Introduction to Statistical Learning

– (Part Three) –

Topic 3. Statistical learning - a high-level overview and illustrative examples

3.4. Assessing model accuracy

Sonja Petrović

Created for ITMS/ITMD/STAT 514

Spring 2021.

Goals of this lecture

- Assessing model accuracy
- Measuring quality of fit
  - from the point of view of classification
  - from the point of view of regression
- Intro to validation set approaches (data partitioning and cross-validation)
  - $\rightarrow$ This will be used for establishing baseline model performance

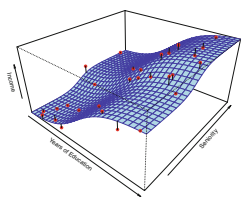
Section 1

Setting the context: estimating f, accuracy & interpretability

Review: estimating f

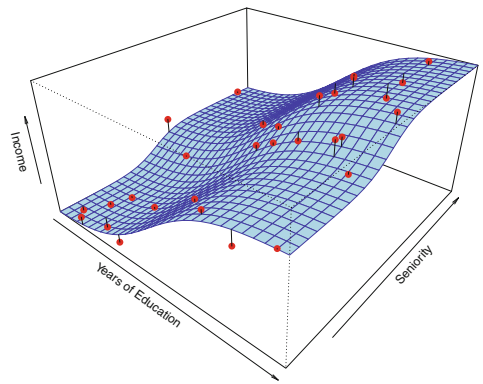
- Observe: a quantitative response Y, p different predictors, $X_1, X_2, \ldots, X_p$.
- Assume: some relationship between Y and $X = (X_1, X_2, ..., X_p)$, which can be written in the very general form

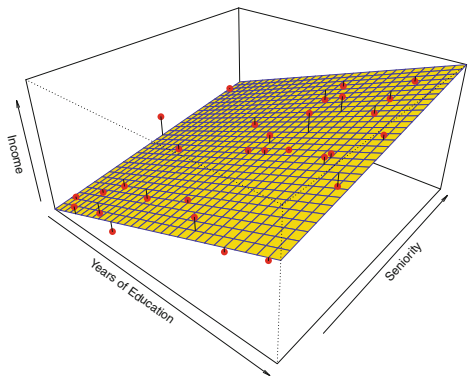
$$Y = f(X) + \epsilon.$$

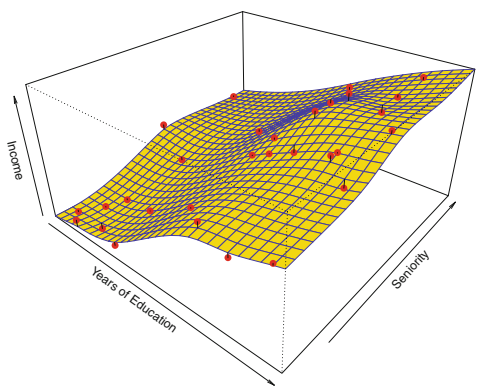


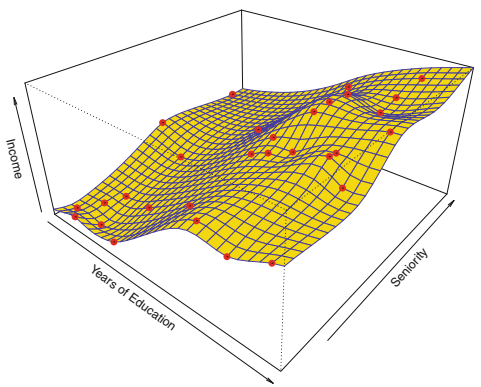
- f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$
- ϵ is a random error term, which is independent of X and has mean zero.
- In this formulation, f represents the *systematic* information that X provides about Y.

[Regression setting][1] f, $Y = f(X) + \epsilon$, $\hat{f}$, $\hat{Y} = \hat{f}(X)$









[1] ISLR book figures.

[Classification setting][2]

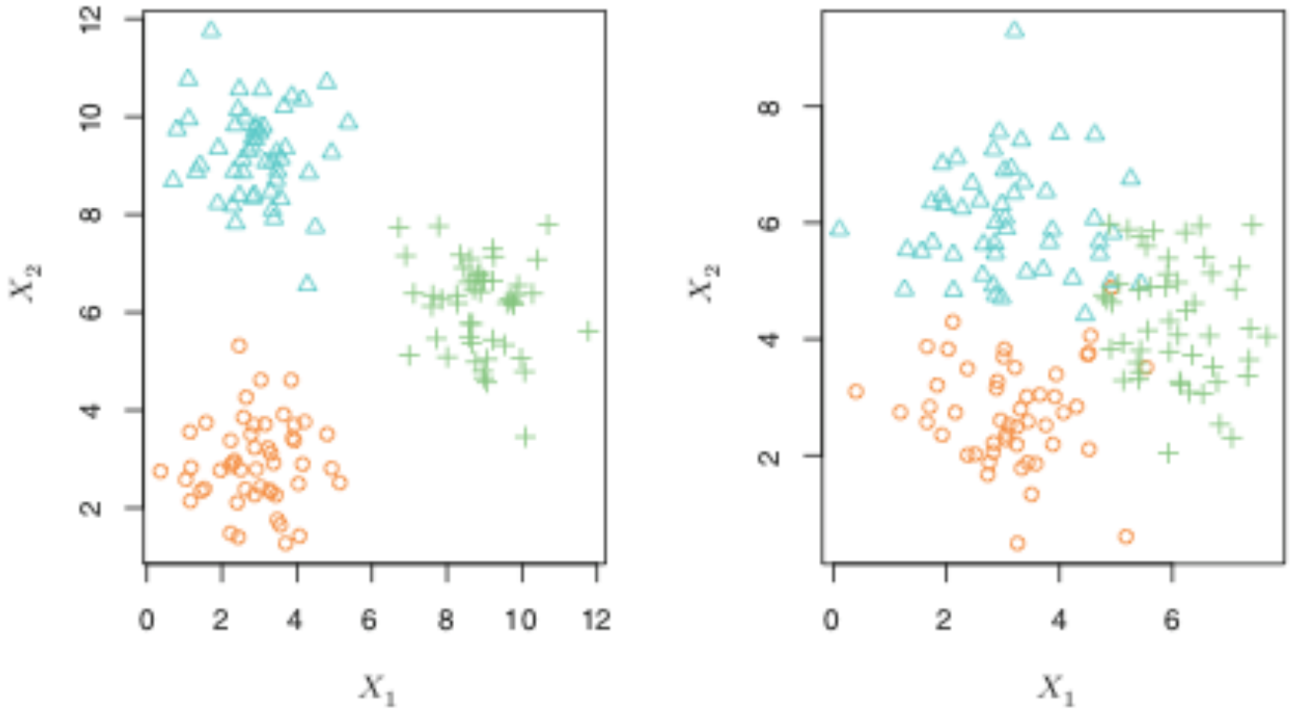


Figure 1: A clustering data set involving three groups. Each group is shown using a different colored symbol. **Left**: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. **Right**: There is some overlap among the groups. Now the clustering task is more challenging.

[2] ISLR fig2.8.

Assessing model accuracy

There is no free lunch in statistics!

No one method dominates all others over all possible data sets.

- Important task: decide, for any given set of data, which method produces the best results.
  - Selecting the best approach can be one of the *most challenging parts* of performing statistical learning in practice.
- Need: measure how well predictions match observed data.
  - $\rightarrow$ quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

Measuring quality of fit - regression setting

Mean squared error (MSE)

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

$\hat{f}(x_i)$ is the prediction that f gives for the ith observation.

- Most commonly-used measure
- Interpretation? [measuring closeness $\hat{f}(x_i) \approx y_i$?]
- *Better name: **training MSE**! [why?]

But... we don't really care about *training MSE*!

Real question:
What is the accuracy of the predictions that we obtain when we apply our method to previously unseen test data?

→ Test data!

Training vs. test data

Example 1

Goal: Develop an algorithm to predict a stock's price based on previous stock returns.

- We can train the method using stock returns from the past 6 months.
- But we don't really care how well our method predicts last week's stock price.
- We instead care about how well it will predict **tomorrow's price** or **next month's price**.

Example 2

Goal: predict diabetes risk for future patients based on their clinical measurements.

- Clinical measurements (e.g. weight, blood pressure, height, age, family history of disease) for a number of patients, + info whether each patient has diabetes.
- Train a statistical learning method to **predict risk of diabetes based on clinical measurements**.
- No interest: whether method accurately predicts diabetes risk for patients used to train the model, since we already know which of those patients have diabetes.

The test MSE

- (x_0, y_0) a *previously unseen test observation*
- **Goal**: $\hat{f}(x_0) \approx y_0$?

Test MSE

$$Ave(y_0 - \hat{f}(x_0))^2$$

average squared prediction error for test observations (x_0, y_0).

Discuss meaning!

Minimizing (test) MSE

How to select a method that does this??

Scenario: test data available

- Set of observations not used to train the statistical model.
- Evaluate test MSE, $Ave(y_0 - \hat{f}(x_0))^2$ on that set.

Minimizing (test) MSE

How to select a method that does this??

Scenario: test data available

- Set of observations not used to train the statistical model.
- Evaluate test MSE, $Ave(y_0 - \hat{f}(x_0))^2$ on that set.

Scenario: no test observations available

- Maybe.... select a model/method that minimizes training MSE, $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$.
  - $\rightarrow$ fundamental problem with this strategy!

$\rightarrow$ let's look at an example [ISLR fig2.9.]:

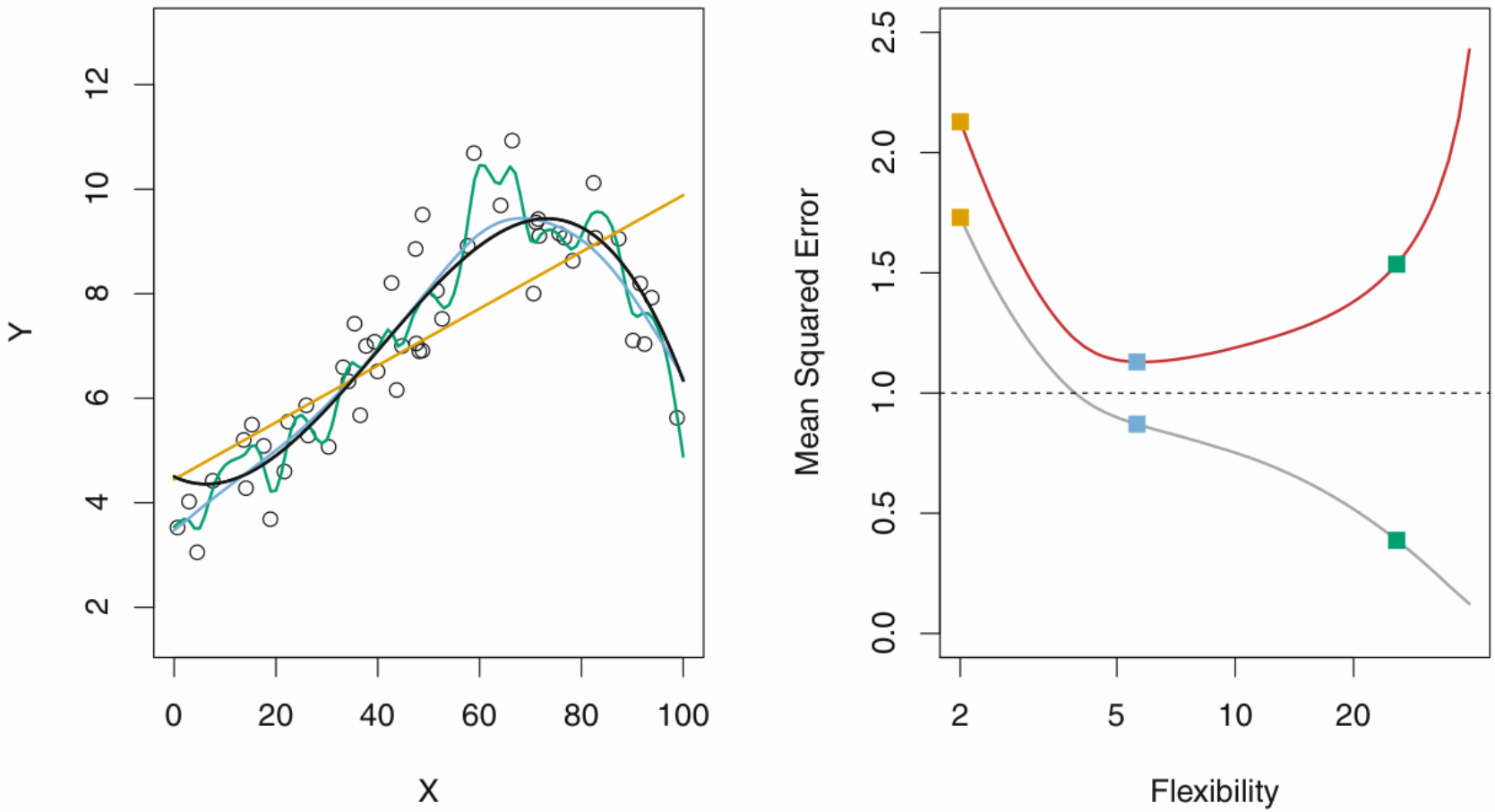


Figure 2: **Left**: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). **Right**: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

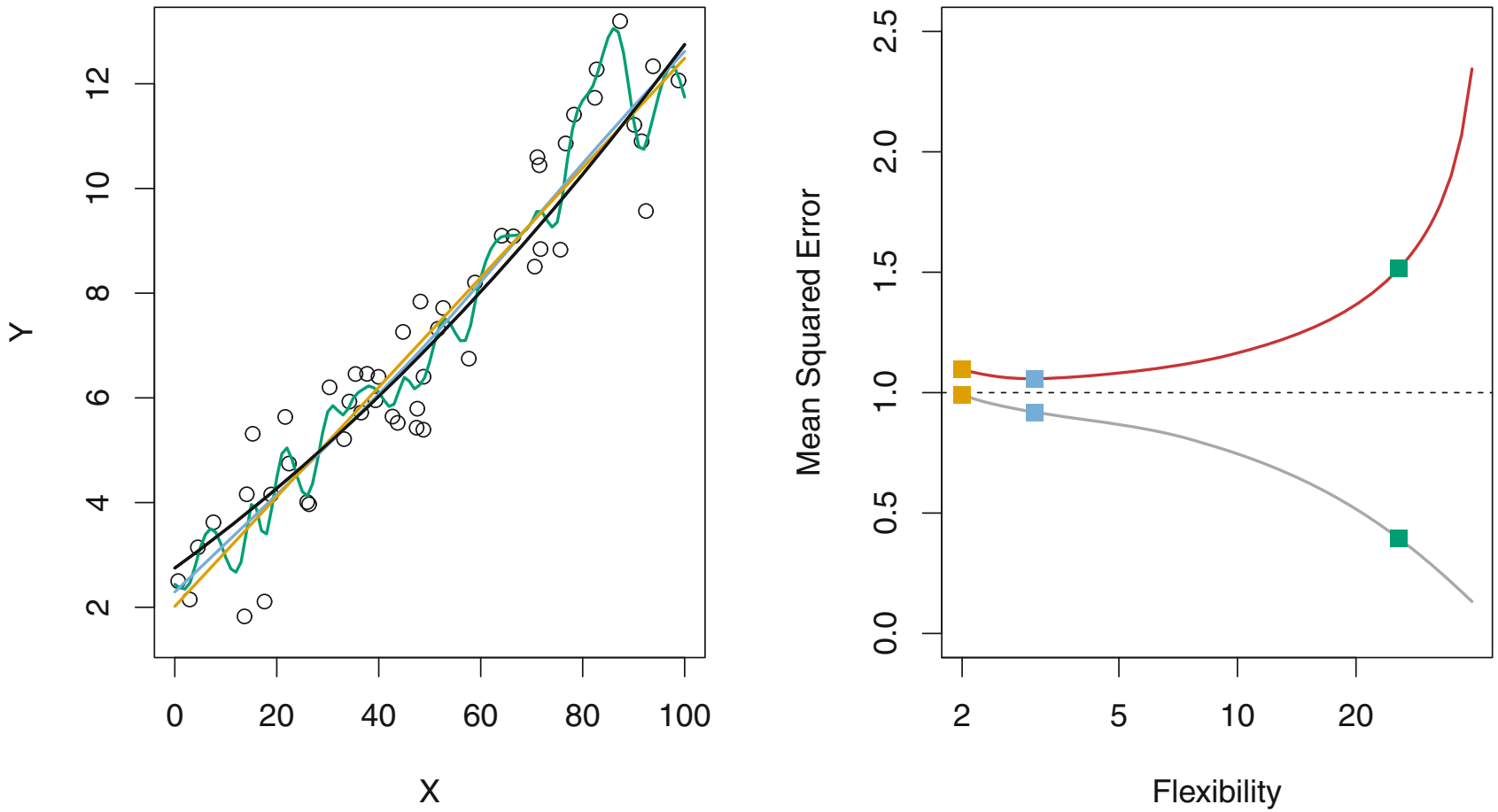


Figure 3: figure 2.10 from ISLR: Same as previous figure, but true f much closer to linear. In this case, linear regression provides a very good fit to the data.

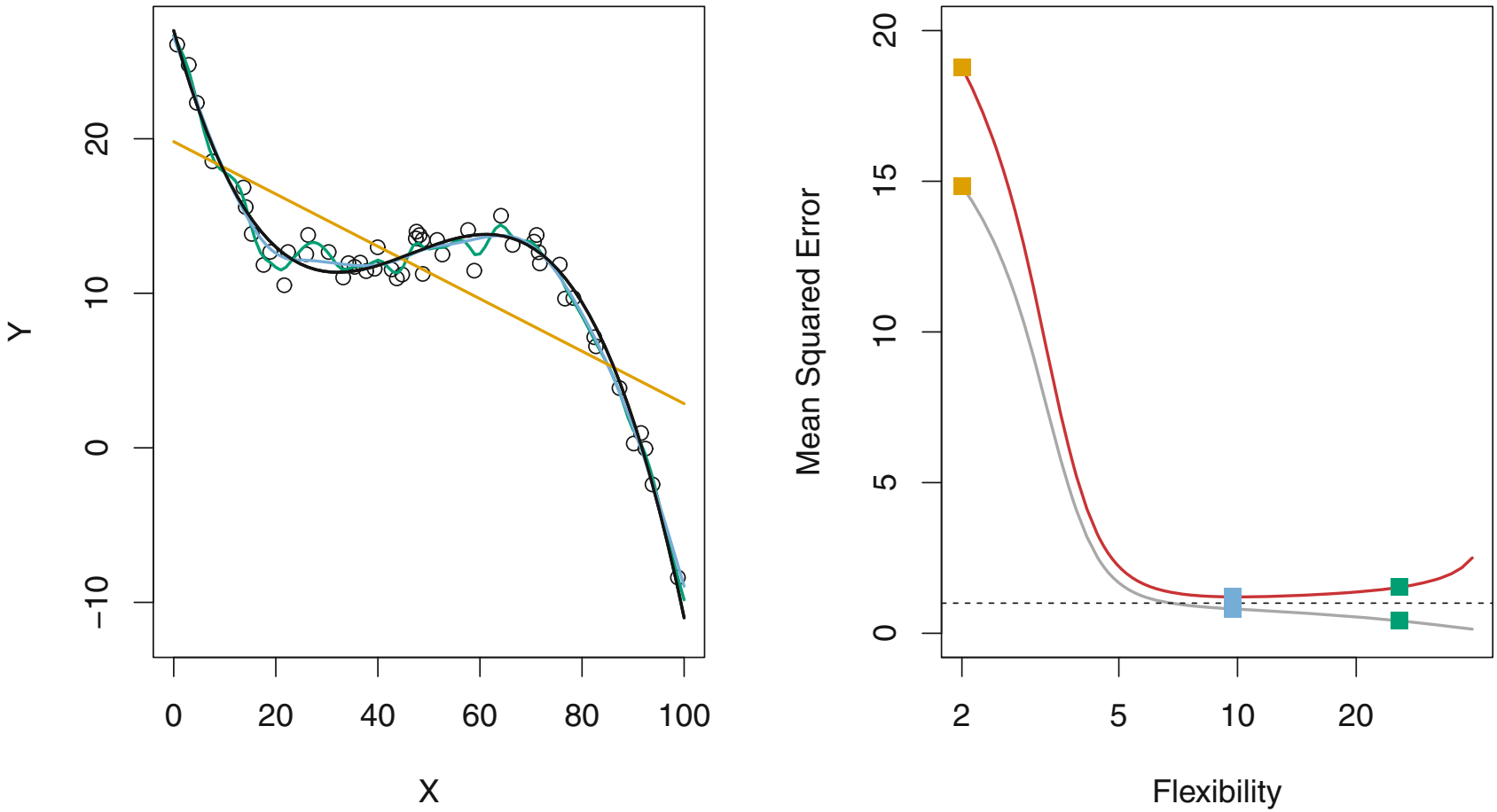


Figure 4: figure 2.11 from ISLR: Same as previous figure, but true f is farther from linear.

A fundamental conclusion

- Increase in model flexibility $\Longrightarrow$
  - decrease in training MSE
  - U-shape in test MSE.
- Small training MSE + large test MSE $\leftrightarrow$ overfitting the data!
- In practice:
  - what to do if no test data available?
  - One example (you will learn later): cross-validation = a method for estimating test MSE using training data.
  - Cf. the data partitioning worksheets!!

Idea behind cross-validation

Cross-validation is essentially one of the resampling methods.

- Remember:
  - Testing error measures average error on measurements that were not used to train the method.
  - Available test data set $\Longrightarrow$ testing error easy to compute.
- Testing error rate needs to be estimated
  - Use a very large designated tst set; or
  - Use the training data!! How?
    - Mathematical adjustment to the training error rate;
    - Cross-validate like this:

Estimate the test error rate by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

Measuring quality of fit - classification setting

- Training error rate (proportion of mistakes made)

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

- computes the fraction of incorrect classifications
- Test error rate

$$Ave(\mathbf{1}(y_0 \neq \hat{y}_0)$$

Notes to remember: * there is a (unattainable!) gold standard (a classifier with the lowest possible error rate) * next best thing: K-nearest neighbors.

illustration...

License

Content of this lecture is based on the first two chapters of the textbook Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *' An Introduction to Statistical Learning: with Applications in R'*. The book is available online.

Part of this lecture notes are extracted from Prof. Alexandra Chouldechova data mining notes CMU-95791, released under a Attribution-NonCommercial-ShareAlike 3.0 United States license.