2.2 Assessing Model Accuracy

Prof. Lauren Perry

Model Accuracy

- No one method is going to be appropriate for every dataset.
- In practice, choosing the best method is one of the most challenging parts of doing statistical work!

Quality of Fit

What do we mean by "quality of fit"?

- Goal: quantify the extent to which the model is able to accurately predict response values Y for a given set of inputs X.

Mean Squared Error

In the regression setting, we usually use *mean squared error* (MSE) to examine quality of fit.

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

where $\hat{f}(x_i)$ is the predicted value and y_i the actual output for the ith observation.

- If predicted values are close to the truth (on average), MSE will be (relatively) small.
- If predicted values are far from the truth, MSE will be large.

Training versus Test Data

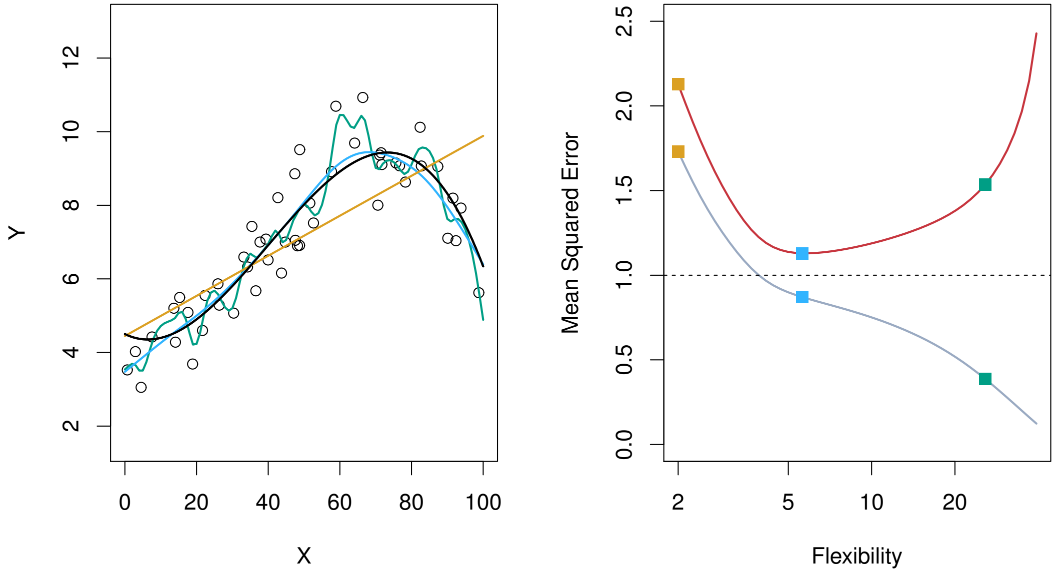
We'd prefer to examine model accuracy using data that was *not* used to construct the model.

- This helps us examine possible overfitting in the original model.

One approach: separate our data into *test* and *training* sets.

- Use the *test data* to construct the model.
- Use the *training data* to examine quality of fit.

The next slide shows data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.



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

The Bias-Variance Trade-off

The U-shape in MSE curves is a result of two competing properties.

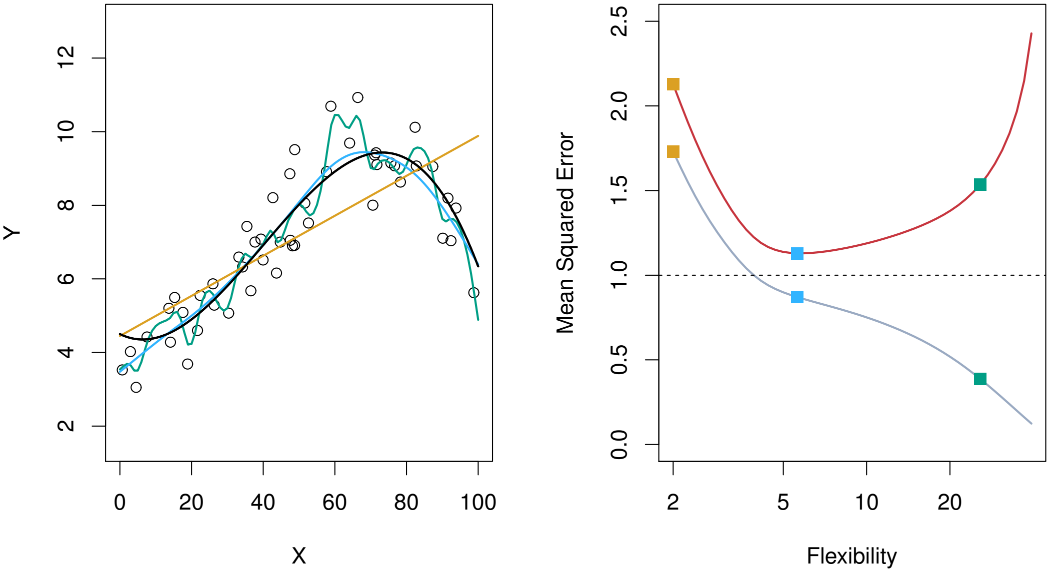
That is, the expected MSE (using test data) can be decomposed into

1. variance of $\hat{f}$
2. squared bias of $\hat{f}$
3. variance of error terms ϵ

So to minimize MSE, we would select a method with low bias and low variance.

Variance of Statistical Learning Methods

- Variance is the amount $\hat{f}$ would change if it were estimated using a different set of training data.
  - Ideally, $\hat{f}$ will not vary too much between training sets.
  - In general, more flexible methods have higher variance.



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

Bias of Statistical Learning Methods

- Bias is the error introduced by approximating a real-life problem.
  - Ex: it is unlikely that any sufficiently complex real-life scenario has a linear relationship.
  - In general, more flexible methods have less bias.

The Bias-Variance Trade-off

More flexible methods = less bias, but more variance.

- For a given problem, minimizing (test data) MSE means finding a *balance* between bias and variance.

The Classification Setting

Here, we examine *error rate*:

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

where $\mathrm{I}()$ is the *indicator function* which evaluates to 1 if its input expression is true, 0 otherwise.

- We can calculate test or training error rate, depending which data is used.

A good classifier is one for which error rate is low.

The Bayes Classifier

Test error rate is minimized when each observation is assigned to the most likely class.

That is, assign observation x_0 to the class j for which

$$P(Y = j | X = x_0)$$

is largest.

- $P(Y = j | X = x_0)$ is the *conditional probability* that $Y = j$ given $X = x_0$.

This is called a *Bayes classifier*.

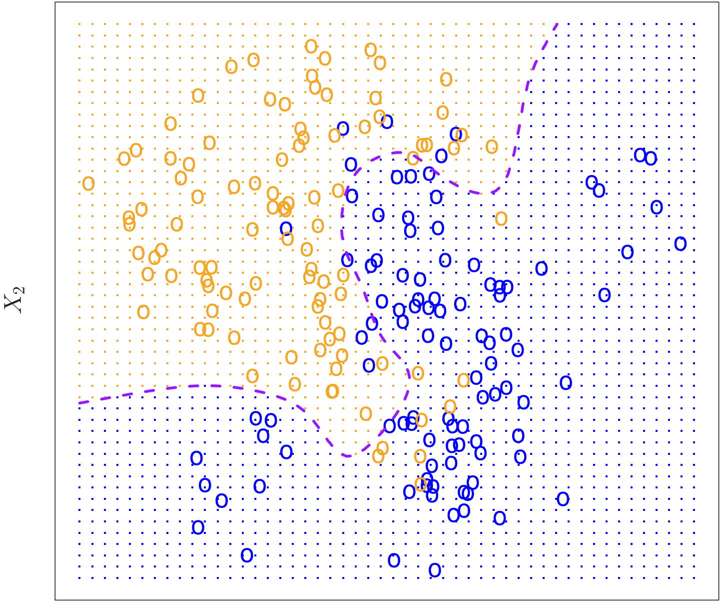
The Bayes Classifier

In a classification problem with only two classes, the Bayes classifier corresponds to

- predicting Class A if $P(Y = a|X = x_0) > 0.5$ and Class B otherwise.

The set of points for which $P(Y = a|X = x_0) = 0.5$ is called the *Bayes decision boundary*.

Example



K-Nearest Neighbors

The Bayes Classifier looks great on paper, but we don't actually know the conditional distribution of $Y|X\dots$

- Bayes Classifier is the unattainable "gold standard" which other models are compared to.

Other models attempt to estimate the distribution of $Y|X$.

- One example is K-Nearest Neighbors (KNN).

K-Nearest Neighbors

Given a positive integer K and some test observation x_0,

1. Identify the K points in the training data which are closest to x_0. Call this set of points N_0.
2. Estimate
$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_o} I(y_i = j)$$
for all classes j.
   - This is the fraction of points in N_0 whose response values $= j$.
3. Classify x_0 to the class with the largest estimated probability from (2).

K-Nearest Neighbors

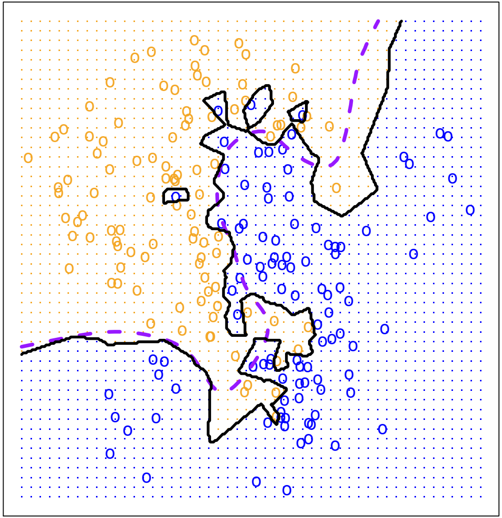
One difficulty: K is set by the user.

- Small values of K may be too flexible.
- Large values of K may not be flexible enough.

As with regression, too much flexibility will lead to overfitting of the training data.

K-Nearest Neighbors

KNN: K=1



KNN: K=100