

Bias-Variance Tradeoff

David Dalpiaz

STAT 430, Fall 2017

Announcements

- Homework 03 released
- Regrade policy
- Style policy?

- Supervised Learning
 - **Regression**
 - Parametric
 - Non-Parametric
 - Classification
- Unsupervised Learning

Regression Setup

Given a random pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$. We would like to “predict” Y with some function of X , say, $f(X)$.

Define the **squared error loss** of estimating Y using $f(X)$ as

$$L(Y, f(X)) \triangleq (Y - f(X))^2$$

We call the expected loss the **risk** of estimating Y using $f(X)$

$$R(Y, f(X)) \triangleq \mathbb{E}[L(Y, f(X))] = \mathbb{E}_{X,Y}[(Y - f(X))^2]$$

Minimizing Risk

After conditioning on X

$$\mathbb{E}_{X,Y} [(Y - f(X))^2] = \mathbb{E}_X \mathbb{E}_{Y|X} [(Y - f(X))^2 \mid X = x]$$

We see that the risk is minimized by the conditional mean

$$f(x) = \mathbb{E}(Y \mid X = x)$$

We call this, the **regression function**.

Given data

$$\mathcal{D} = (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$$

our goal is to find some \hat{f} that is a good estimate of the regression function f .

Expected Prediction Error

$$\text{EPE}(Y, \hat{f}(X)) \triangleq \mathbb{E}_{X, Y, \mathcal{D}} \left[(Y - \hat{f}(X))^2 \right]$$

Reducible and Irreducible Error

$$\begin{aligned} \text{EPE}(Y, \hat{f}(x)) &= \mathbb{E}_{Y|X, \mathcal{D}} \left[(Y - \hat{f}(X))^2 \mid X = x \right] \\ &= \mathbb{E}_{Y|X, \mathcal{D}} \left[(Y - \hat{f}(X))^2 \mid X = x \right] \\ &= \underbrace{\mathbb{E}_{\mathcal{D}} \left[(f(x) - \hat{f}(x))^2 \right]}_{\text{reducible error}} + \underbrace{\mathbb{V}_{Y|X} [Y \mid X = x]}_{\text{irreducible error}} \end{aligned}$$

Bias and Variance

Recall the definition of the **bias** of an estimator.

$$\text{bias}(\hat{\theta}) \triangleq \mathbb{E} [\hat{\theta}] - \theta$$

Also recall the definition of the **variance** of an estimator.

$$\mathbb{V}(\hat{\theta}) = \text{var}(\hat{\theta}) \triangleq \mathbb{E} [(\hat{\theta} - \mathbb{E} [\hat{\theta}])^2]$$

Bias and Variance

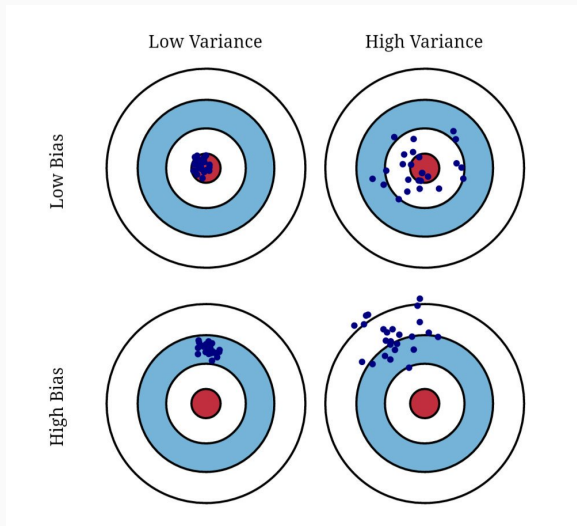


Figure 1: Dartboard Analogy of Bias and Variance

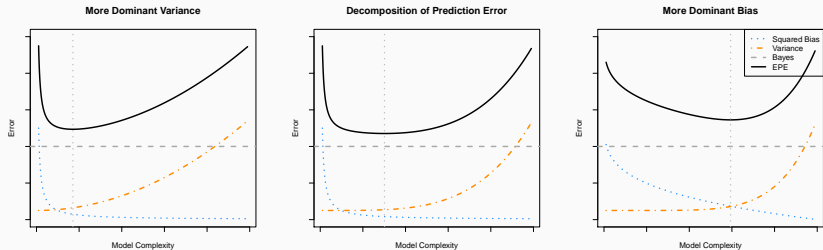
Bias-Variance Decomposition

$$\begin{aligned}\text{MSE}(f(x), \hat{f}(x)) &= \mathbb{E}_{\mathcal{D}} \left[(f(x) - \hat{f}(x))^2 \right] \\ &= \underbrace{(f(x) - \mathbb{E}[\hat{f}(x)])^2}_{\text{bias}^2(\hat{f}(x))} + \mathbb{E} \left[\underbrace{(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2}_{\text{var}(\hat{f}(x))} \right]\end{aligned}$$

Bias-Variance Decomposition

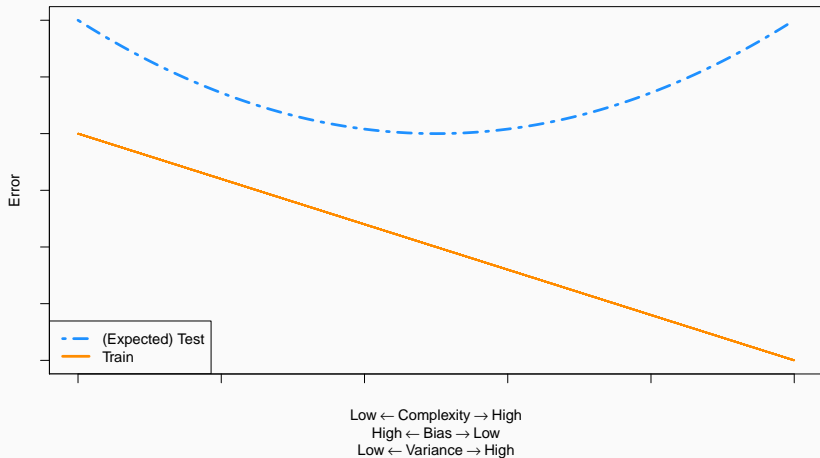
$$\text{MSE} \left(f(x), \hat{f}(x) \right) = \text{bias}^2 \left(\hat{f}(x) \right) + \text{var} \left(\hat{f}(x) \right)$$

Bias-Variance Decomposition



Expected Test Error

Error versus Model Complexity



Simulation Study, Regression Function

We will illustrate these decompositions, most importantly the bias-variance tradeoff, through simulation. Suppose we would like to train a model to learn the true regression function function

$$f(x) = x^2$$

```
f = function(x) {  
  x ^ 2  
}
```

Simulation Study, Regression Function

More specifically, we'd like to predict an observation, Y , given that $X = x$ by using $\hat{f}(x)$ where

$$\mathbb{E}[Y | X = x] = f(x) = x^2$$

and

$$\mathbb{V}[Y | X = x] = \sigma^2.$$

Simulation Study, Data Generating Process

To carry out a concrete simulation example, we need to fully specify the **data generating process**. We do so with the following R code.

```
get_sim_data = function(f, sample_size = 100) {  
  x = runif(n = sample_size, min = 0, max = 1)  
  y = rnorm(n = sample_size, mean = f(x), sd = 0.3)  
  data.frame(x, y)  
}
```

Simulation Study, Models

Using this setup, we will generate datasets, \mathcal{D} , with a sample size $n = 100$ and fit four models.

$$\text{predict}(\text{fit0}, x) = \hat{f}_0(x) = \hat{\beta}_0$$

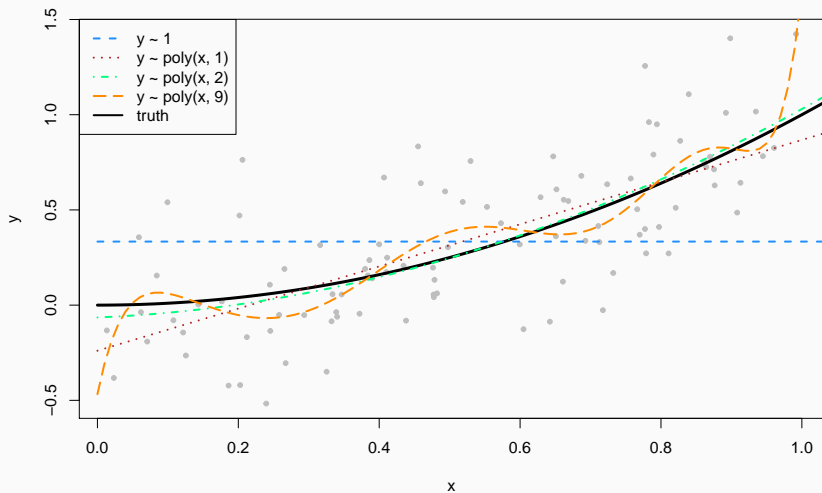
$$\text{predict}(\text{fit1}, x) = \hat{f}_1(x) = \hat{\beta}_0 + \hat{\beta}_1x$$

$$\text{predict}(\text{fit2}, x) = \hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1x + \hat{\beta}_2x^2$$

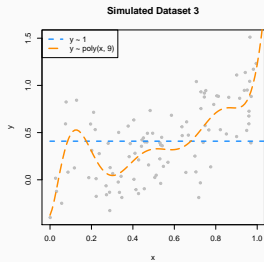
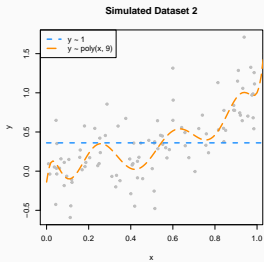
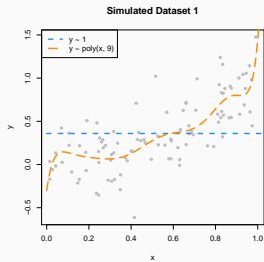
$$\text{predict}(\text{fit9}, x) = \hat{f}_9(x) = \hat{\beta}_0 + \hat{\beta}_1x + \hat{\beta}_2x^2 + \dots + \hat{\beta}_9x^9$$

Simulation Study, Trained Models

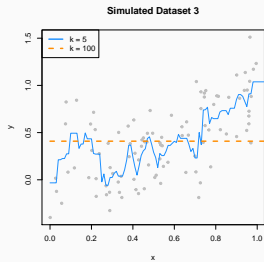
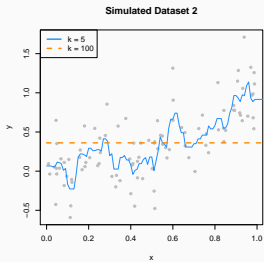
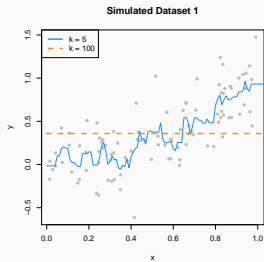
Four Polynomial Models fit to a Simulated Dataset



Simulation Study, Repeated Training



Simulation Study, KNN



Simulation Study, Setup

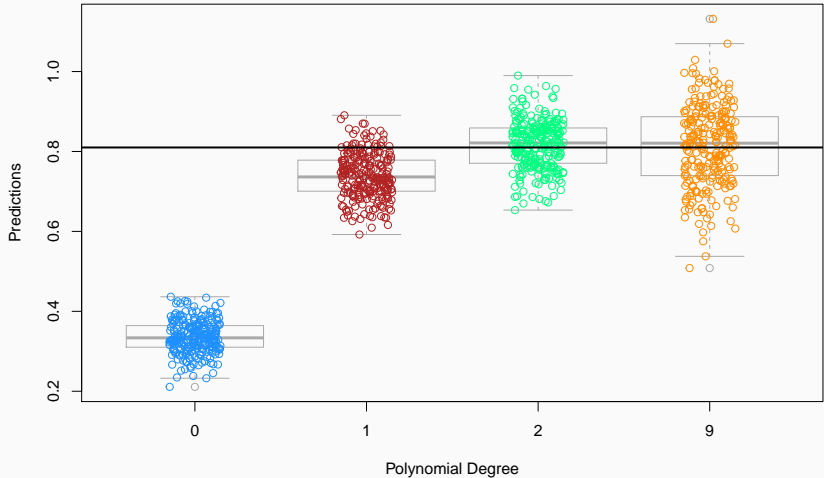
```
set.seed(1)
n_sims = 250
n_models = 4
x = data.frame(x = 0.90)
predictions = matrix(0, nrow = n_sims, ncol = n_models)
```

Simulation Study, Running Simulations

```
for(sim in 1:n_sims) {  
  
  sim_data = get_sim_data(f)  
  
  # fit models  
  fit_0 = lm(y ~ 1, data = sim_data)  
  fit_1 = lm(y ~ poly(x, degree = 1), data = sim_data)  
  fit_2 = lm(y ~ poly(x, degree = 2), data = sim_data)  
  fit_9 = lm(y ~ poly(x, degree = 9), data = sim_data)  
  
  # get predictions  
  predictions[sim, 1] = predict(fit_0, x)  
  predictions[sim, 2] = predict(fit_1, x)  
  predictions[sim, 3] = predict(fit_2, x)  
  predictions[sim, 4] = predict(fit_9, x)  
}
```

Simulation Study, Results

Simulated Predictions for Polynomial Models



- As complexity *increases*, **bias decreases**.
- As complexity *increases*, **variance increases**.

$$\begin{aligned} \text{MSE} \left(f(0.90), \hat{f}_k(0.90) \right) &= \underbrace{\left(\mathbb{E} \left[\hat{f}_k(0.90) \right] - f(0.90) \right)^2}_{\text{bias}^2(\hat{f}_k(0.90))} \\ &+ \underbrace{\mathbb{E} \left[\left(\hat{f}_k(0.90) - \mathbb{E} \left[\hat{f}_k(0.90) \right] \right)^2 \right]}_{\text{var}(\hat{f}_k(0.90))} \end{aligned}$$

Estimation Using Simulation

$$\widehat{\text{MSE}} \left(f(0.90), \hat{f}_k(0.90) \right) = \frac{1}{n_{\text{sims}}} \sum_{i=1}^{n_{\text{sims}}} \left(f(0.90) - \hat{f}_k(0.90) \right)^2$$

$$\widehat{\text{bias}} \left(\hat{f}(0.90) \right) = \frac{1}{n_{\text{sims}}} \sum_{i=1}^{n_{\text{sims}}} \left(\hat{f}_k(x0.90) \right) - f(0.90)$$

$$\widehat{\text{var}} \left(\hat{f}(0.90) \right) = \frac{1}{n_{\text{sims}}} \sum_{i=1}^{n_{\text{sims}}} \left(\hat{f}_k(0.90) - \frac{1}{n_{\text{sims}}} \sum_{i=1}^{n_{\text{sims}}} \hat{f}_k(0.90) \right)^2$$

Simulation Study, Results

Degree	Mean Squared Error	Bias Squared	Variance
0	0.22643	0.22476	0.00167
1	0.00829	0.00508	0.00322
2	0.00387	0.00005	0.00381
9	0.01019	0.00002	0.01017

- Note that, $\hat{f}_g(x)$ is unbiased
- Some live coding