Model Flexibility and Cross-Validation

Damek Davis
School of ORIE, Cornell University
ORIE 4740 Lec 7–8 (Feb 15, 17)

Announcements

Recap

Statistical learning: $Y = \underbrace{f(\vec{X})}_{\text{Explained part}} + \underbrace{\epsilon}_{\text{Unexplained part}}$

Training dataset: Fit model $\hat{f}$ using $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$

Test dataset: Make prediction on new data points $\vec{x}_{n+1}, \vec{x}_{n+2}, \ldots$

Regression: Response Y is continuous

- Linear regression
- More flexible: add non-linear terms $X_1 X_2$, X_1^3, $\log X_2$, $\sqrt{X_4}$, etc

Classification: Y is categorical/qualitative

- Logistic regression. More flexibility by adding non-linear terms.
- K-Nearest Neighbor. More flexibility by decreasing K

Evaluation: Training error vs Test Error

Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.

Evaluation: Training error vs Test Error

Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.
- Important: errors on the test dataset $(\tilde{x}_i, \tilde{y}_i), i = 1, 2, \ldots, m$.
  - previously unseen data
  - not used to build $\hat{f}$.

Evaluation: Training error vs Test Error

Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad\qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.
- Important: errors on the test dataset $(\tilde{x}_i, \tilde{y}_i), i = 1, 2, \ldots, m$.
  - previously unseen data
  - not used to build $\hat{f}$.
  - $\implies$ measure testing MSE or testing Classification error.

Evaluation: Training error vs Test Error

Do we care about training error or test error?

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

- Computed on the training dataset $(x_i, y_i), i = 1, 2, \ldots, n$.
- Important: errors on the test dataset $(\tilde{x}_i, \tilde{y}_i), i = 1, 2, \ldots, m$.
  - previously unseen data
  - not used to build $\hat{f}$.
  - $\Longrightarrow$ measure testing MSE or testing Classification error.

  How to measure testing error?

A.

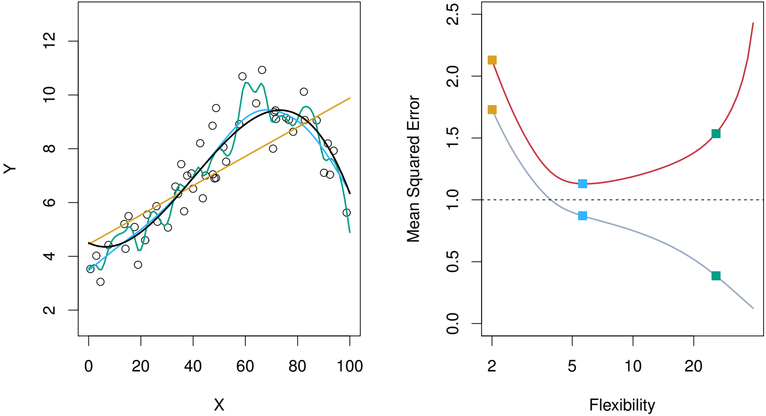
$$\boxed{\text{MSE} = \frac{1}{m}\sum_{i=1}^{n}(\tilde{y}_i - \hat{f}(\tilde{x}_i))^2 \qquad \frac{1}{m}\sum_{i=1}^{m} I(\tilde{y}_i \neq \hat{f}(\tilde{x}_i))}$$

B.

$$\boxed{\text{MSE} = \frac{1}{m}\sum_{i=1}^{n}(\tilde{y}_i - f(\tilde{x}_i))^2 \qquad \frac{1}{m}\sum_{i=1}^{m} I(\tilde{y}_i \neq f(\tilde{x}_i))}$$

Training Error vs. Testing Error

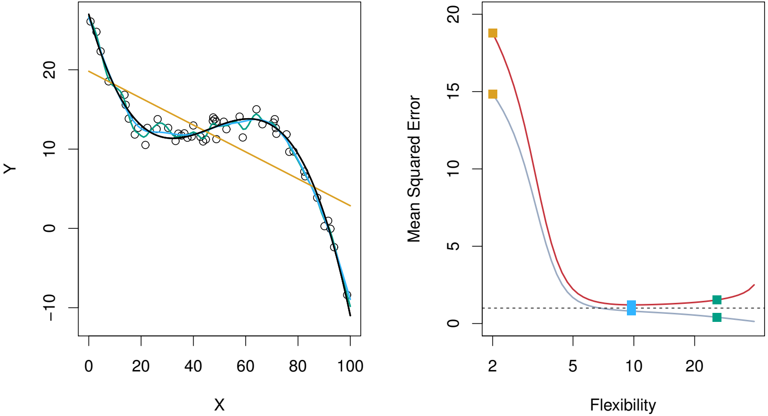
The regression setting:



Red: test error. Gray: training error

Training Error vs. Testing Error

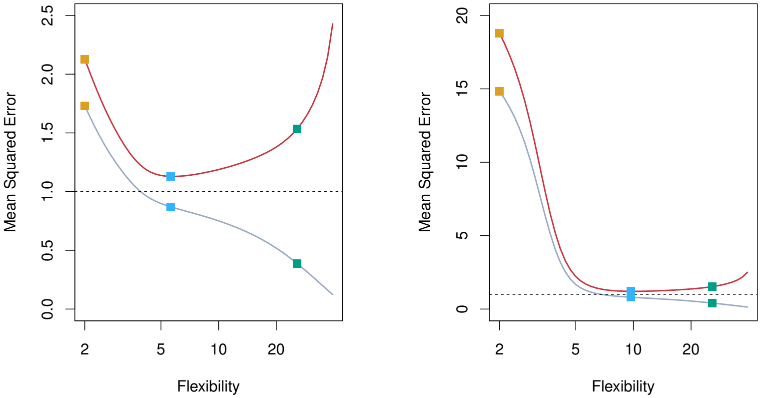
The regression setting:



Red: test error. Gray: training error

Training Error vs. Testing Error

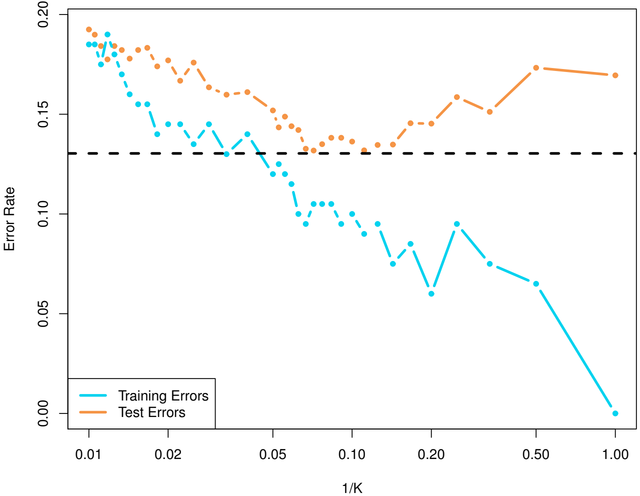
The regression setting:



Red: test error. Gray: training error

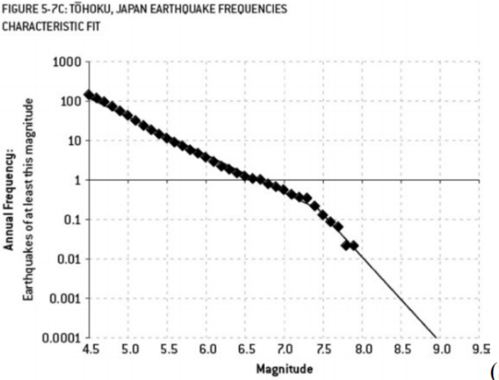
Training Error vs. Testing Error

The classification setting: KNN



large K low flexibility small K high flexibility

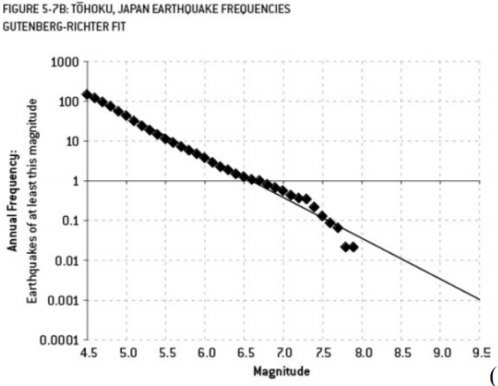
Fukushima



(Silver, N, 2012)

$$\text{frequency} = f(\text{magnitude}) + \epsilon$$

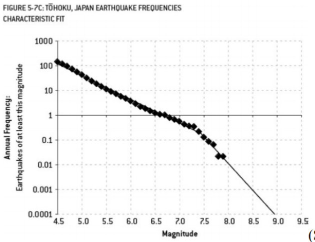
Fukushima



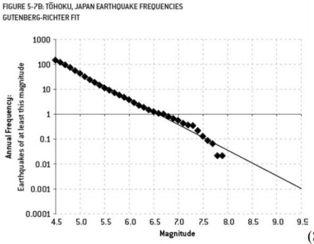
(Silver, N, 2012)

$$\log(\text{frequency}) = \beta_0 + \beta_1 \text{magnitude} + \epsilon$$

Fukushima



(Silver, N, 2012)



(Silver, N, 2012)

- GR Fit. 9.0 earthquake once every 300 years.
- Fukushima Team. 9.0 earthquake once every 13000 years.

  - Fukushima only built to withstand 8.6 earthquake (2.5× weaker).

Brian Stacey, Fukushima: The Failure of Predictive Models

Bias-Variance Tradeoff

(Reading: ISLR Section 2.2.2)

Say:

- θ is an unknown quantity we want to estimate
- $\hat{\theta}$ is an estimator of θ (computed from data, random)

Bias-Variance Tradeoff

(Reading: ISLR Section 2.2.2)

Say:

- θ is an unknown quantity we want to estimate
- $\hat{\theta}$ is an estimator of θ (computed from data, random)

Proved in HW0:

The mean squared error (MSE) decomposes as:

$$\underbrace{\mathbb{E}(\hat{\theta}-\theta)^2}_{\text{MSE}} = \underbrace{\left(\mathbb{E}\hat{\theta}-\theta\right)^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{\theta})}_{\text{variance}}$$

Bias-Variance Tradeoff

(Reading: ISLR Section 2.2.2)

Say:

- θ is an unknown quantity we want to estimate
- $\hat{\theta}$ is an estimator of θ (computed from data, random)

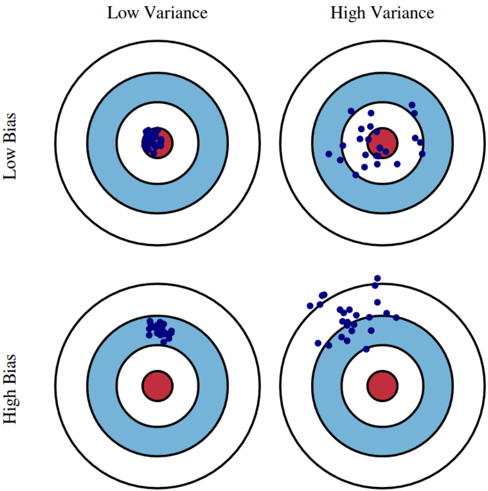
Proved in HW0:

The mean squared error (MSE) decomposes as:

$$\underbrace{\mathbb{E}(\hat{\theta}-\theta)^2}_{\text{MSE}} = \underbrace{\left(\mathbb{E}\hat{\theta}-\theta\right)^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{\theta})}_{\text{variance}}$$

$$\text{MSE} = \text{bias}^2 + \text{variance}$$

Bias-Variance Tradeoff



Source:
`http://scott.fortmann-roe.com/docs/BiasVariance.html`

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}(\hat{\theta}-\theta)^2}_{\text{MSE}} = \underbrace{\left(\mathbb{E}\hat{\theta}-\theta\right)^2}_{\text{bias}^2} + \underbrace{\mathsf{Var}(\hat{\theta})}_{\text{variance}}$$

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}(\hat{\theta}-\theta)^2}_{\text{MSE}} = \underbrace{\left(\mathbb{E}\hat{\theta}-\theta\right)^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{\theta})}_{\text{variance}}$$

- Applied to regression:

Want to predict *y* for a fixed test data point *x*

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}(\hat{\theta}-\theta)^2}_{\text{MSE}} = \underbrace{\left(\mathbb{E}\hat{\theta}-\theta\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{variance}}$$

- Applied to regression:

Want to predict y for a fixed test data point x

y is generated from the true model $y = f(x) + \epsilon$
(ϵ has zero mean and independent of everything else)

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}(\hat{\theta} - \theta)^2}_{\text{MSE}} = \underbrace{\left(\mathbb{E}\hat{\theta} - \theta\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{variance}}$$

- Applied to regression:

Want to predict y for a fixed test data point x

y is generated from the true model $y = f(x) + \epsilon$
(ϵ has zero mean and independent of everything else)

Estimate using some model $\hat{f}$: $\hat{y} = \hat{f}(x)$

Then:

$$\text{test error} = \mathbb{E}(\hat{f}(x) - y)^2 = \underbrace{\left(\mathbb{E}\hat{f}(x) - f(x)\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(\hat{f}(x))}_{\text{variance}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible error}}$$

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

$$\mathbb{E}(\hat{f}(x) - y)^2$$

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

$$
\begin{aligned}
&\mathbb{E}(\hat{f}(x) - y)^2 \\
=&\mathbb{E}\Big[\mathbb{E}\big[(\hat{f}(x) - y)^2 \big| y\big]\Big] \qquad \text{(tower property of conditional expectation)}
\end{aligned}
$$

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

$$
\begin{aligned}
&\mathbb{E}(\hat{f}(x) - y)^2 \\
=&\mathbb{E}\Big[\mathbb{E}\big[(\hat{f}(x) - y)^2 \big| y\big]\Big] && \text{(tower property of conditional expectation)} \\
=&\mathbb{E}\Big[\big(\mathbb{E}\hat{f}(x) - y\big)^2 + \mathrm{Var}(\hat{f}(x))\Big] && \text{(hw0)}
\end{aligned}
$$

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

$$
\begin{aligned}
&\mathbb{E}(\hat{f}(x) - y)^2 \\
=&\mathbb{E}\Big[\mathbb{E}\big[(\hat{f}(x) - y)^2 \big| y\big]\Big] && \text{(tower property of conditional expectation)} \\
=&\mathbb{E}\Big[\big(\mathbb{E}\hat{f}(x) - y\big)^2 + \mathrm{Var}(\hat{f}(x))\Big] && \text{(hw0)} \\
=&\mathbb{E}\big(\mathbb{E}\hat{f}(x) - f(x) - \epsilon\big)^2 + \mathrm{Var}(\hat{f}(x))
\end{aligned}
$$

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

$$
\begin{aligned}
&\mathbb{E}(\hat{f}(x) - y)^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[(\hat{f}(x) - y)^2 \middle| y\right]\right] \qquad \text{(tower property of conditional expectation)} \\
=&\mathbb{E}\left[\left(\mathbb{E}\hat{f}(x) - y\right)^2 + \mathrm{Var}(\hat{f}(x))\right] \qquad \text{(hw0)} \\
=&\mathbb{E}\left(\mathbb{E}\hat{f}(x) - f(x) - \epsilon\right)^2 + \mathrm{Var}(\hat{f}(x)) \\
=&\mathbb{E}\left(\mathbb{E}\hat{f}(x) - f(x)\right)^2 + 2\mathbb{E}\left[\left(\mathbb{E}\hat{f}(x) - f(x)\right)\epsilon\right] + \mathbb{E}\epsilon^2 + \mathrm{Var}(\hat{f}(x))
\end{aligned}
$$

Proof (Optional)

Two sources of randomness:

- The *test* data y is random, because $y = f(x) + \epsilon$ and ϵ is random (x is fixed)
- The estimator $\hat{f}$ is random, because it is fitted on random *training* data

$$
\begin{aligned}
& \mathbb{E}(\hat{f}(x) - y)^2 \\
=& \mathbb{E}\Big[\mathbb{E}\big[(\hat{f}(x) - y)^2 \big| y\big]\Big] && \text{(tower property of conditional expectation)} \\
=& \mathbb{E}\Big[\big(\mathbb{E}\hat{f}(x) - y\big)^2 + \mathrm{Var}(\hat{f}(x))\Big] && \text{(hw0)} \\
=& \mathbb{E}\big(\mathbb{E}\hat{f}(x) - f(x) - \epsilon\big)^2 + \mathrm{Var}(\hat{f}(x)) \\
=& \mathbb{E}\big(\mathbb{E}\hat{f}(x) - f(x)\big)^2 + 2\mathbb{E}\Big[\big(\mathbb{E}\hat{f}(x) - f(x)\big)\epsilon\Big] + \mathbb{E}\epsilon^2 + \mathrm{Var}(\hat{f}(x)) \\
=& \underbrace{\big(\mathbb{E}\hat{f}(x) - f(x)\big)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible error}} + \underbrace{\mathrm{Var}(\hat{f}(x))}_{\text{variance}} && (\epsilon \perp \hat{f}, \mathbb{E}[\epsilon] = 0)
\end{aligned}
$$

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}(\hat{f}(x) - y)^2}_{\text{test error}} = \underbrace{\left(\mathbb{E}\hat{f}(x) - f(x)\right)^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{f}(x))}_{\text{variance}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error}}$$

- Less flexible model:

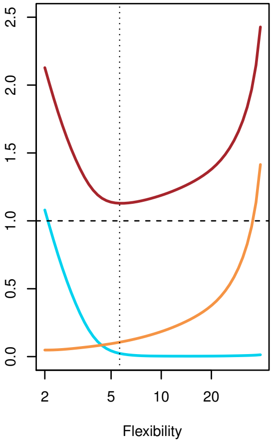
$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

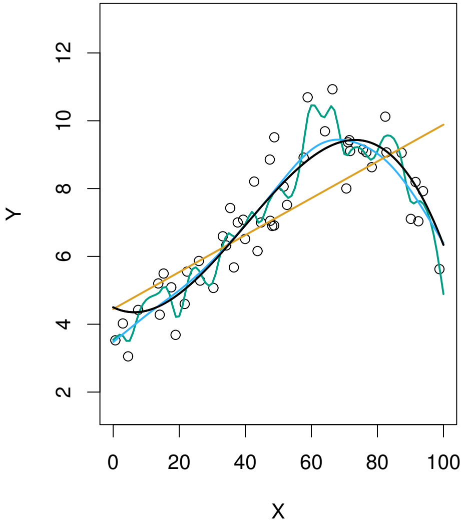
bias high, variance low

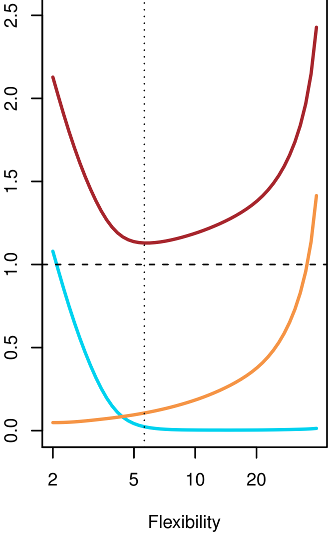
- More flexible model:

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3$$

bias low, variance high







Variance $\mathrm{Var}(\hat{f}(x))$ how much $\hat{f}$ changes when fitted using different datasets

Bias $\mathbb{E}\left[\hat{f}(x)\right] - f(x)$ difference b/w the true function f and the expectation of the model $\hat{f}$

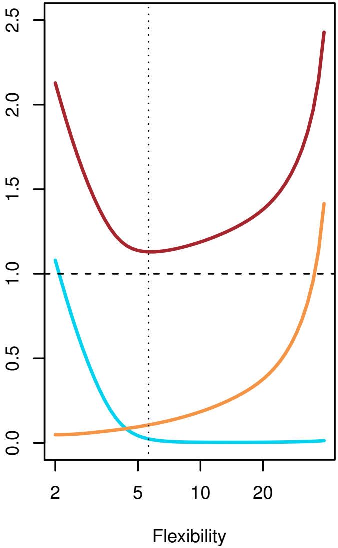
Bias-Variance Tradeoff

$$\text{test error} = \text{bias}^2 + \text{variance} + c$$

Bias: Large if model is not flexible enough

Variance: Large if model is too flexible

- True for regression and classification



Bias-Variance Tradeoff

In the Fukushima disaster, the engineering model had...

A. High Bias

B. High Variance

Bias-Variance Tradeoff

Scenario 1

- Training error is much lower than desired testing error

Scenario 2

- Training error is much higher than desired testing error.

In which scenario should you increase the flexibility of your model?

A. Scenario 1

B. Scenario 2

Bias-Variance Tradeoff for KNN

Which of the following is true about the KNN classifier?

A. As K grows, we expect higher bias and lower variance.

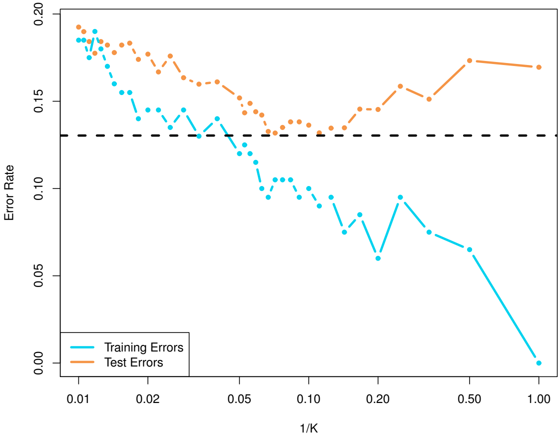
B. As K shrinks, we expect higher bias and lower variance.

Bias-Variance Tradeoff for KNN

Which of the following is true about the KNN classifier?

A. As *K* grows, we expect higher bias and lower variance.

B. As *K* shrinks, we expect higher bias and lower variance.



Bias-Variance Tradeoff

Suppose you try two classifiers: KNN and Logistic Regression. Which would we expect to have lower bias?

A. KNN

B. Logistic Regression

C. Depends on the value of K

Bias-Variance Tradeoff

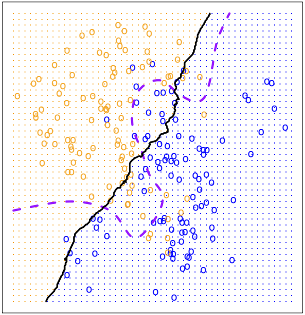
Suppose you try two classifiers: KNN and Logistic Regression. Which would we expect to have lower bias?

A. KNN

B. Logistic Regression

C. Depends on the value of *K*

KNN: K=100



Bias-Variance Tradeoff Mystery

- **In practice:** neural networks have 0 training error, but good test error!

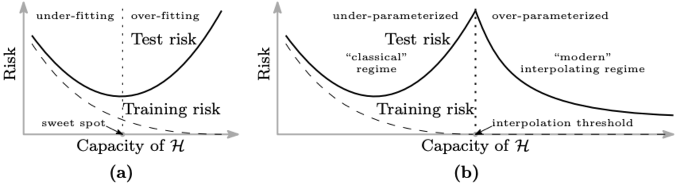
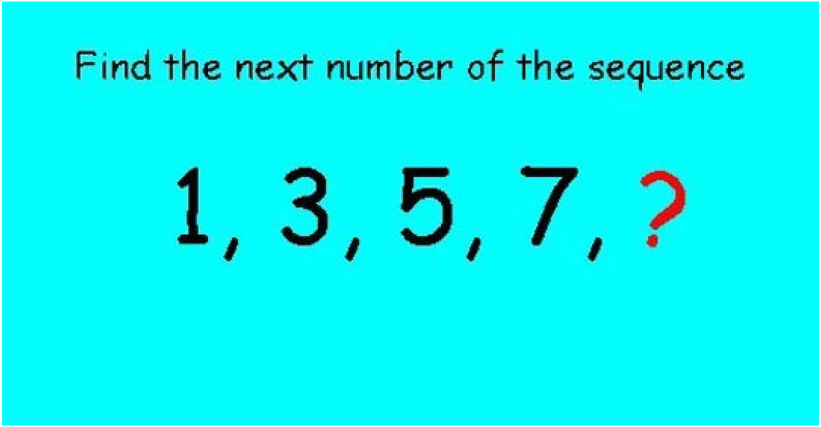


Figure 1: **Curves for training risk (dashed line) and test risk (solid line).** (**a**) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (**b**) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high capacity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

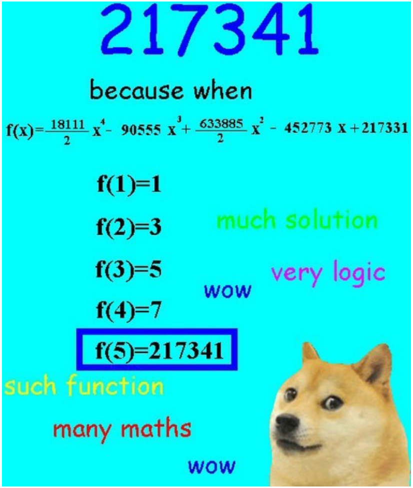
Belkin et al. Reconciling modern machine learning practice and the bias-variance trade-off

Training vs Testing Error



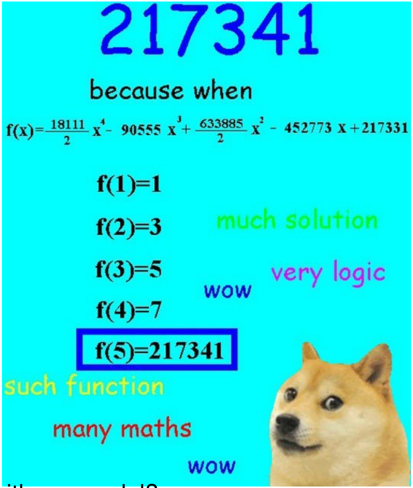
- training data = $\{(1,1),(2,3),(3,5),(5,7)\}$,
- $\hat{f}(i) = i$th odd number $= 2(i-1)+1$. (linear model)
- Perfect training MSE

Training vs Testing Error



- testing data = $\{(5, 27341)\}$
- Testing MSE = $(9-27341)^2 = 747{,}038{,}224$

Training vs Testing Error



What went wrong with our model?

A. High bias

B. High variance

Bias-Variance Tradeoff

$$\underbrace{\mathbb{E}(\hat{f}(x) - y)^2}_{\text{test error}} = \underbrace{\left(\mathbb{E}\hat{f}(x) - f(x)\right)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}(\hat{f}(x))}_{\text{variance}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible error}}$$

- Less flexible model:

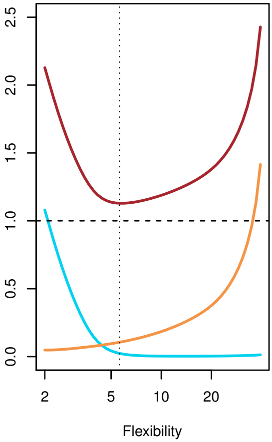
$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

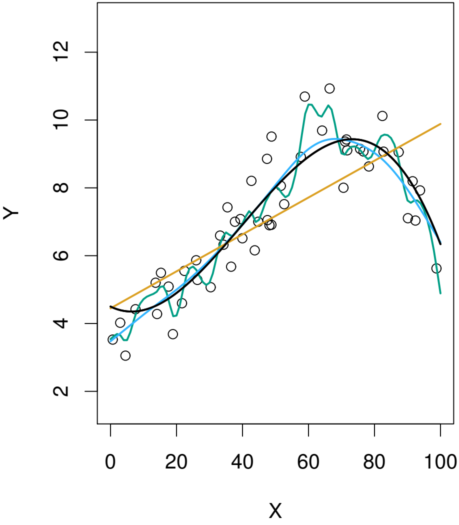
bias high, variance low

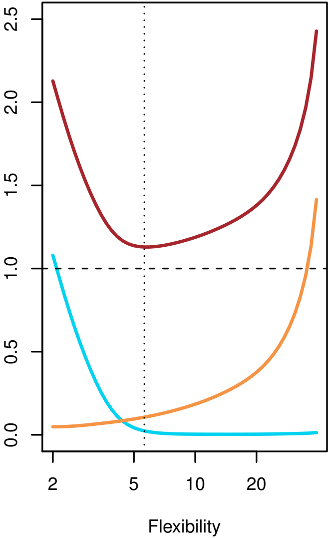
- More flexible model:

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3$$

bias low, variance high







Variance	$\mathrm{Var}(\hat{f}(x))$	how much $\hat{f}$ changes when fitted using different datasets
Bias	$\mathbb{E}\hat{f}(x) - f(x)$	difference b/w the true function f and the expectation of the model $\hat{f}$

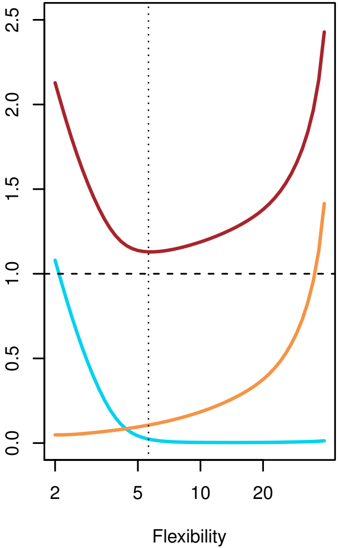
Bias-Variance Tradeoff

$$\text{test error} = \text{bias}^2 + \text{variance} + c$$

Bias: Large if model is not flexible enough

Variance: Large if model is too flexible

- True for regression and classification



What to try next?

You have a training set with one predictor variable and one response variable. You fit a model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

and get perfect training error. On the the other hand, you are astonished to find out that your model performs really poorly on the test data. Which model should you try next?

Choose one:

A. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

B. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$

What to try next?

You have a training set with one predictor variable and one response variable. You fit a model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

and get perfect training error. On the the other hand, you are astonished to find out that your model performs really poorly on the test data. Which model should you try next?

Choose one:

A. $Y = \beta_0 + \beta_1 \exp(X)$

B. $Y = \beta_0 + \beta_1 \log(X)$

C. ?????

Cross-validation

- Want to minimize test error

(by using the right amount of model flexibility)

- Test error is unknown

(when we fit the model)

Cross-validation

- Want to minimize test error

(by using the right amount of model flexibility)

- Test error is unknown

(when we fit the model)

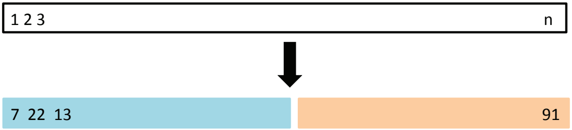
Cross-Validation: a way to <u>estimate</u> the test error

The Validation Set Approach

(ISLR Sec 5.1.1)

(Randomly) split dataset into two subsets:

- A training set: to fit the model $\hat{f}$
- A validation set (aka hold-out set): to estimate the test error



A Simple Approach

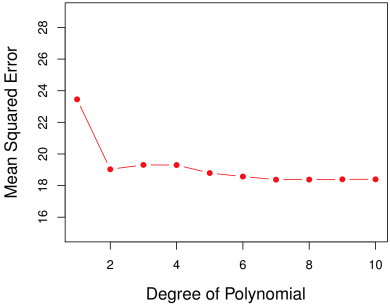
Example (ISLR Sec 5.3.1): `Auto` dataset

```
> train = sample(392,196)

> lm.fit = lm(mpg~horsepower, data=Auto, subset=train)
> mean( (mpg - predict(lm.fit, Auto))[-train]^2 )

> lm.fit2 = lm(mpg~poly(horsepower,2), data=Auto, subset=train)
> mean( (mpg - predict(lm.fit2, Auto))[-train]^2 )
```

A Simple Approach

Example (ISLR Sec 5.3.1): `Auto` dataset

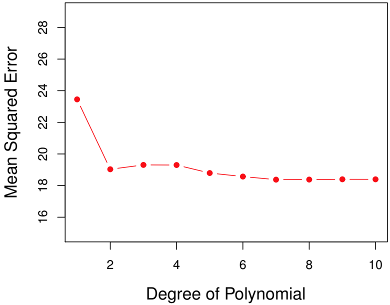
```
> train = sample(392,196)

> lm.fit = lm(mpg~horsepower, data=Auto, subset=train)
> mean( (mpg - predict(lm.fit, Auto))[-train]^2 )

> lm.fit2 = lm(mpg~poly(horsepower,2), data=Auto, subset=train)
> mean( (mpg - predict(lm.fit2, Auto))[-train]^2 )
```

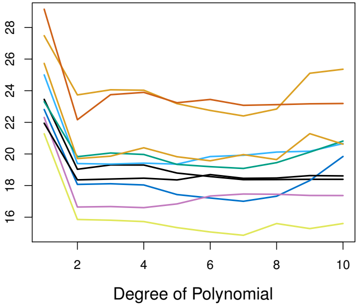


A Simple Approach

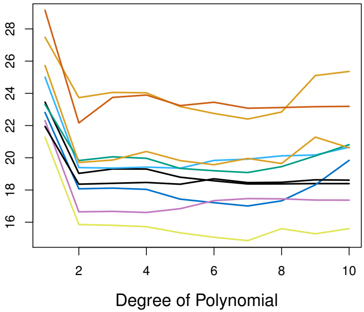
Example (ISLR Sec 5.3.1): `Auto` dataset

```
> train = sample(392,196)

> lm.fit = lm(mpg~horsepower, data=Auto, subset=train)
> mean( (mpg - predict(lm.fit, Auto))[-train]^2 )

> lm.fit2 = lm(mpg~poly(horsepower,2), data=Auto, subset=train)
> mean( (mpg - predict(lm.fit2, Auto))[-train]^2 )
```





The Validation Set Approach



Issues:

- Estimate of test error is highly variable
- Wasteful! Use only half of the data to fit models

Validation set approach

Suppose you fit a model $\hat{f}$ using the whole training set.

On average, would you expect the validation set approach to overestimate or to underestimate the testing error $\mathbb{E}(\hat{f}(x) - f(x))^2$?

A. overestimate

B. underestimate

C. unclear

Validation set approach

Suppose you fit a model $\hat{f}$ using the whole training set.

On average, would you expect the validation set approach to overestimate or to underestimate the testing error $\mathbb{E}(\hat{f}(x) - f(x))^2$?

A. overestimate

B. underestimate

C. unclear

Leave-One-Out Cross-Validation (LOOCV) (ISLR 5.1.2)

Leave-One-Out Cross-Validation (LOOCV) (ISLR 5.1.2)

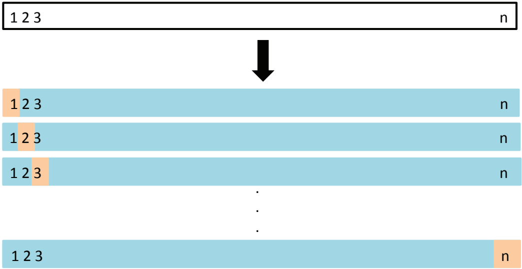
- Split n data points into:
  - a training set of $n - 1$ points
  - a validation set of 1 point

Leave-One-Out Cross-Validation (LOOCV) (ISLR 5.1.2)

- Split n data points into:
  - a training set of $n-1$ points
  - a validation set of 1 point
- Consider all n possible ways of splitting

Estimate test error by averaging:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{MSE}_i}_{\text{Error on Sample } i}$$



LOOCV

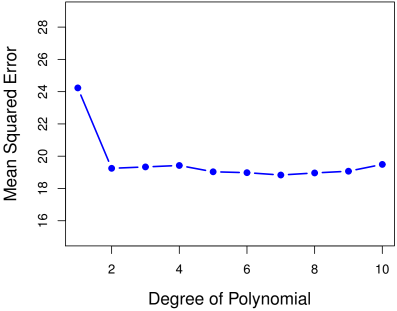
Example (ISLR Sec 5.3.1): `Auto` dataset

```
> library(boot)
> for (i in 1:5){
+     glm.fit = glm(mpg∼poly(horsepower, i), data=Auto)
+     cv.error[i] = cv.glm(Auto, glm.fit)$delta[1]
+ }
> cv.error
[1] 24.23 19.25 19.33 19.42 19.03
```

LOOCV

Example (ISLR Sec 5.3.1): `Auto` dataset

```
> library(boot)
> for (i in 1:5){
+     glm.fit = glm(mpg~poly(horsepower, i), data=Auto)
+     cv.error[i] = cv.glm(Auto, glm.fit)$delta[1]
+ }
> cv.error
[1] 24.23 19.25 19.33 19.42 19.03
```

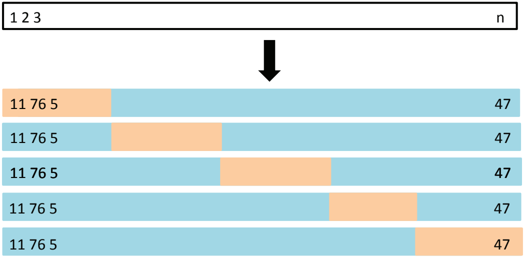


k-Fold Cross-Validation (ISLR Sec 5.1.3)

- (Randomly) split dataset into k subsets (folds) of equal size:
- Use $(k-1)$ subsets to fit model
- Use the remaining 1 subset as validation set

Get k estimates of test error; average:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \underbrace{\mathrm{MSE}_i}_{\text{Error on block } i}$$



k-fold cross validation

Suppose that the number of training samples, denoted by n, is a multiple of k. Then in each “round" of k-fold cross validation, your training set has size:

A. n/k

B. $n - n/k$

C. $(n - n/k) - 1$

D. $k - 1$

k-Fold Cross-Validation (ISLR Sec 5.1.3)

- (Randomly) split dataset into k subsets (folds) of equal size:
- Use $(k-1)$ subsets to fit model
- Use the remaining 1 subset as validation set

Get k estimates of test error; average:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

k-Fold Cross-Validation (ISLR Sec 5.1.3)

- (Randomly) split dataset into k subsets (folds) of equal size:
- Use $(k-1)$ subsets to fit model
- Use the remaining 1 subset as validation set

Get k estimates of test error; average:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

- LOOCV: equivalent to $k = n$

k-Fold Cross-Validation (ISLR Sec 5.1.3)

- (Randomly) split dataset into k subsets (folds) of equal size:
- Use $(k-1)$ subsets to fit model
- Use the remaining 1 subset as validation set

Get k estimates of test error; average:

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i$$

- LOOCV: equivalent to $k = n$
- Typically $k = 5$ or $k = 10$

k-Fold Cross-Validation (ISLR Sec 5.1.3)

- (Randomly) split dataset into k subsets (folds) of equal size:
- Use $(k-1)$ subsets to fit model
- Use the remaining 1 subset as validation set

Get k estimates of test error; average:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

- LOOCV: equivalent to $k = n$
- Typically $k = 5$ or $k = 10$
- Faster than LOOCV

k-Fold Cross-Validation

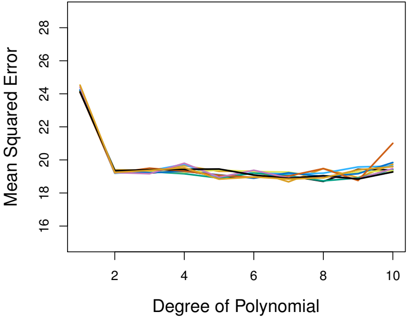
Example (ISLR Sec 5.3.3): `Auto` dataset

```
> for (i in 1:10){
+     glm.fit = glm(mpg~poly(horsepower, i), data=Auto)
+     cv.error[i] = cv.glm(Auto, glm.fit, K=10)$delta[1]
+ }
> cv.error
[1] 24.21 19.19 19.31 19.34 18.88 19.02 18.90 19.71 18.95 19.50
```

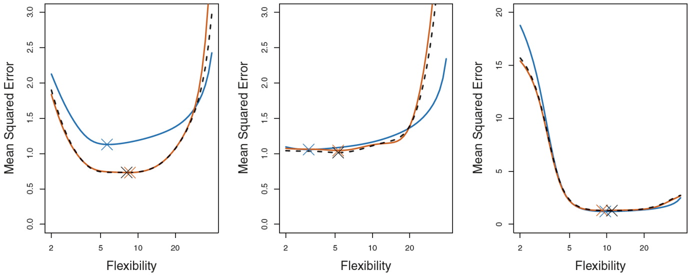
k-Fold Cross-Validation

Example (ISLR Sec 5.3.3): `Auto` dataset

```
> for (i in 1:10){
+     glm.fit = glm(mpg~poly(horsepower, i), data=Auto)
+     cv.error[i] = cv.glm(Auto, glm.fit, K=10)$delta[1]
+ }
> cv.error
[1] 24.21 19.19 19.31 19.34 18.88 19.02 18.90 19.71 18.95 19.50
```



k-Fold Cross-Validation



10 fold CV and MSE and LOOCV

Suppose we want to apply cross validation to the following estimation procedure:

1. **Stage 1:** Fit a linear regression model with full set of p predictors.
2. **Stage 2:** Throw out the predictors with the r smallest values (in magnitude).
3. **Stage 3:** Fit a linear regression model with the smaller set of $p - r$ predictors.

We wish to use cross validation to help us choose which r produces the best test error, with r ranging from 1 to p.

True or False: A correctly implemented cross validation procedure will possibly choose a different set of r predictors for every fold in the CV procedure.

A. True

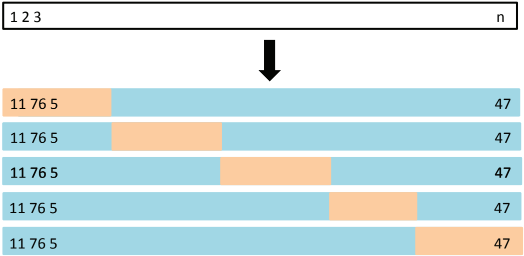
B. False

Cross-Validation on Classification Problems

(ISLR Sec 5.1.5)

Works in the exact same way:

$$\mathrm{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} (\text{Misclassification rate in the } i\text{-th split})$$



Example: estimate test errors for

- Logistic regression with different high order terms, or
- K Nearest Neighbor for different K (not the same k above)

Summary

test error $=$ bias^2 $+$ variance

- Bias decreases with model flexibility
- Variance increases with model flexibility

Summary

test error $=$ bias^2 $+$ variance

- Bias decreases with model flexibility
- Variance increases with model flexibility

Cross-Validation: A general way to estimate test error

- Validation set approach: One subset for training, the rest for validation
- LOOCV: $(n-1)$ points for training, 1 point for validation
- k-Fold CV: Split into k folds; $(k-1)$ folds for training, 1 for validation

Summary

test error $=$ bias^2 $+$ variance

- Bias decreases with model flexibility
- Variance increases with model flexibility

Cross-Validation: A general way to estimate test error

- Validation set approach: One subset for training, the rest for validation
- LOOCV: $(n-1)$ points for training, 1 point for validation
- k-Fold CV: Split into k folds; $(k-1)$ folds for training, 1 for validation

Choosing flexibility:

- Estimate test errors for models of different flexibility
- Pick the one with lowest error

(Optional) A popular interview question

An unknown quantity:

$$\theta = \text{test error}$$

(Optional) A popular interview question

An unknown quantity:

$$\theta = \text{test error}$$

Can be estimated by *k*-fold CV:

$$\hat{\theta} = \mathsf{CV}_{(k)}$$

(Optional) A popular interview question

An unknown quantity:

$$\theta = \text{test error}$$

Can be estimated by *k*-fold CV:

$$\hat{\theta} = \mathsf{CV}_{(k)}$$

Bias-variance equation applies here!

$$\underbrace{\mathbb{E}(\hat{\theta} - \theta)^2}_{\text{error in estimating the test error}} \quad = \quad \underbrace{\left(\mathbb{E}\hat{\theta} - \theta\right)^2}_{\text{bias}^2 \text{ of } \mathsf{CV}_{(k)}} \quad + \quad \underbrace{\mathsf{Var}(\hat{\theta})}_{\text{variance of } \mathsf{CV}_{(k)}}$$

(Optional) A popular interview question

An unknown quantity: $\theta =$ test error

Estimated by *k*-fold CV: $\hat{\theta} = \text{CV}_{(k)}$

Bias-variance equation applies here!

$$\underbrace{\mathbb{E}(\hat{\theta}-\theta)^2}_{\text{error in estimating the test error}} \quad = \quad \underbrace{\left(\mathbb{E}\hat{\theta}-\theta\right)^2}_{\text{bias}^2 \text{ of CV}_{(k)}} \quad + \quad \underbrace{\text{Var}(\hat{\theta})}_{\text{variance of CV}_{(k)}}$$

(Optional) A popular interview question

An unknown quantity: $\theta =$ test error

Estimated by k-fold CV: $\hat{\theta} = \mathrm{CV}_{(k)}$

Bias-variance equation applies here!

$$\underbrace{\mathbb{E}(\hat{\theta} - \theta)^2}_{\text{error in estimating the test error}} = \underbrace{\left(\mathbb{E}\hat{\theta} - \theta\right)^2}_{\text{bias}^2 \text{ of } \mathrm{CV}_{(k)}} + \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{variance of } \mathrm{CV}_{(k)}}$$

Question: If we increase k, how will the bias and variance change?

A. Bias $\nearrow$, variance $\searrow$.

B. Bias $\searrow$, variance $\nearrow$.

C. They both stay the same.

(Optional) A popular interview question

An unknown quantity: $\theta =$ test error

Estimated by k-fold CV: $\hat{\theta} = \mathrm{CV}_{(k)}$

Bias-variance equation applies here!

$$\underbrace{\mathbb{E}(\hat{\theta} - \theta)^2}_{\text{error in estimating the test error}} = \underbrace{\left(\mathbb{E}\hat{\theta} - \theta\right)^2}_{\text{bias}^2 \text{ of } \mathrm{CV}_{(k)}} + \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{variance of } \mathrm{CV}_{(k)}}$$

Question: If we increase k, how will the bias and variance change?

A. Bias $\nearrow$, variance $\searrow$.

B. Bias $\searrow$, variance $\nearrow$.

C. They both stay the same.

Answer: B; see ISLR Sec 5.1.4

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
Slides based on Yudong Chen's Slides.
Some images due to Machine learning @ Berkeley Group