}y + y^2}$$

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

$$\text{bias}^2$$

$$\boxed{\text{bias} \triangleq \bar{y} - y}$$

MSE = Variance + Bias^2

$$\mathbb{E}\Big[\big(\hat{y}(x^{tr}, y^{tr}) - y\big)^2 \Big] = \mathbb{E}\Big[(\hat{y} - y)^2 \Big]$$

$$= \mathbb{E}\Big[\hat{y}^2 - 2\hat{y}y + y^2\Big]$$

$$= \mathbb{E}\Big[\hat{y}^2\Big] - 2\bar{y}y + y^2$$

$$= \boxed{\mathbb{E}\Big[\hat{y}^2\Big] - \bar{y}^2} + \bar{y}^2 - 2\bar{y}y + y^2$$

$$\boxed{\mathrm{Var}[X] \triangleq \mathbb{E}[X^2] - \mathbb{E}^2}$$

$$= \boxed{\mathrm{Var}(\hat{y})} + (\bar{y} - y)^2$$

Punchline

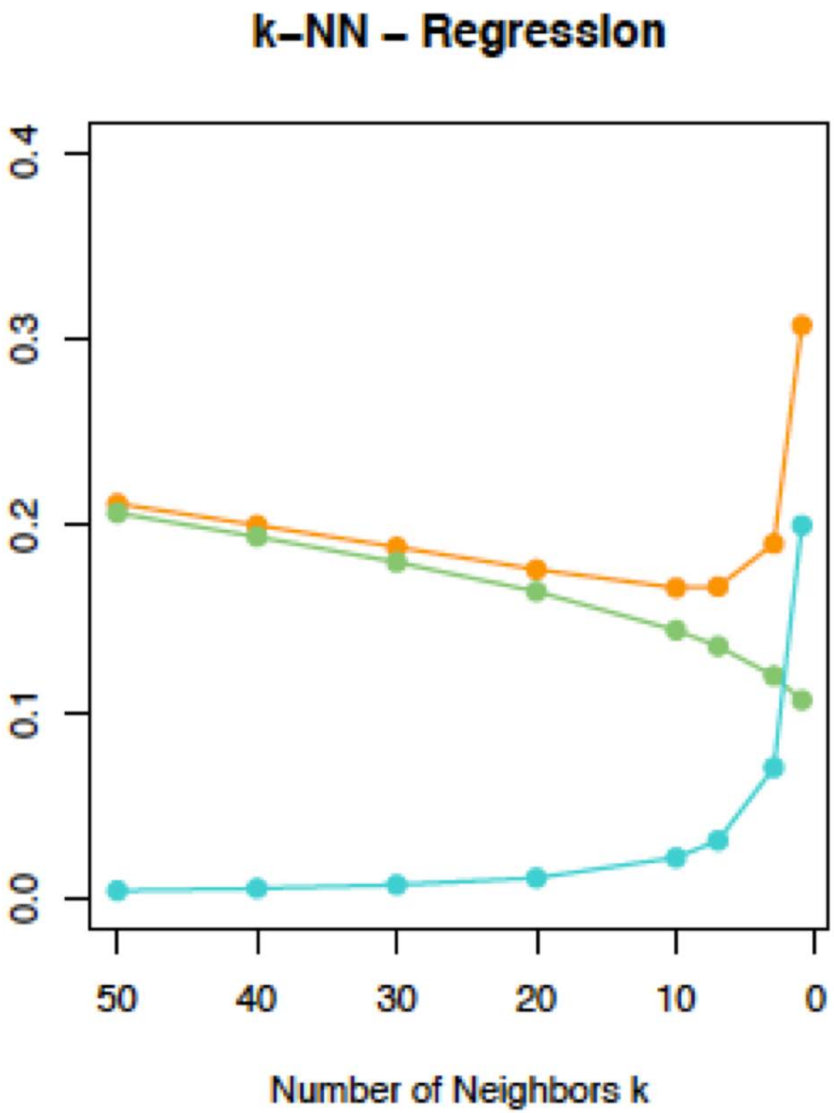
mean squared error = variance + bias^2

We can use this framing to explain tradeoffs of different prediction approaches on finite training datasets.

Toy example: ESL Fig. 7.3

Figure 7.3 shows the bias–variance tradeoff for two simulated examples. There are 80 observations and 20 predictors, uniformly distributed in the hypercube $[0, 1]^{20}$. The situations are as follows:

Left panels: Y is 0 if $X_1 \leq 1/2$ and 1 if $X_1 > 1/2$, and we apply k-nearest neighbors.



More flexible

total error

bias

Error due to inability of average model to capture true predictive relationship

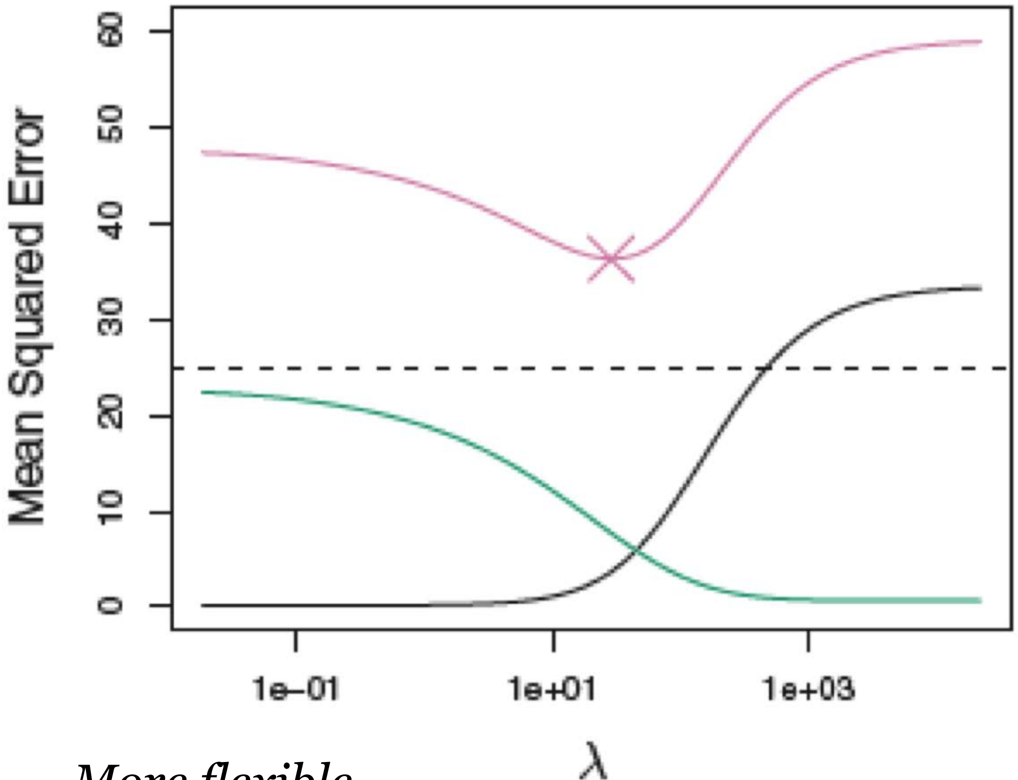
variance

Error due to estimating from a single finite sample

Toy example: ISL Fig. 6.5

Why Does Ridge Regression Improve Over Least Squares?

Ridge regression's advantage over least squares is rooted in the *bias-variance trade-off*. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias. This is illustrated in the left-hand panel of Figure 6.5, using a simulated data set containing $p = 45$ predictors and $n = 50$ observations. The green curve in the left-hand panel



More flexible

total error

bias

Error due to inability of average fit to capture true predictive relationship

variance

Error due to estimating from a single finite sample

Can Also Treat True Y as R.V.

$$Y = f(X) + \epsilon$$

True signal function

Noise Random Variable
Symmetric (zero mean)

Often, Gaussian

The Final MSE decomposition

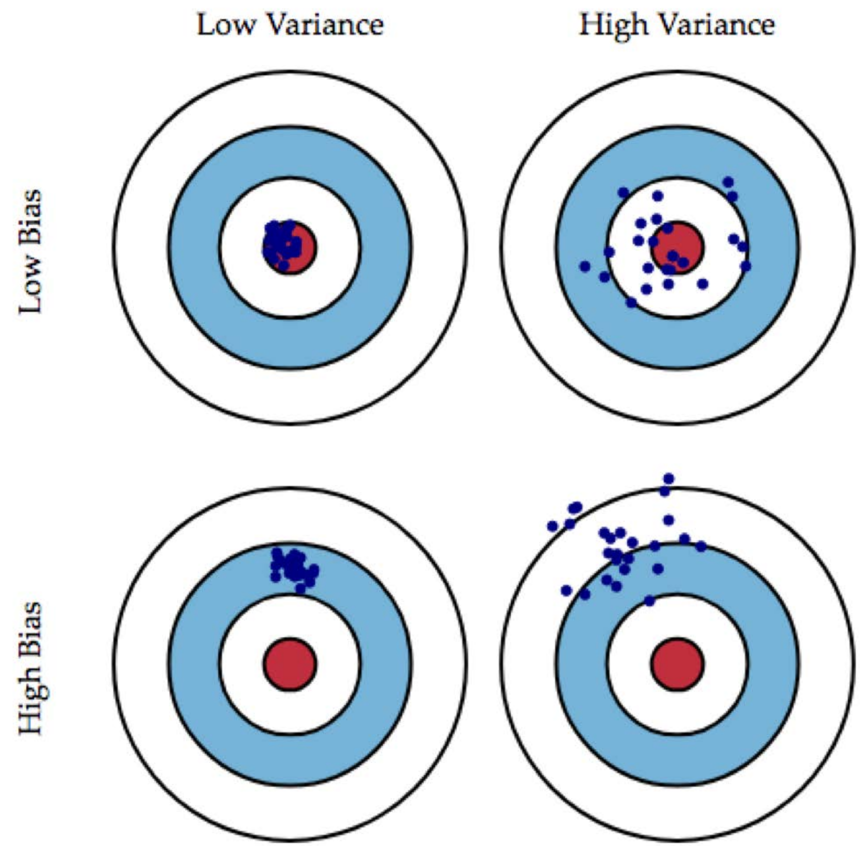
$$\mathbb{E}[MSE] = \mathrm{Var}(\hat{y}) + \mathrm{bias}^2 + \text{irreducible error}$$

For more, see Sec. 7.3 of ESL textbook…

As in Chapter 2, if we assume that $Y = f(X) + \varepsilon$ where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$, we can derive an expression for the expected prediction error of a regression fit $\hat{f}(X)$ at an input point $X = x_0$, using squared-error loss:

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}. \qquad (7.9)
\end{aligned}$$

Bias and Variance



Credit: Scott Fortmann-Roe
http://scott.fortmann-roe.com/docs/BiasVariance.html