Assessing Model Accuracy

(ISLR 2.2)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

1. Measuring the Quality of Fit
2. Bias-Variance Trade-Off
3. The Classification Setting

Measuring quality of fit

Suppose we have a regression problem.

- One common measure of accuracy is the *training MSE*

$$MSE_{\mathsf{Tr}} = \frac{1}{n_{\mathsf{Tr}}} \sum_{i \in \mathsf{Tr}} [y_i - \hat{f}(x_i)]^2$$

- But usually we don't care about model fitting for training data, we want to predict new data — "test data".
- *Test MSE*

$$MSE_{\mathsf{Te}} = \frac{1}{n_{\mathsf{Te}}} \sum_{i \in \mathsf{Te}} [y_i - \hat{f}(x_i)]^2$$

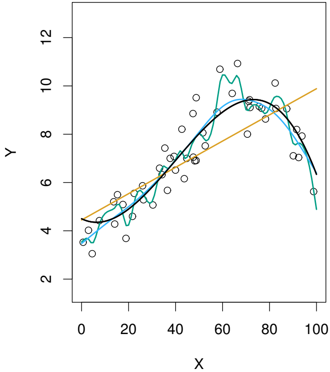
Note: $\hat{f}$ is estimated by the training data.

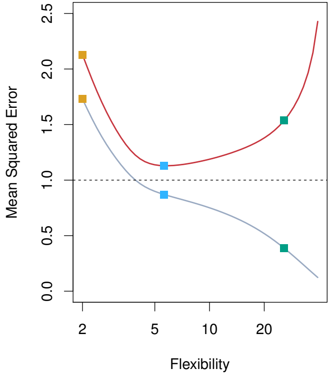
Training vs. test MSE's

There is no guarantee that the method with the smallest training MSE will have the smallest test (i.e. new data) MSE.

- $\hat{f}$ is usually chosen such that MSE_{Tr} is small.
- In general the more flexible a method is the lower its training MSE will be. It will fit or explain the training data very well.
- However, the test MSE may in fact be higher for a more flexible method than for a simple approach like linear regression.

Examples with different levels of flexibility: example 1

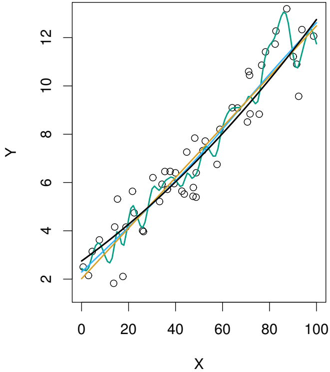


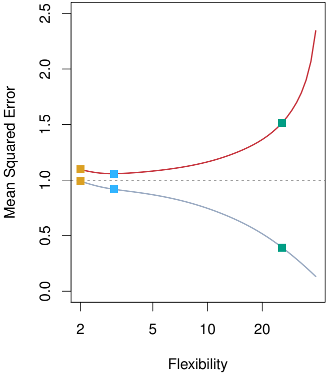


- Black: truth
- Orange: linear regression
- Blue: smoothing spline ✓
- Green: smoothing spline (more flexible)

- Red: test MSE
- Grey: training MSE
- Dashed: Minimum possible test MSE (irreducible error)

Example 2: smooth truth

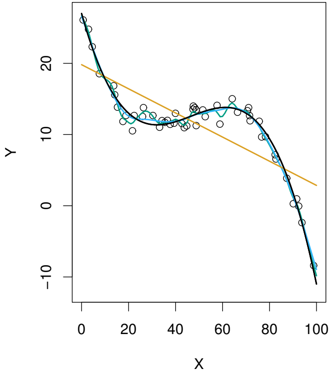


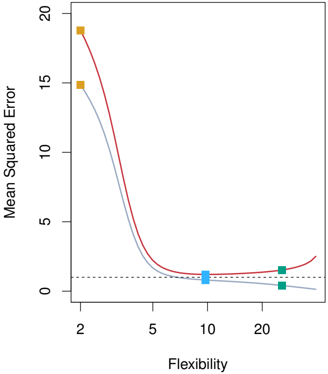


- Black: truth
- Orange: linear regression ✓
- Blue: smoothing spline ✓
- Green: smoothing spline (more flexible)

- Red: test MSE
- Grey: training MSE
- Dashed: Minimum possible test MSE (irreducible error)

Example 3: wiggly truth with low noise





- Black: truth
- Orange: linear regression
- Blue: smoothing spline ✓
- Green: smoothing spline (more flexible) ✓

- Red: test MSE
- Grey: training MSE
- Dashed: Minimum possible test MSE (irreducible error)

Bias-variance trade-off

Let (x_0, y_0) be a test data point. The expected test MSE at x_0 is

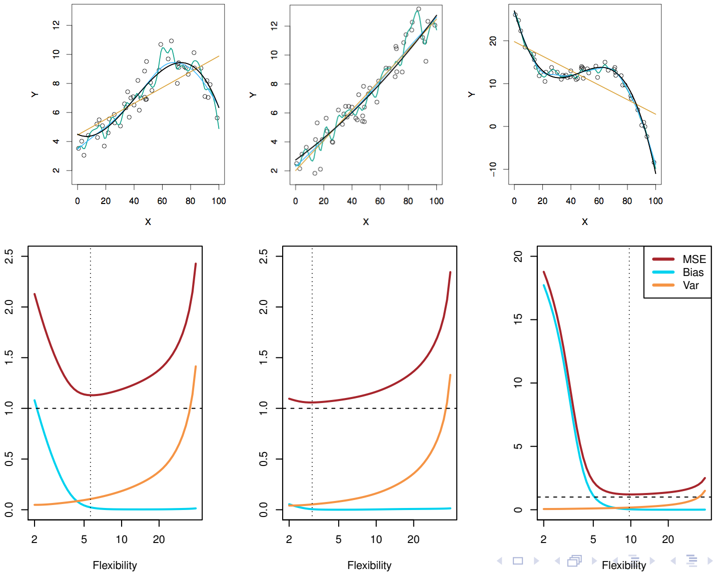
$$E[y_0 - \hat{f}(x_0)]^2 = \underbrace{V[\hat{f}(x_0)]}_{\text{Variance}} + \underbrace{[E(\hat{f}(x_0)) - f(x_0)]^2}_{[\text{Bias}(\hat{f}(x_0))]^2} + V(\epsilon)$$

- The expected test MSE $\geq V(\epsilon)$, the irreducible error.
- Variance: the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- Bias: introduced by approximating a real-life problem. In general, more flexible methods result in less bias.

As the flexibility of $\hat{f}$ increase,

$$\text{Variance} \uparrow \quad \text{Bias} \downarrow$$

Test MSE, bias and variance



Classification problems

- For a regression problem, we used the MSE to assess the accuracy of the statistical learning method
- For a classification problem, we can use the *error rate*:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

  Here $I(A)$ is an indicator function of the event A. It equals 1 if A occurs, otherwise, it equals 0.
- The error rate represents the fraction of misclassifications.
- Training error rate vs. test error rate.

Bayes classifier

Suppose the conditional distribution $P(Y \mid X)$ is known (which is usually not possible), then *the Bayes classifier* at x_0 is

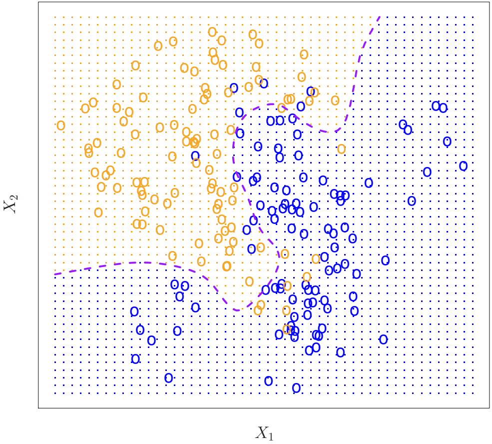
$$\hat{C}(x_0) = \arg\max_j P(Y = j \mid X = x_0)$$

Bayes error rate at x_0

$$1 - \max_j P(Y = j \mid X = x_0)$$

- The Bayes error rate is the lowest possible error rate that could be achieved if somehow we knew exactly what the “true” probability distribution of the data looked like.

Bayes decision boundary

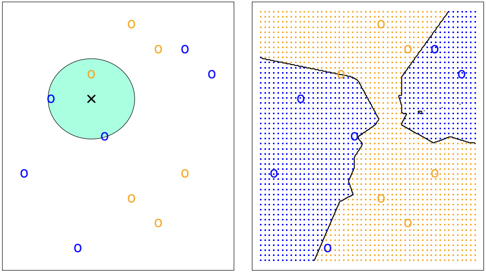


If $Y \in \{0, 1\}$, then $\hat{C}(X) = j$ if $P(Y = j \mid X = x_0) > 0.5$, for $j = 0, 1$.

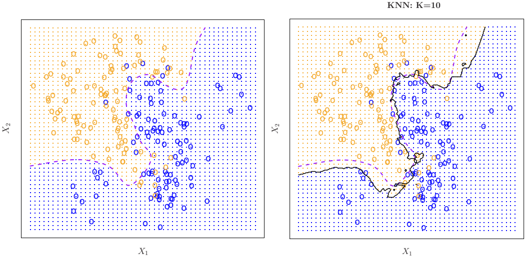
K-nearest neighbors (KNN)

For any given $X = x_0$, we find the K closest neighbors to x_0 in the training data, denoted by $\mathcal{N}_0$, and examine their corresponding Y:

$$\hat{C}(x_0) = \arg\max_j \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$



- KNN is simple, but it can achieves results close to the Bayes classifier.



- Choice of K matters! Smaller K: more flexible