Model Accuracy

Some trade-offs

- Prediction accuracy versus interpretability.

  - Linear models are easy to interpret; thin-plate splines are not.

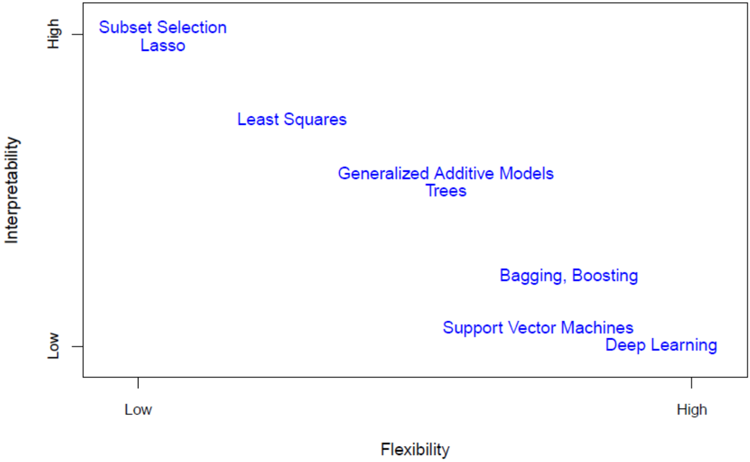
- Good fit versus over-fit or under-fit.

  - How do we know when the fit is just right?

- Parsimony versus black-box.

  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Some trade-offs



Model Accuracy

- Regression model : f, $Y \sim f(X)$

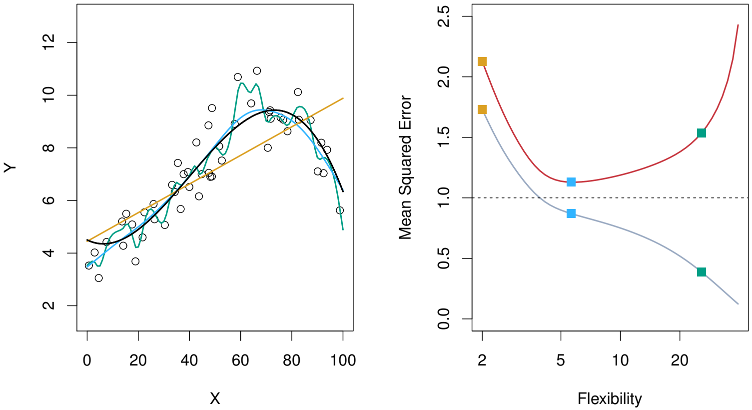
  - Training data : 모형 적합을 위하여 사용한 데이터,

    $\mathsf{Tr} = \{x_i, y_i\}_{i=1}^{n}$

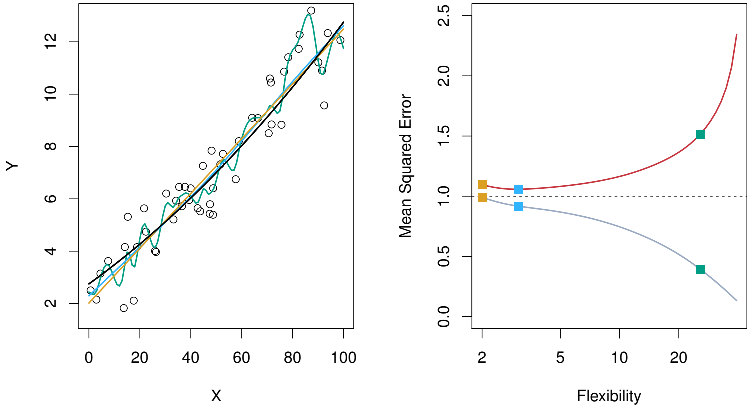
    $$\mathsf{MSE}_{\mathsf{Tr}} = \mathsf{Ave}_{i \in \mathsf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2$$

  - Test data : 아직 관측되지 않은 데이터, $\mathsf{Te} = \{x_i, y_i\}_{i=1}^{M}$

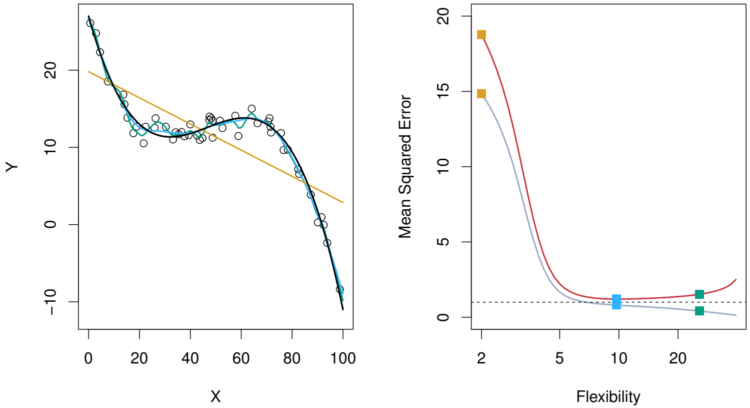
    $$\mathsf{MSE}_{\mathsf{Te}} = \mathsf{Ave}_{i \in \mathsf{Te}} \left[y_i - \hat{f}(x_i) \right]^2$$



- Black line : 실제 분포, Red curve : MSE_{Te}, Grey curve : MSE_{Tr}



- Black line : 실제 분포 - 직선에 가까움 (smoother)



- Black line : 실제 분포 - 비선형 (more flexible)

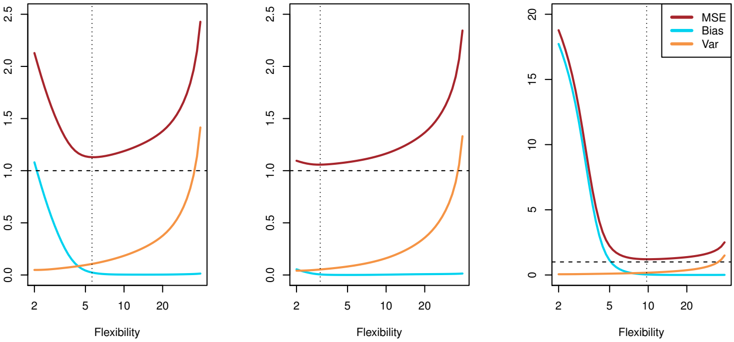
Bias-Variance Trade-off

- True model : $Y = f(X) + \epsilon$,
- $\hat{f}(x)$: training data를 이용하여 적합한 모델
- test data : (x_0, y_0)

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

- Typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a bias-variance trade-off.

Bias-Variance Trade-off



Classification

- Classification model : C, Y $C(X)$

- 오분류율 (misclassification error rate)

$$\mathsf{Err}_{\mathsf{Te}} = \mathsf{Ave}_{i \in \mathsf{Te}} I[y_i \neq \hat{C}(x_i)]$$

- Bayes classifier :

$$p_k(x_0) = P(Y = k | X = x_0)$$

의 값을 가장 크게 해주는 범주 k로 분류

K-Nearest Neighbors : KNN

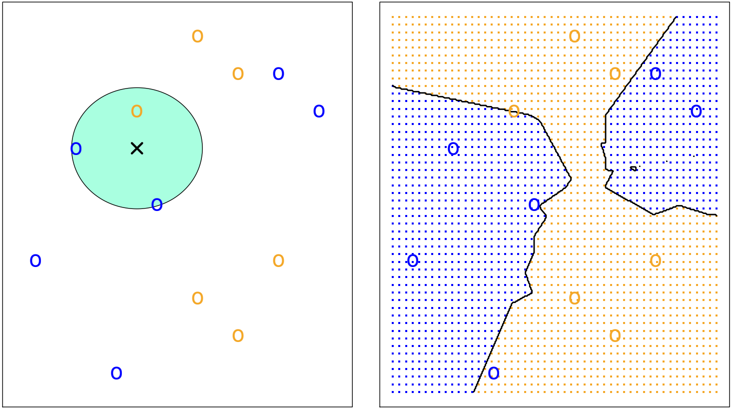
- 가장 가까운 K개의 데이터를 탐색하여 분류 또는 예측
- $\mathcal{N}_0$: x_0로부터 가장 가까운 K개의 관측값
- 분류 :

$$P(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(Y_i = j)$$

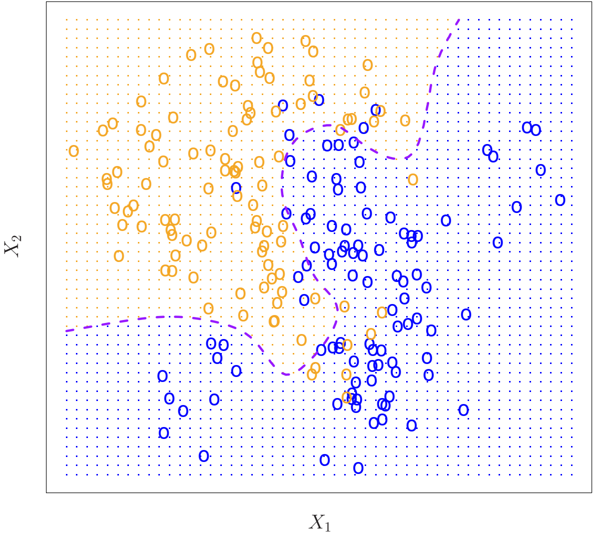
- 예측 :

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

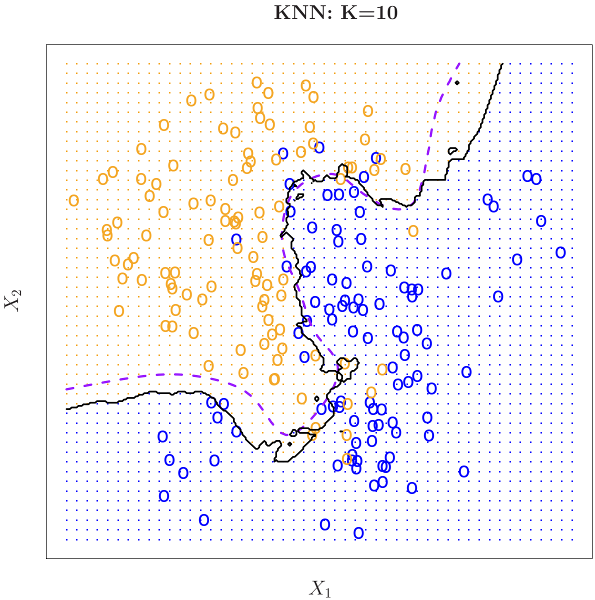
KNN (K=3)



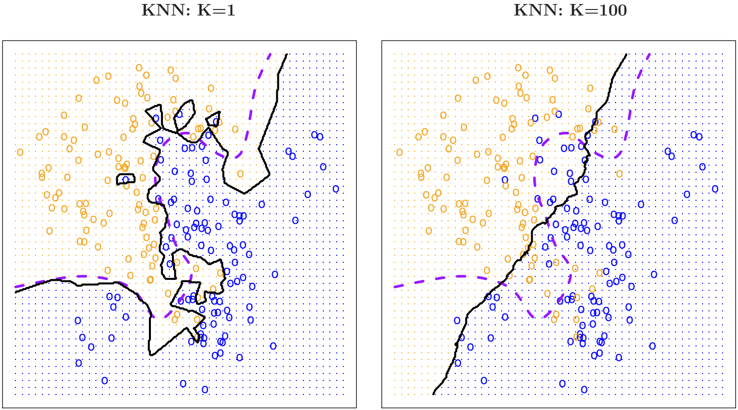
KNN : Example



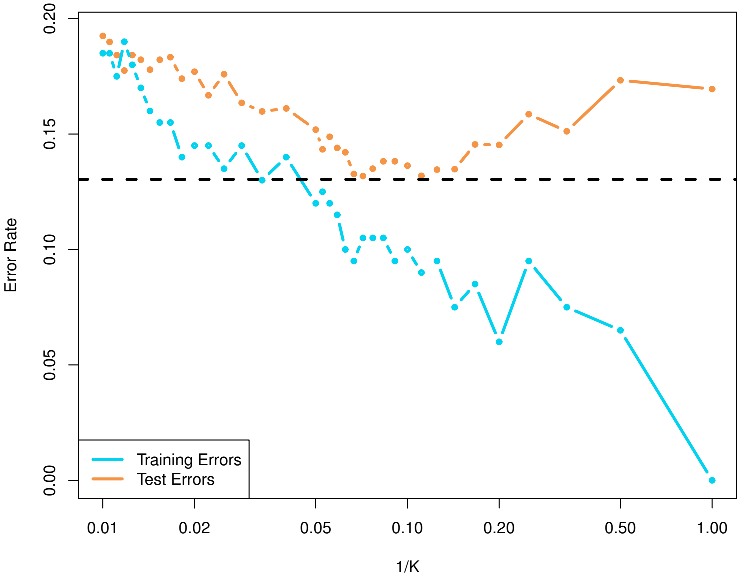
KNN : Example (K=10)



KNN : Example



KNN : Example


Error Rate
0.00
0.05
0.10
0.15
0.20
1/K
0.01
0.02
0.05
0.10
0.20
0.50
1.00
Training Errors
Test Errors