Resampling methods

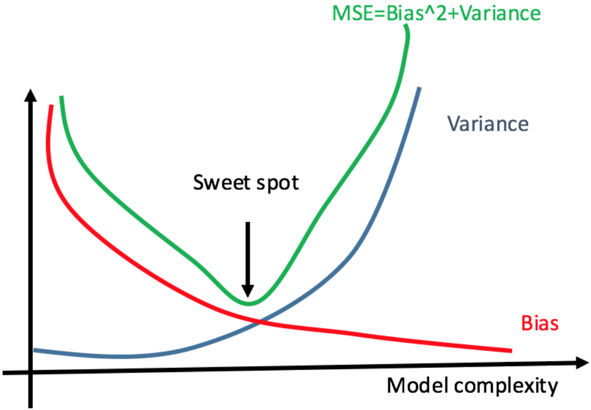
Dr. Lan Du

Faculty of Information Technology, Monash University, Australia

FIT5149 week 5

Outline

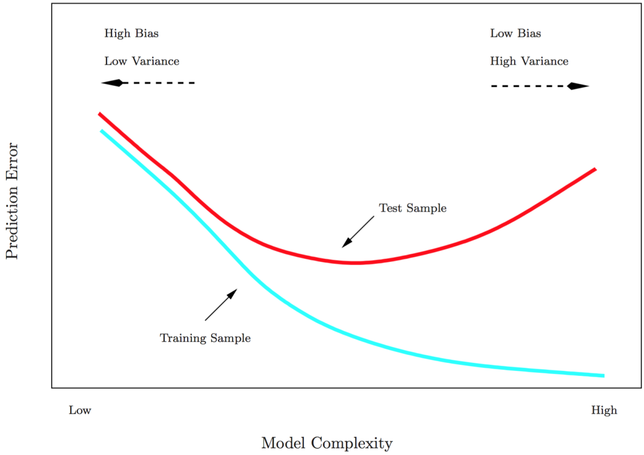
Bias-Variance



Motivation

- To draw many samples from the training set and refit the model on each sample to get better information on the model
- Extra information that is not available from fitting the model only once
- To examine how the resulting fits are different
- Two the most commonly used resampling methods:
  - Cross Validation
    - Be used to estimate the test error associated with a given statistical learning method
    - in order to evaluate its performance, or to select the appropriate level of flexibility
  - Bootstrap
    - To provide a measure of accuracy of a parameter estimate or of a given statistical learning method
- The process of evaluating the performance of a model is known as **model assessment**
- The process of selecting the proper level of flexibility for a model is known as **model selection**

Test and Training Errors

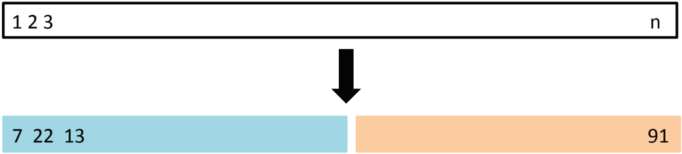


Test and Training Errors

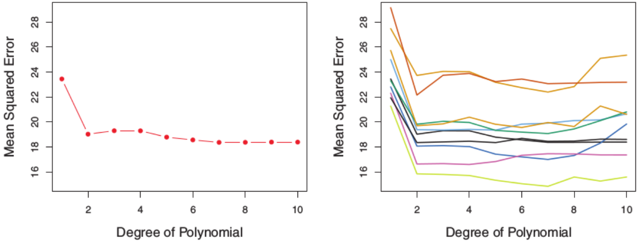
- What if there is not a large enough test set to estimate the test error rate!?
  - A number of techniques can be used to estimate this quantity using the available training data.
  - In this section, we consider a class of methods that estimate the test error rate by holding out a subset of the training observations from the fitting process,
  - And then applying the statistical learning method to those held out observations.
- Distinguish between
  - Quantitative response variable: regression models
  - Qualitative response variable: classification models

The Validation Set Approach

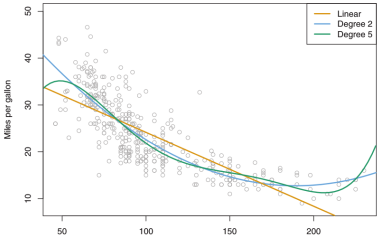
- Aim: to **estimate** the **test error** associated with fitting a particular statistical learning method on a set of observations
- The validation set approach
  - randomly dividing the available set of observations into two parts: a **training set** and a **validation set** or **hold-out set**
  - The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
  - The resulting validation set error rate (typically assessed using MSE in the case of a quantitative response) provides an estimate of the test error rate.



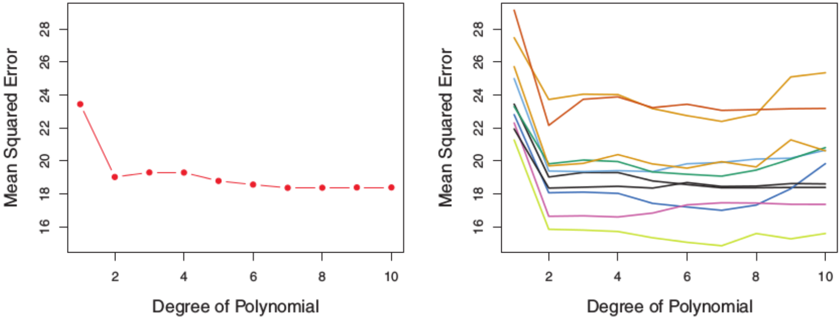
Error from Validation Set



- L: Error estimates for a single 50-50 split into training and validation
- R: validation method was repeated ten times 50-50 random splits



Error from Validation Set



- Left: Error estimates for a single split into training and validation data sets
- Right: validation method was repeated ten times, each time using a different random split
- Based on the variability among these curves, all that we can conclude with any confidence is that the linear fit is not adequate for this data.

Disadvantages

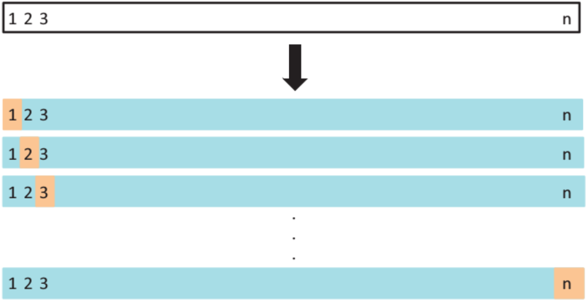
MONASH University

- The validation set approach is conceptually simple and is easy to implement.
- But it has two potential drawbacks
  1. the validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set
  2. In the validation approach, only a subset of the observations (those that are included in the training set rather than in the validation set) used to fit the model
     - Since statistical methods tend to perform worse when trained on fewer observations,
     - This suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

Leave-One-Out Cross-Validation

- Repeating this approach n times produces n squared errors, $MSE_1, \ldots, MSE_n$. The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

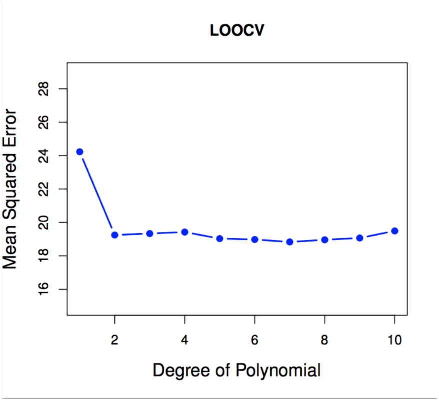


LOOCV: Advantages

- It has far less bias
  - We repeatedly fit the statistical learning method using training sets that contain $n - 1$ observations
  - Almost as many as are in the entire data set
  - In the validation set approach, in which the training set is typically around half the size of the original data set
  - The LOOCV approach tends not to overestimate the test error rate as much as the validation set approach does
- Performing LOOCV multiple times will always yield the same results
  - The validation approach will yield different results when applied repeatedly due to randomness in the training/validation set splits
  - There is no randomness in the training/validation set splits.

Test LOOCV on the Auto Data

- To obtain an estimate of the test set MSE
- From fitting a linear regression model to predict mpg using polynomial functions of horsepower
- The LOOCV error curve



LOOCV: Disadvantages

- LOOCV has the potential to be expensive to implement
- the model has to be fit n times
- If n is large and each individual model is slow to fit!!
- However,
  - LOOCV is a very general method, and can be used with any kind of predictive modeling
  - we could use it with logistic regression or linear discriminant analysis

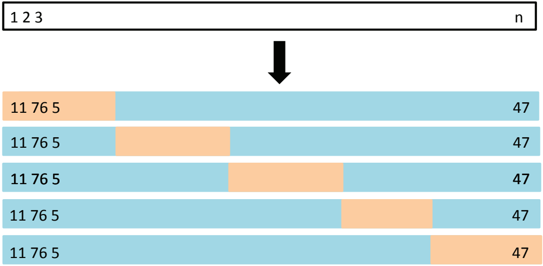
k-fold CV

- An alternative to LOOCV is k-fold CV
- Randomly dividing the set of observations into k groups, or folds, of approximately equal size
- The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds.
- The mean squared error, MSE_1, is then computed on the observations in the held-out fold
- This procedure is repeated k times
- Each time, a different group of observations is treated as a validation set
- The k-fold CV estimate is computed by averaging these values,

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i$$

5-fold CV

- A set of n observations is randomly split into five non-overlapping groups
- Each of these fifths acts as a validation set (shown in beige),
- And the remainder as a training set (shown in blue)
- The test error is estimated by averaging the five resulting MSE estimates

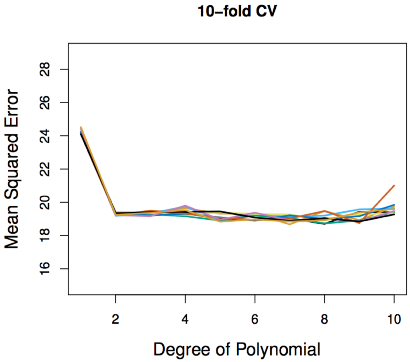


k-fold CV

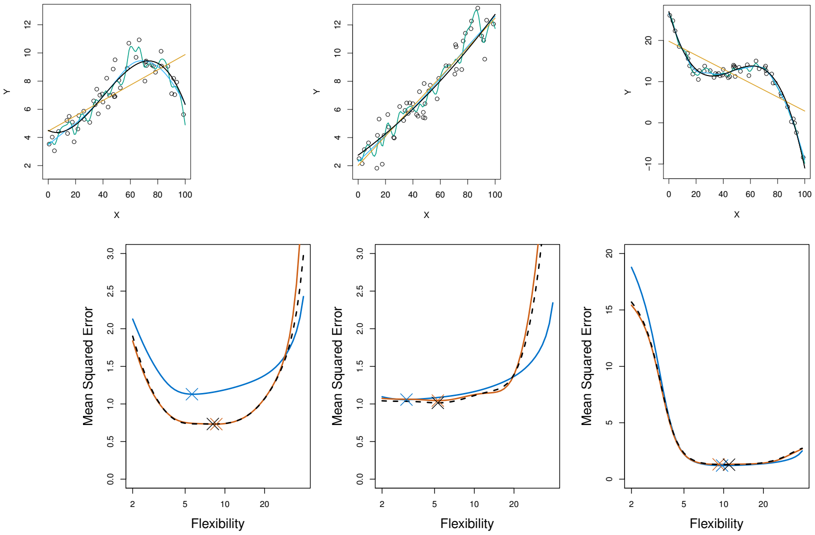
- LOOCV is a special case of k-fold CV in which k is set to equal n
- In practice, one typically performs k-fold CV using $k = 5$ or $k = 10$
- LOOCV requires fitting the statistical learning method n times
- Performing 10-fold CV requires fitting the learning procedure only ten times
- Other non-computational advantages to performing k-fold CV, involve bias-variance trade-off

Error from k-fold CV

- Nine different 10-fold CV estimates for the Auto data set
- Each resulting from a different random split of the observations into 10 folds
- There is some variability in the CV estimates as a result of the variability in how the observations are divided into ten folds
- Variability is lower than the validation set approach in estimating test error



k-fold CV: Example



Blue: true test MSE, black: LOOCV, and Orange: 10-fold CV

Model Assessment and Model Selection

- When we perform cross-validation, our goal might be to determine how well a given statistical learning procedure can be expected to perform on independent data
- the actual estimate of the test MSE is of interest
- But at other times we are interested only in the location of the minimum point in the estimated test MSE curve
- This is because we might be performing cross-validation on a number of statistical learning methods, or on a single method using different levels of flexibility,
- In order to identify the method that results in the lowest test error
- The location of the minimum point in the estimated test MSE curve is important,
- But the actual value of the estimated test MSE is not
- Despite the fact that they sometimes underestimate the true test MSE, all of the CV curves come close to identifying the correct level of flexibility

Classification

- Cross-validation can also be a very useful approach in the classification setting when Y is qualitative
- Rather than using MSE to quantify test error, we instead use the number of misclassified observations
- The LOOCV error rate

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i$$

- $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$
- The k-fold CV error rate and validation set error rates are defined analogously.

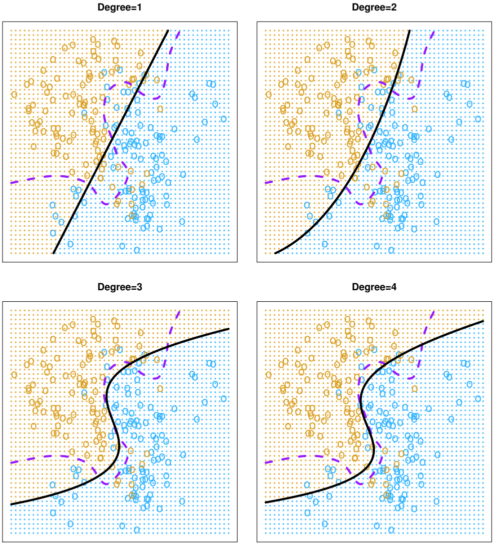
Example

- Top-left: the black solid line shows the estimated decision boundary from fitting a standard logistic regression model
- This is simulated data, we can compute the true test error rate
- Which is 0.201 and so is substantially larger than the Bayes error rate of 0.133
- Logistic regression does not have enough flexibility to model the Bayes decision boundary
- In logistic regression, we get non-linear decision boundary by using polynomial functions of the predictors
- We can fit a quadratic logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

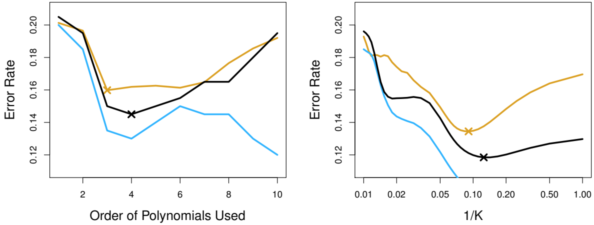
- Top-right displays the resulting decision boundary, which is now curved
- The test error rate has improved only slightly, to 0.197

Example



Example

- Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data
- Left: Logistic regression using polynomial functions of the predictors
- The order of the polynomials used is displayed on the x-axis.
- Right: The KNN classifier with different values of K, the number of neighbors used in the KNN classifier



Outline

The Bootstrap

- Used to quantify the uncertainty associated with a given estimator or statistical learning method
- Example: can be used to estimate the standard errors of the coefficients from a linear regression fit
- For linear regression, not a big deal! why?
- It can be easily applied to a wide range of statistical learning methods

A Toy Example

- We wish to determine the best investment allocation
- Invest a fixed sum of money in two financial assets that yield returns of X and Y (random quantities)
- Invest a fraction of α in X and $1-\alpha$ in Y
- There is variability associated with the returns on these two assets
- Choose α to minimize the total risk, or variance, of our investment

$$\text{minimize Var}\left(\alpha X + (1-\alpha)Y\right)$$

- It is proven that

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

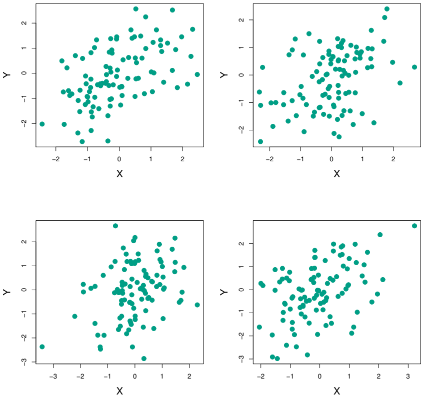
- Where $\sigma_x^2 = \text{Var}(X)$, $\sigma_x^2 = \text{Var}(X)$, $\sigma_{XY} = \text{Cov}(X,Y)$
- We don't know $\sigma_x^2 = \text{Var}(X)$, $\sigma_x^2 = \text{Var}(X)$, $\sigma_{XY} = \text{Cov}(X,Y)$ in reality!
- What we can do then??

A Toy Example

- We estimate them: $\hat{\sigma}_X^2 = \mathrm{Var}(X)$, $\hat{\sigma}_X^2 = \mathrm{Var}(X)$, $\hat{\sigma}_{XY} = \mathrm{Cov}(X, Y)$
- Using a data set that contains past measurements for X and Y
- We can then estimate the value of α that minimizes the variance of our investment
$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$
- Simulated 100 pairs of returns for the investments X and Y
- These return are used to estimate $\sigma_x^2 = \mathrm{Var}(X)$, $\sigma_x^2 = \mathrm{Var}(X)$, $\sigma_{XY} = \mathrm{Cov}(X, Y)$
- These estimates are substituted to find $\hat{\alpha}$

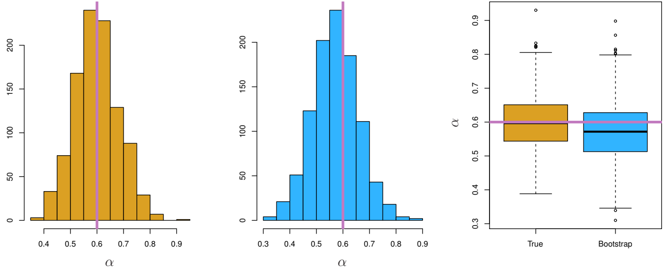
A Toy Example

- Each panel displays 100 simulated returns for investments X and Y
- From left to right and top to bottom, the resulting estimates for α are 0.576, 0.532, 0.657, and 0.651



Accuracy of α

- Natural question: quantify the accuracy of our estimate of α
- To estimate the standard deviation of $\hat{\alpha}$:
  - the process of simulating 100 paired observations of X and Y
  - and estimating α using
  - 1,000 times
  - we obtained $\hat{\alpha}_1, \ldots, \hat{\alpha}_{1000}$



Accuracy of α

MONASH University

- For these simulations the parameters were set $\sigma_X^2 = 1,\ \ \sigma_X^2 = 1,\ \sigma_{XY} = 1.25$
- we know that $\alpha = 0.6$ (solid vertical line on the histogram)
- The mean is

$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996$$

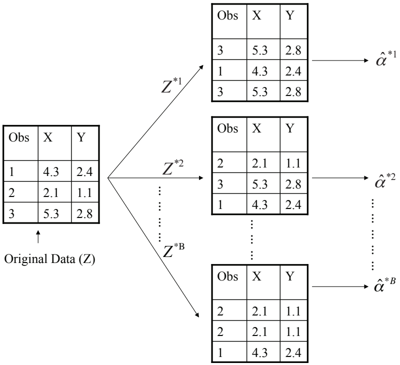
- the standard deviation of the estimates

$$\sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$$

- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $\mathrm{SE}(\hat{\alpha}) \approx 0.083$
- So roughly speaking
  - for a random sample from the population
  - we would expect $\hat{\alpha}$ to differ from α by approximately 0.08, on average.

Bootstrap Example

- a small sample containing $n = 3$ observations
- Each bootstrap data set contains n observations, sampled with replacement from the original data set
- Each bootstrap data set is used to obtain an estimate of α



Bootstrap Example

- a simple data set, which we call Z, that contains only $n = 3$ observations
- We randomly select n observations from the data set in order to produce a bootstrap data set
- The sampling is performed with replacement
- the same observation can occur more than once in the bootstrap data set
- Repeat B times
  - different bootstrap data sets $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$
  - estimates of α are $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$
  - standard error of these bootstrap estimates

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

  - This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set

Summary

- Cross Validation
- Bootstrap
- Reading materials:
  - "Resampling Methods", Chapter 5 of "Introduction to Statistical Learning", 6th edition
- References:
  - Figures in this presentation were taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani