The Idea of Fitting a Predictive Model

Linear Regression and kNN regression

Ramesh Srinivasan

October 3, 2024

Multivariate Data Structure

A data matrix of the form below is at the heart of any data science project.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & a_{1M} \\ \vdots & \ddots & & \\ x_{N1} & \dots & & a_{NM} \end{bmatrix}$$

- In this data matrix there are M columns corresponding to the M different variables being measured.
- In this data matrix there are N rows corresponding to N observations
- If there are multiple data collection, there may be K such matrices.
- The goal is to take the data in this matrix and:
  1. *Classification/Regression* Use the data to predict another variable which is like a Response. In psychology, this is usually behavior. In Neuroscience, the data above is from the brain.
  2. *Clustering* Use the data to learn about supgroups of either N observations or M observables.
  3. *Latent Variable Models* Learn about hidden variables that are generating the observed M variables because they provide better theiretical intuition.

Book Notation

- We will use Y to denote a response or target that we wish to predict. X will be the prediction variables.
- The goal is to obtain a model $Y = f(X) + \epsilon$ where ϵ captures measurement errors and failures of the model and X to capture variability in Y (due to unmeasured factors).
- With a good model we will be able to predict Y for new observations X.
- Depending on the complexity of f, we may be able to understand how each component Xj of X affects Y.
- For a scientist, learning about f is important because it allows you to formulate a hypothesis for confirmatory experiments.

Assesing Model Accuracy: Training and Test data

- Suppose we fit a model $\hat{f}(x)$ to some **training** data $Tr = \{x_i, y_i\}$ of size N.
- We could compute the average squared **fitting** error over Tr:

$$MSE_{Tr} = Ave_{i \in Tr}[y_i - \hat{f}(x_i)]^2$$

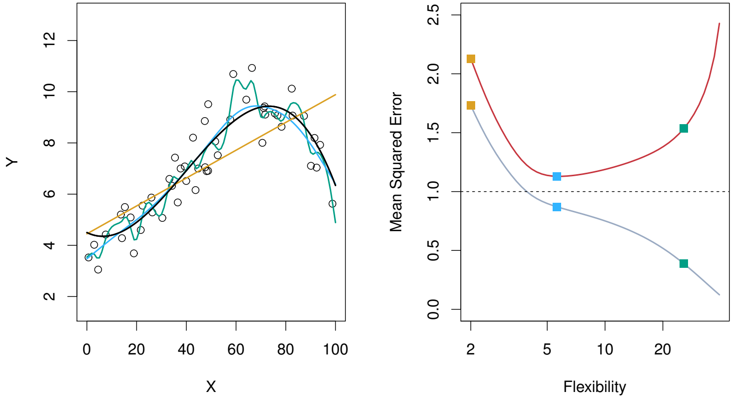
It's often the case that we minimized this square error in order to choose $\hat{f}$.

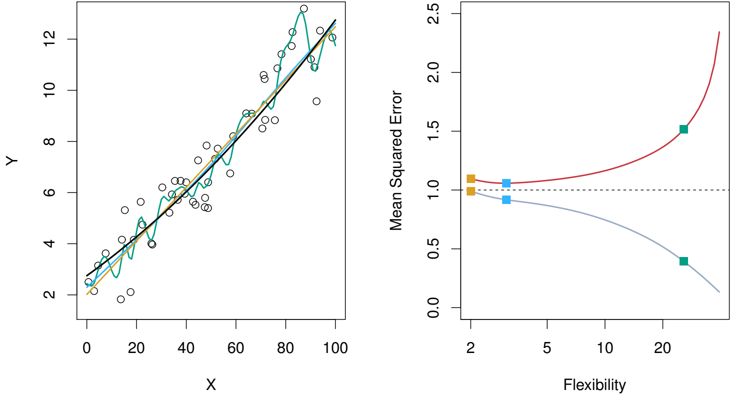
- It would be better if we had more data, which we label **test** data $Te = \{x_i, y_i\}$ of size K. Then we could measure a average squared **prediction** error.

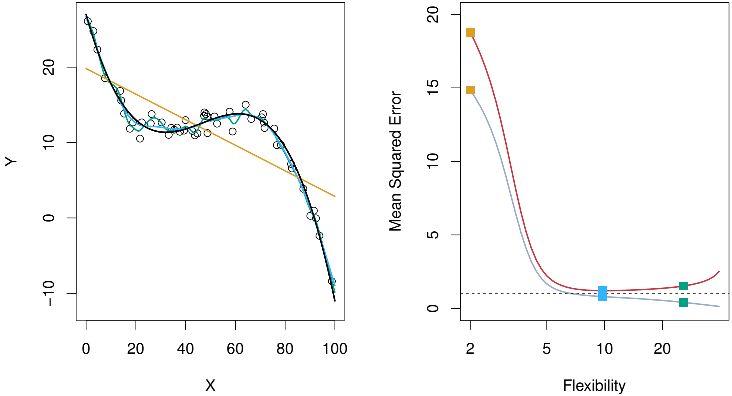
$$MSE_{Te} = Ave_{i \in Te}[y_i - \hat{f}(x_i)]^2$$

Critically, when we perform this test, we do not update $\hat{f}$ based on the test data.

- The design of the training and test data sets are a critical component of robust statistical learning. This includes the choice of N and K, and the choice of how we design sampling the training and test data from our data matrix.


Y
X
0
20
40
60
80
100
2
4
6
8
10
12
Mean Squared Error
Flexibility
2
5
10
20
0.0
0.5
1.0
1.5
2.0
2.5


Y
X
2
4
6
8
10
12
0
20
40
60
80
100
Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20

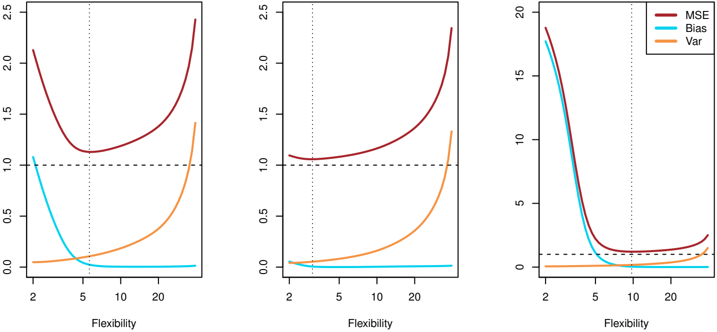

Y
X
0
20
40
60
80
100
−10
0
10
20
Mean Squared Error
Flexibility
2
5
10
20
0
5
10
15
20

Bias-Variance Trade-off

- Suppose we have fit a model $\hat{f}(x)$ to some training data Tr, and let (x_0, y_0) be a new test observation drawn from the population.
- If the true model is $Y = f(X) + \epsilon$ with $f(x) = E(Y|X = x))$

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon)$$

- Note, $Bias(\hat{f}(x_0)) = E[\hat{f}(x0)]f(x0)$ -> Error in the model.
- $Var(\hat{f}(x_0))$ is the variability due to the variability in Tr.
- Typically as the flexibility of $\hat{f}$ increases, the variance term increases, and the bias term decreases.


MSE
Bias
Var
2.5
2.0
1.5
1.0
0.5
0.0
2
5
10
20
Flexibility
20
15
10
5
0

Linear Regression

- We assume a model $Y = \beta_0 + \beta_1 X + \epsilon$ wheren β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term
- Let $\hat{y}_i = \hat{b_0} + \hat{b_1} x_i$ be the prediction for sample x_i.
- Then $e_i = y_i - \hat{y}_i$ represents the error (or residual)
- The residual sum of squares(RSS) is $RSS = e_1^2 + e_2^2 + \; + e_n^2$

R squared

- We have the resdidual sum of squares, a measure of goodness of fit is

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- In the case of a simple linear regression with 1 predictor variable x, R^2 is simply the squared correlation coefficient.