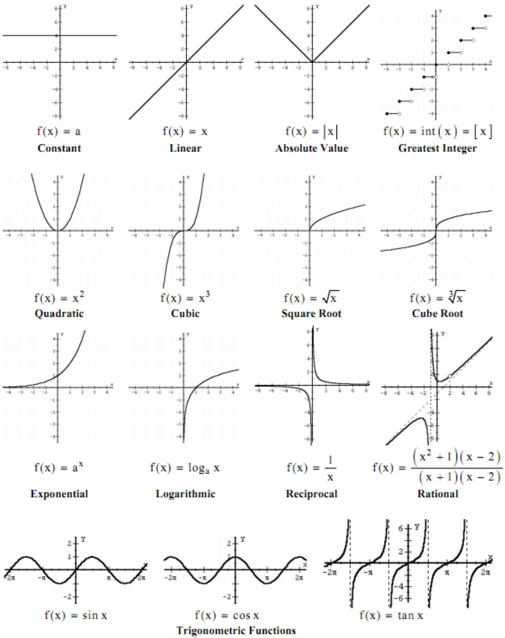
Regularization/Shrinkage

Schwartz

November 8, 2017

Functions



f(x) = a
Constant
f(x) = x
Linear
f(x) = |x|
Absolute Value
f(x) = int(x) = [x]
Greatest Integer
f(x) = x²
Quadratic
f(x) = x³
Cubic
f(x) = √x
Square Root
f(x) = ∛x
Cube Root
f(x) = aˣ
Exponential
f(x) = logₐ x
Logarithmic
f(x) = 1/x
Reciprocal
f(x) = (x² + 1)(x − 2) / (x + 1)(x − 2)
Rational
f(x) = sin x
f(x) = cos x
f(x) = tan x
Trigonometric Functions

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
   - Regression: predict a real-valued outcome
   - Classification: predict a categorical class

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
   - Regression: predict a real-valued outcome
   - Classification: predict a categorical class

1. Build a classification or regression model

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
    - Regression: predict a real-valued outcome
    - Classification: predict a categorical class

1. Build a classification or regression model
    a. Identify and collect relevant features *and associated outcomes*
    b. Fit model $\iff$ *capture features and outcomes associations*

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
   - Regression: predict a real-valued outcome
   - Classification: predict a categorical class

1. Build a classification or regression model
   a. Identify and collect relevant features *and associated outcomes*
   b. Fit model $\iff$ *capture features and outcomes associations*

2. Assess performance

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
   - Regression: predict a real-valued outcome
   - Classification: predict a categorical class

1. Build a classification or regression model
   a. Identify and collect relevant features *and associated outcomes*
   b. Fit model $\iff$ *capture features and outcomes associations*

2. Assess performance
   - Iterate model/feature specification phase if insufficient
   - Deploy if sufficient

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
   - Regression: predict a real-valued outcome
   - Classification: predict a categorical class

1. Build a classification or regression model
   a. Identify and collect relevant features *and associated outcomes*
   b. Fit model $\iff$ *capture features and outcomes associations*

2. Assess performance
   - Iterate model/feature specification phase if insufficient
   - Deploy if sufficient

So this works if features are available *before* prediction is needed...

What is the *Predictive Modeling Workflow?*

0. You want to make a decision based on an informed guess
   - Regression: predict a real-valued outcome
   - Classification: predict a categorical class

1. Build a classification or regression model
   a. Identify and collect relevant features *and associated outcomes*
   b. Fit model $\iff$ *capture features and outcomes associations*

2. Assess performance
   - Iterate model/feature specification phase if insufficient
   - Deploy if sufficient

So this works if features are available *before* prediction is needed...

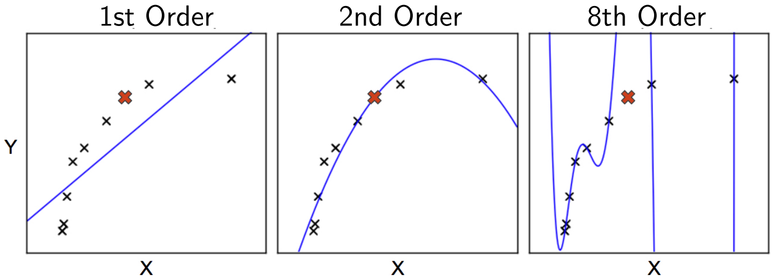
Have X to guess $\hat{Y}$ to help decision making

Model Complexity

1. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1$
2. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1 + x^2\hat{\beta}_2$

   $\vdots$

3. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1 + x^2\hat{\beta}_2 + x^3\hat{\beta}_3 + x^4\hat{\beta}_4 + x^5\hat{\beta}_5 + x^6\hat{\beta}_6 + x^7\hat{\beta}_7 + x^8\hat{\beta}_8$

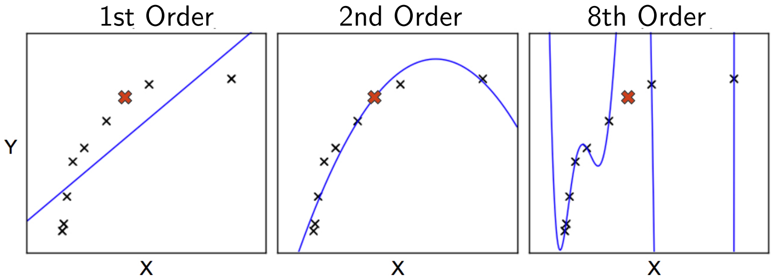


Model Complexity

1. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1$
2. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1 + x^2\hat{\beta}_2$

$\vdots$

3. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1 + x^2\hat{\beta}_2 + x^3\hat{\beta}_3 + x^4\hat{\beta}_4 + x^5\hat{\beta}_5 + x^6\hat{\beta}_6 + x^7\hat{\beta}_7 + x^8\hat{\beta}_8$



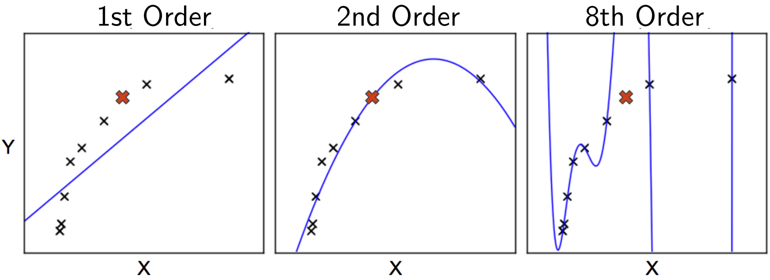
Model fit to the data always improves until *perfect* data fit

Model Complexity

1. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1$
2. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1 + x^2\hat{\beta}_2$

   $\vdots$

3. $\hat{Y} = \hat{\beta}_0 + x\hat{\beta}_1 + x^2\hat{\beta}_2 + x^3\hat{\beta}_3 + x^4\hat{\beta}_4 + x^5\hat{\beta}_5 + x^6\hat{\beta}_6 + x^7\hat{\beta}_7 + x^8\hat{\beta}_8$



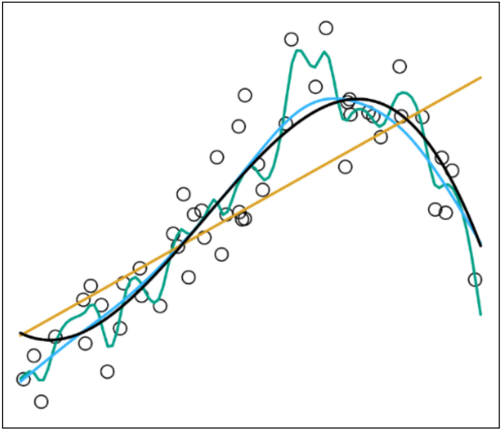
Model fit to the data always improves until *perfect* data fit

R^2 can only get better – never worse – with more features

Variance and Bias

<u>Variance</u>: (1) the volatility of a model prediction from data set to data set; (2) the amount of flexibility/susceptibility the model has to being influenced by idiosyncratic outliers

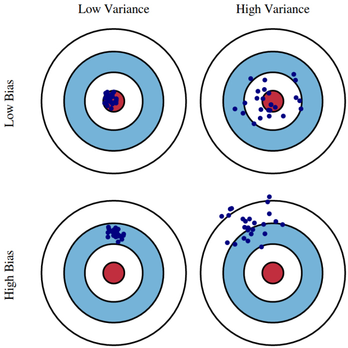
<u>Bias</u>: the rigidity/inability/limitations of the model to flexibly capture complex but true data associations

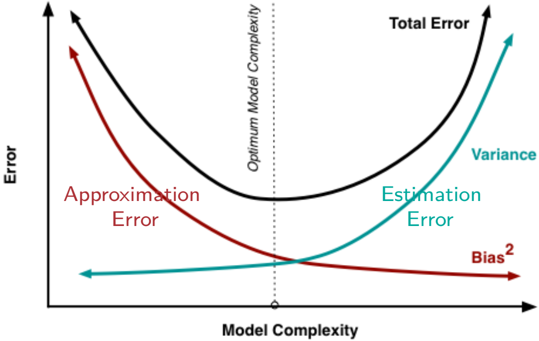


Bias and *Variance* characterize models *robustness* – a neutral word

Variance and Bias and Tradeoff

- In *Machine Learning*, *bias* and *variance* refer to performance accuracy characteristics over hypothetical random data sets

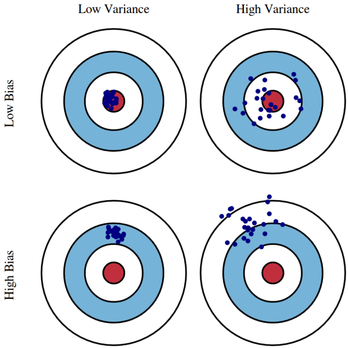


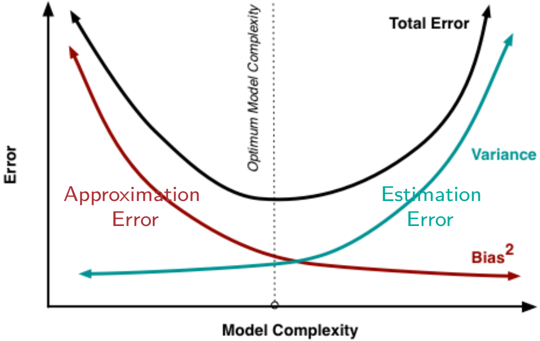


Variance and Bias and Tradeoff

- In *Machine Learning*, *bias* and *variance* refer to performance accuracy characteristics over hypothetical random data sets

  Model Complexity: too simple=bias & too flexible=variance



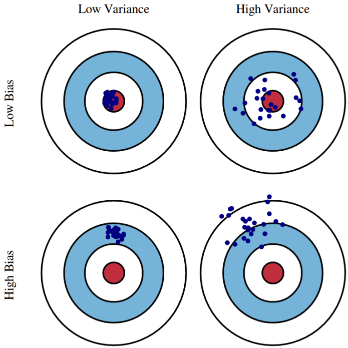


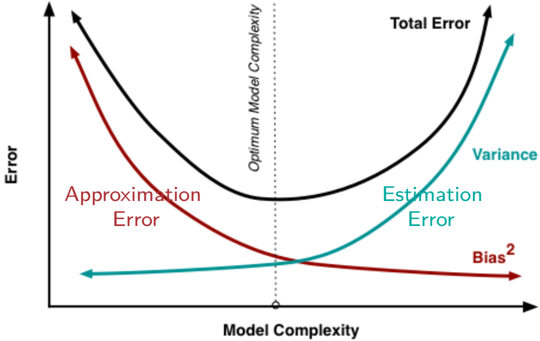
Variance and Bias and Tradeoff

- In *Machine Learning*, *bias* and *variance* refer to performance accuracy characteristics over hypothetical random data sets

  Model Complexity: too simple=bias & too flexible=variance

  The Machine Learning objective is finding the right balance





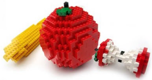
Partition of Variation

Let $y_i = \theta + \epsilon_i$ with $\theta = f(x_0)$ and $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$
For estimator $\hat{\theta} = \hat{f}(x_0)$,

$$
\begin{aligned}
MSE &= \frac{1}{n}\sum_i \left(y_i - \hat{\theta}\right)^2 \approx \mathrm{E}\left[\left(y_i - \hat{\theta}\right)^2\right] \\
&= \mathrm{E}\left[\left(y_i - \theta + \theta - E[\hat{\theta}] + E[\hat{\theta}] - \hat{\theta}\right)^2\right] \\
&= \mathrm{E}\left[\left((y_i - \theta) + (\theta - E[\hat{\theta}]) + (E[\hat{\theta}] - \hat{\theta})\right)^2\right] \\
&\overset{!}{=} \mathrm{E}\left[(y_i - \theta)^2\right] + \mathrm{E}\left[\left(\theta - E[\hat{\theta}]\right)^2\right] + \mathrm{E}\left[\left(E[\hat{\theta}] - \hat{\theta}\right)^2\right] \\
&= \sigma_\epsilon^2 + \left(\mathrm{E}[\hat{\theta}] - \theta\right)^2 + \sigma_{\hat{\theta}}^2 \\
&= \text{Residual Variance} + \text{Model Bias}^2 + \text{Model Variance}
\end{aligned}
$$

Model Generalizability

- The question isn't how well does the model fit data you have
  - It can fit data you have exactly as closely as you want


(and once you've got the model you shake off the data and just use the model)

Model Generalizability

- The question isn't how well does the model fit data you have
  - It can fit data you have exactly as closely as you want

(and once you've got the model you shake off the data and just use the model)

- The question is how well does your model *generalize*?

  i.e., how well will it perform "in the wild"

Model Generalizability

- The question isn't how well does the model fit data you have
  - It can fit data you have exactly as closely as you want

(and once you've got the model you shake off the data and just use the model)

- The question is how well does your model *generalize*?

  i.e., how well will it perform "in the wild"

- **How will you be able to know how well you'll to do??**

Model Generalizability

- The question isn't how well does the model fit data you have
  - It can fit data you have exactly as closely as you want

(and once you've got the model you shake off the data and just use the model)

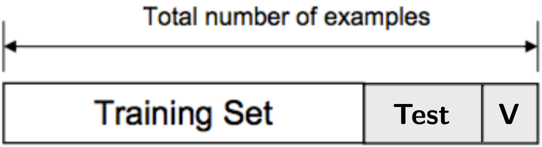
- The question is how well does your model *generalize*?

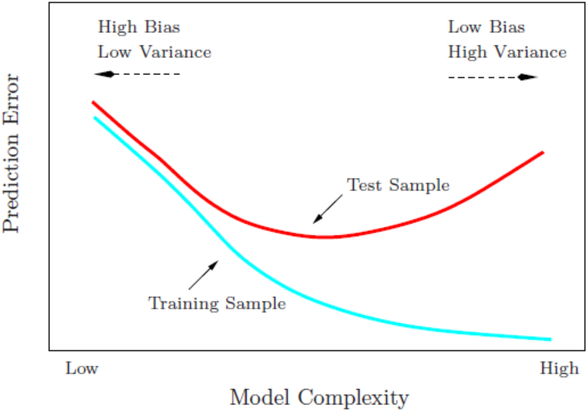
  i.e., how well will it perform "in the wild"

- **How will you be able to know how well you'll to do??**

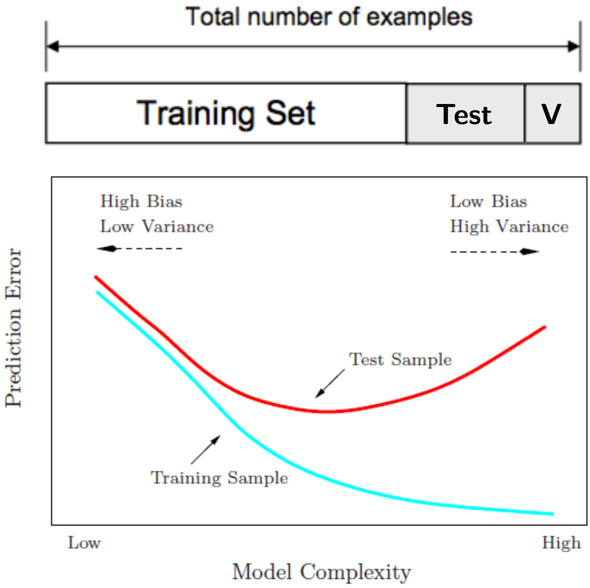
Train/Test split

The Train/Test Split



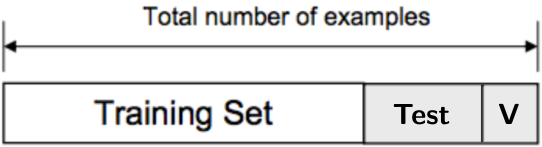


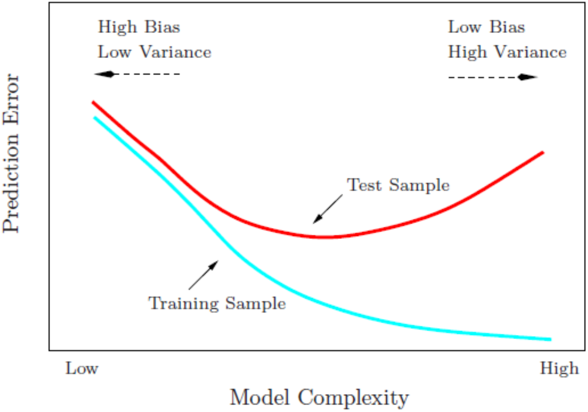
The Train/Test Split



- How do we choose?
- Sampling Variability?
- Why do we need V?

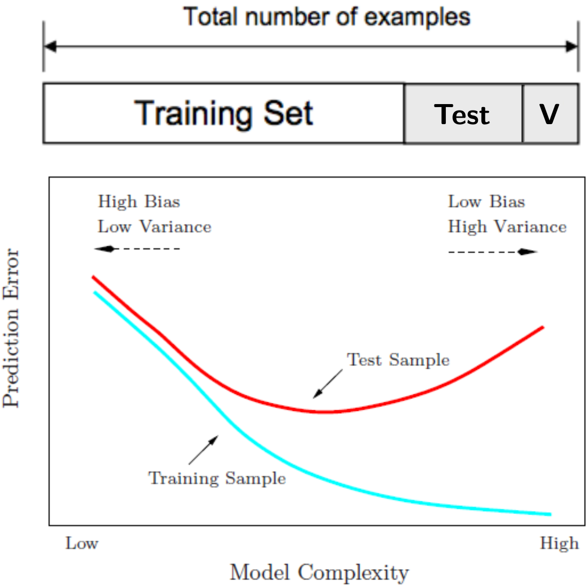
The Train/Test Split





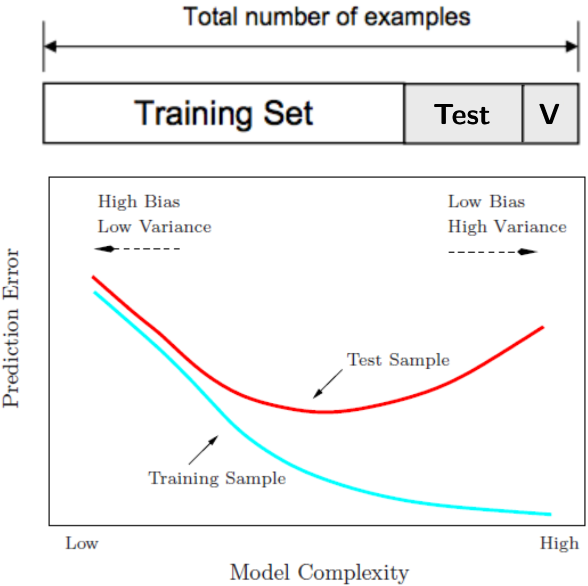
- How do we choose?
  Widening gap means less generalizability
- Sampling Variability?
- Why do we need V?

The Train/Test Split



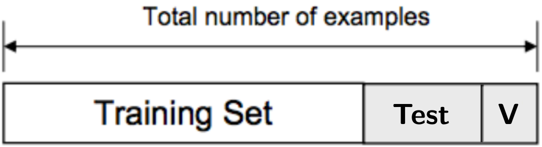
- How do we choose?
  Widening gap means less generalizability
  Minimum test error means best prediction
- Sampling Variability?
- Why do we need V?

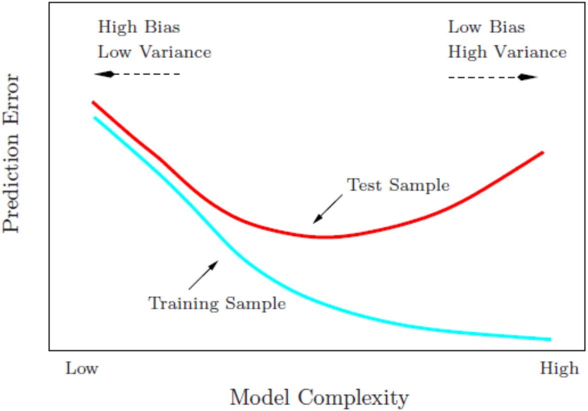
The Train/Test Split



- How do we choose?
  Widening gap means less generalizability
  Minimum test error means best prediction
- Sampling Variability?
  Occam's razor + this is one train/test split...
- Why do we need V?

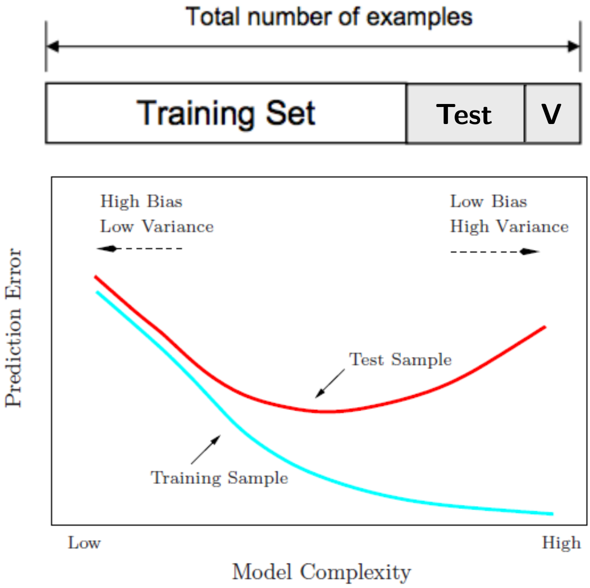
The Train/Test Split





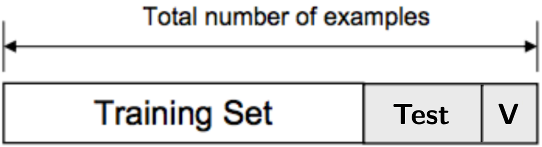
- How do we choose?
  Widening gap means less generalizability
  Minimum test error means best prediction
- Sampling Variability?
  Occam's razor + this is one train/test split...
  <Hold this thought>
- Why do we need V?

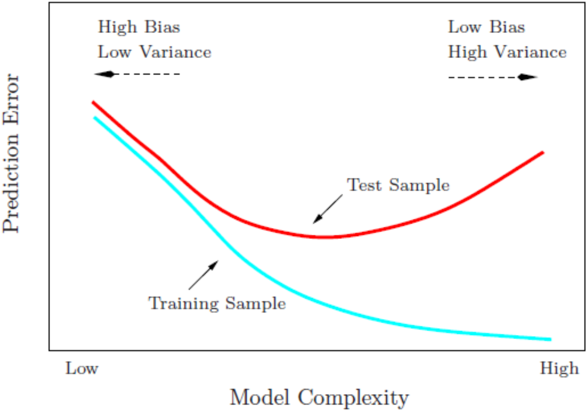
The Train/Test Split



- How do we choose?
  Widening gap means less generalizability
  Minimum test error means best prediction
- Sampling Variability?
  Occam's razor + this is one train/test split...
  <Hold this thought>
- Why do we need V?
  Complexity choice is "fit" from test data

The Train/Test Split





- How do we choose?
  Widening gap means less generalizability
  Minimum test error means best prediction
- Sampling Variability?
  Occam's razor + this is one train/test split...
  <Hold this thought>
- Why do we need V?
  Complexity choice is "fit" from test data
  The *validation* set V actually "tests wild"

Quiz

What is the “prediction error” on the previous slide?
(let’s have two regression and two classification examples)

Quiz

What is the "prediction error" on the previous slide?

(let's have two regression and two classification examples)

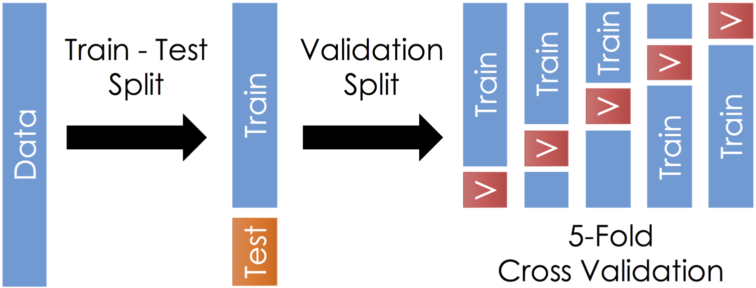
$$RMSE = \sqrt{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}$$

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)(\hat{Y}_i - \bar{\hat{Y}}_i)^2}{\hat{\sigma}_Y \hat{\sigma}_{\hat{Y}}}$$

$$Accuracy = \frac{\sum_{i=1}^{n} 1_{[Y_i = \hat{Y}_i]}}{n}$$

$$Sensitivity = \frac{\sum_{i=1}^{n} 1_{[Y_i = \hat{Y}_i]} 1_{[Y_i = 1]}}{\sum_{i=1}^{n} 1_{[Y_i = 1]}}$$

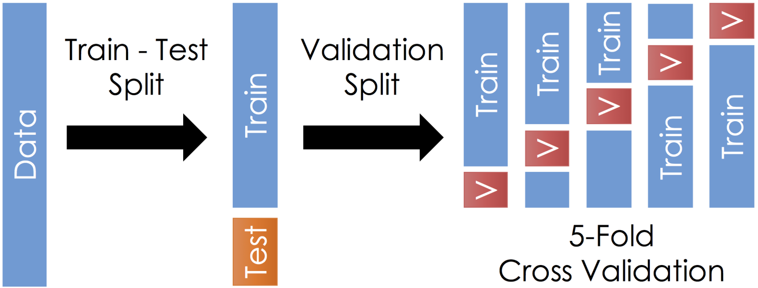
K-Folds Cross Validation



Benefits?

Challenges?

K-Folds Cross Validation

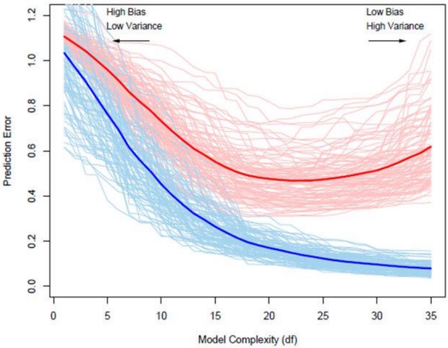


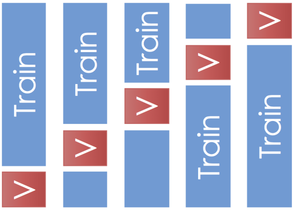
Benefits?

- Uses all the data as "validation" set

Challenges?

K-Folds Cross Validation





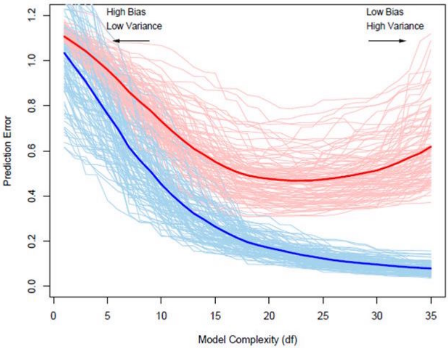
5-Fold
Cross Validation

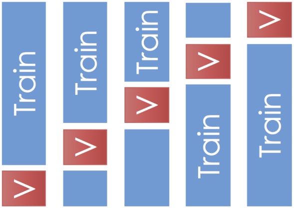
Benefits?

- Uses all the data as "validation" set
- Shows variation in sample accuracy scores

Challenges?

K-Folds Cross Validation





5-Fold
Cross Validation

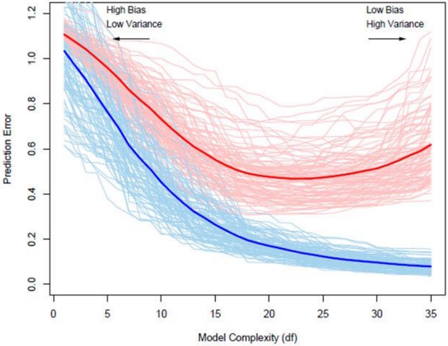
Benefits?

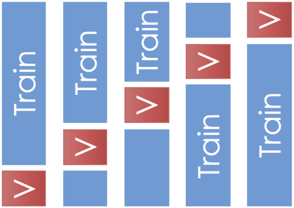
- Uses all the data as "validation" set
- Shows variation in sample accuracy scores

Challenges?

- I just fit K models... which do I use?

K-Folds Cross Validation





5-Fold
Cross Validation

Benefits?

- Uses all the data as “validation” set
- Shows variation in sample accuracy scores

Challenges?

- I just fit K models… which do I use?
- Refit with all data at a generalizable complexity level

Overfitting

1. How do you make this simpler/more complex?

$$\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + \cdots$$

Overfitting

1. How do you make this simpler/more complex?

$$\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + \cdots$$

2. What does "dampening"/"suppressing" the β_j towards 0 do?

Overfitting

1. How do you make this simpler/more complex?

$$\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + \cdots$$

2. What does "dampening"/"suppressing" the β_j towards 0 do?

3. Why might we like this?

Overfitting

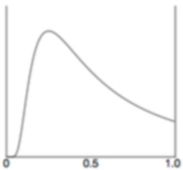
1. How do you make this simpler/more complex?

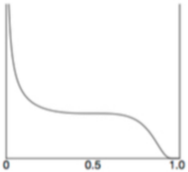
$$\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + \cdots$$

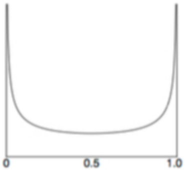
2. What does "dampening"/"suppressing" the β_j towards 0 do?

3. Why might we like this?

4. Suppose every β_j was 1...
   what do you think about the following *shrinkage* profiles?







Bias/Variance Tradeoff