Bias Variance Tradeoff

DS 301

Iowa State University

Today's Agenda

- Bias-variance tradeoff
- Introduction to Multiple Linear Regression

Recap

Regression models

General setup:

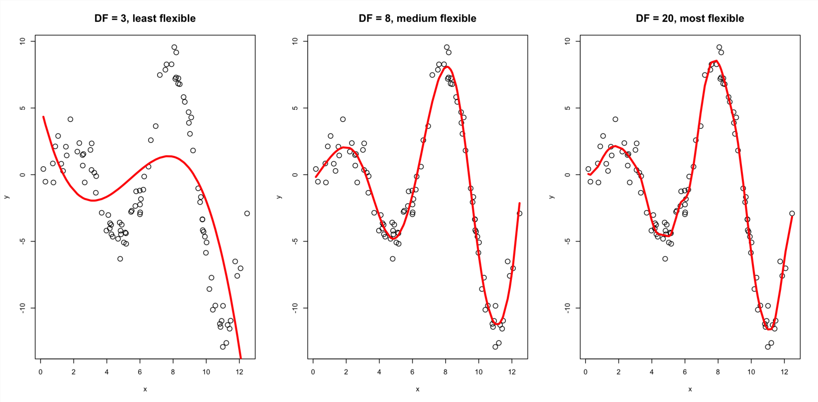
$$Y = f(X) + \epsilon$$

- Y: quantitative response.
- $X_1, X_2, \ldots X_p$: p different predictors
- We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$:
- Our goal: to estimate (learn) the function f, using a set of training data:

$$\hat{Y} = \hat{f}(X)$$

where $\hat{f}$ represents our estimate for f and $\hat{Y}$ represents the resulting prediction for Y.

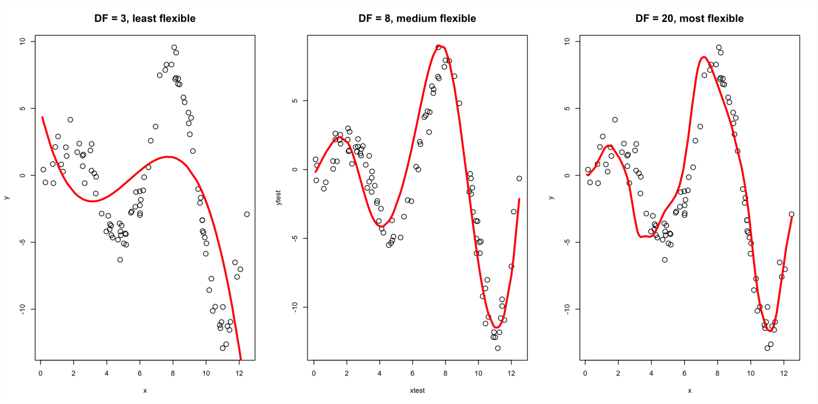
Training error



- least flexible model's training error: 16.92441
- medium flexible model's training error: 0.8542847
- most flexible model's ~~prediction~~ error: 0.6513902

training

Test error



- least flexible model's test error: 16.71268
- medium flexible model's test error: 3.579566
- most flexible model's test error: 5.47645

Mean squared error

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

Problem: $f(x)$ is unknown.
Goal: Estimate $f(x)$ from the data: $\hat{f}(x)$.

We need some way to measure how well a regression model actually matches the observed data.

In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by

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

Training MSE

Training data set is the data you used to build your model. The MSE evaluated on this data set is referred to as the **training MSE.**

$$\text{training MSE} = \frac{1}{n}\sum_{i=1}^{n}(\underline{y_i} - \underline{\hat{f}(x_i)})^2.$$

$(x_i, y_i) \rightarrow \hat{f}(x_i)$

Test MSE

Test data set is some previously unseen data that were not used to train the model. The MSE evaluated on the test set is referred to as the **test MSE**.

$$\text{test MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{f}(x_i'))^2.$$

(x_i', y_i') , $i = 1, \ldots, m$.

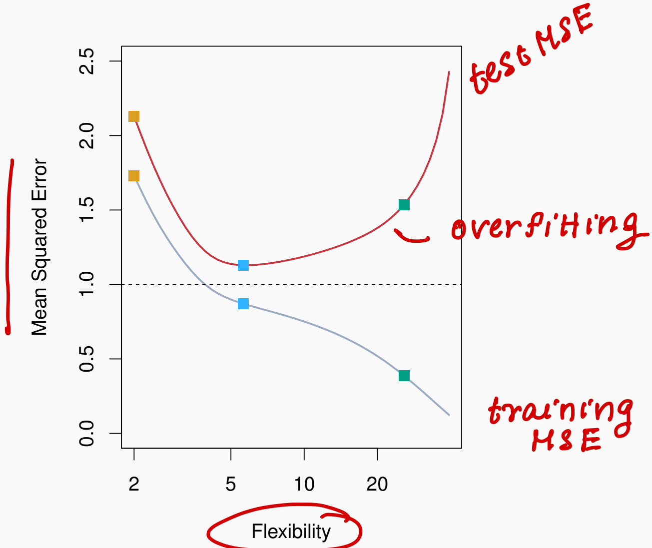
$\hat{f}(x_i)$ is our trained model

$\Downarrow$

evaluate it on our test set.

Tradeoff

Our goal in prediction is to select a method that minimizes the test MSE. Low training MSE does not imply low test MSE.



Bias-variance tradeoff

The U-shape observed in the test MSE curves turns out to be the result of two competing properties of statistical learning methods: **bias** and **variance**.

let x_0 be a fixed test point.

Then $y_0 = \underline{f}(x_0) + \varepsilon$.

$\hat{f}$ is estimated from our training data $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

The expected test MSE can be decomposed as:

$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}(\hat{f}(x_0)) + \mathrm{Var}(\varepsilon).$$

Bias-variance tradeoff

Test MSE is an estimate for the expected test MSE.

$$\text{test MSE} : \frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{f}(x_i'))^2$$

$$\text{expected test MSE} : E(y_i' - \hat{f}(x_i'))^2$$

What is the difference between $E(X)$ and $\bar{X}$?

Toy example:

training data

x_i	y_i
2	3
11	10
6.5	8
2	4

$\hat{f}(X) = 2 + 0.5(X)$.

test data

	x_i'	y_i'	$\hat{f}(x_i')$
obs. 1	3	5	$2+0.5(3) = 1.5$
obs. 2	10	11	$2+10(5) = 7$

test MSE

$$\frac{1}{2}\left((5-3.5)^2 + (11-7)^2 \right) = \boxed{}$$

expected test MSE?

$$E(y_i' - \hat{f}(x_i'))^2:$$

suppose I had another training set and estimated the model $\hat{f}(X) = 2 + 0.75X$.

Then I could evaluate this on my test set.

If I could repeat this many many many times then I would the true difference btwn y_i and its prediction.

compute this over all possible values of x_i

$\Longrightarrow$ expected test MSE.

This expected test MSE is a theoretical quantity (we cannot calculate it from real data).

we study it because it gives us insight on how statistical learning methods behave.

Bias-variance tradeoff : expected test MSE = $\text{var}(\hat{f}(x_0)) + \text{bias}(\hat{f}(x_0)) + \text{var}(\varepsilon)$

Var($\hat{f}(x_0)$): the amount by which $\boxed{\hat{f}(x_0)}$ would change if we estimated it using a different training set.

↳ if small changes in training set lead to large changes in $\hat{f}$

↳ suggests $\hat{f}$ is modeling noise and not underlying signal.

↳ high variance (overfitting)

More flexible methods tend to have high variance.

Bias-variance tradeoff

Bias($\hat{f}(x_0)$): refers to the error introduced by estimating f.

$E(\hat{f}(x_0)) - f(x_0)$.

Measures deviation of average prediction from the truth.

- unbiased estimate: $E(\hat{f}(x_0)) = f(x_0)$
- More flexible methods tend to have low bias.

↳ $\mathrm{var}(\varepsilon)$: irreducible error

$\Rightarrow$ expected test MSE $= \mathrm{var}(\hat{f}(x_0)) + \mathrm{bias}(\hat{f}(x_0))^2$ + irreducible error.

Bias-variance tradeoff