MSSC 6250 / Statistical Machine Learning

Instructor: Mehdi Maadooliat

RESAMPLING METHODS

- Chapter 05

Department of Mathematics, Statistics and Computer Science

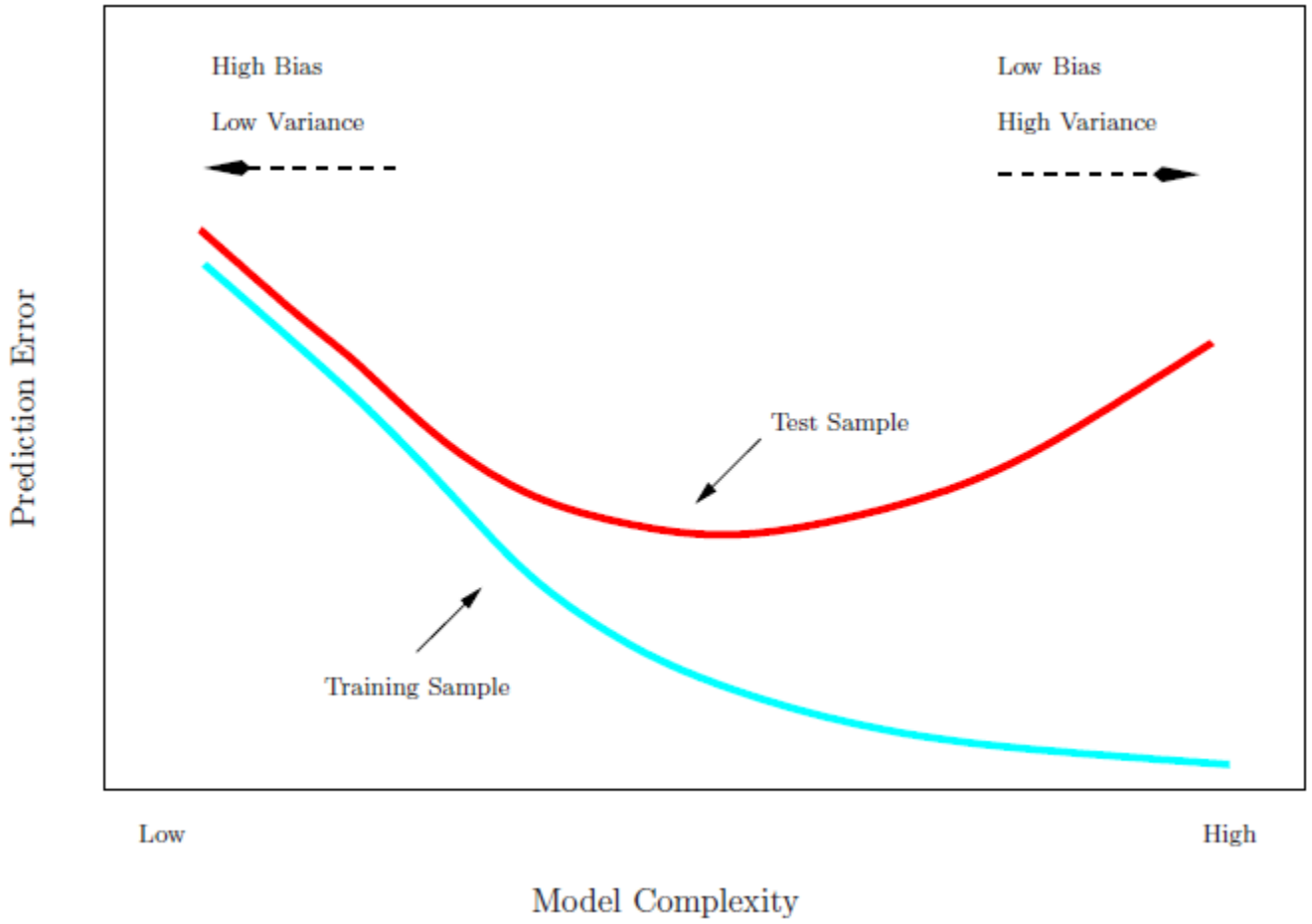
OUTLINE

- Cross Validation
  - The Validation Set Approach
  - Leave-One-Out Cross Validation
  - K-fold Cross Validation
  - Bias-Variance Trade-off for k-fold Cross Validation
  - Cross Validation on Classification Problems
- Bootstrap

Training Error versus Test error

- Recall the distinction between the *test error* and the *training error:*
- The *test error* is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the *training error* can be easily calculated by applying the statistical learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can *dramatically underestimate* the latter.
- Best solution: a large designated test set. Often not available

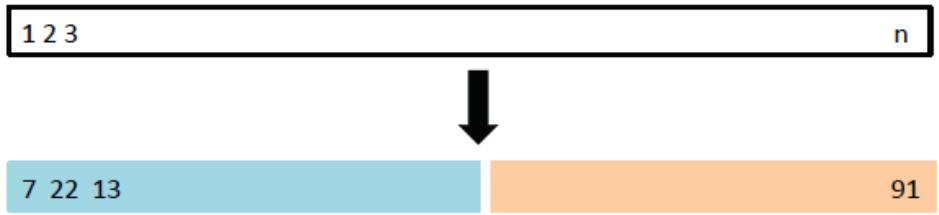
Training- versus Test-Set Performance



- Here we instead consider a class of methods that estimate the test error 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

VALIDATION-SET APPROACH

- Here we randomly divide the available set of samples into two parts: a *training set* and a *validation* 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.

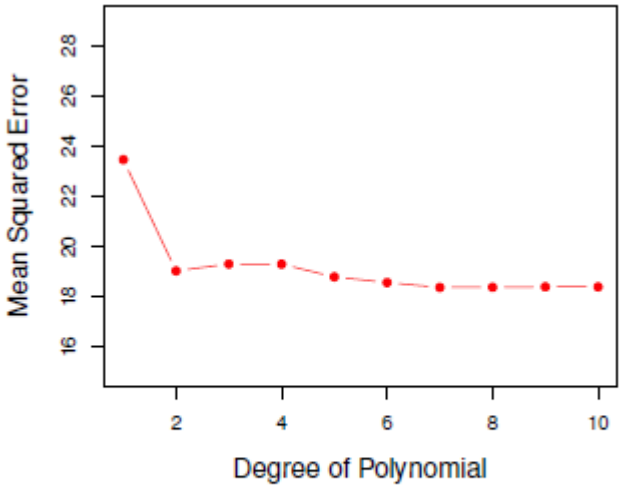


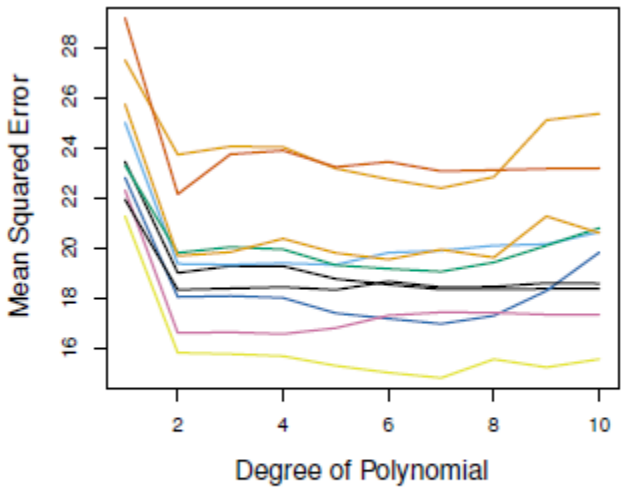
A random splitting into two halves: left part is training set, right part is validation set

- The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

EXAMPLE: AUTOMOBILE DATA

- Want to compare linear vs higher-order polynomial terms in a linear regression
- We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.





Left panel shows single split; *right panel shows multiple splits*

DRAWBACKS OF VALIDATION SET APPROACH

- the validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

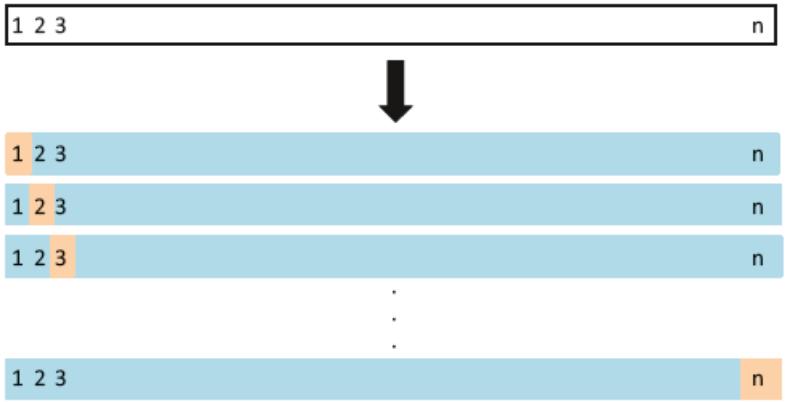
- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.

- This suggests that the validation set error may tend to *overestimate* the test error for the model fit on the entire data set.

K-FOLD CROSS-VALIDATION

- *Widely used approach* for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.

- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.



- This is done in turn for each part $k = 1, 2, \ldots K$, and then the results are combined.

The Details

- Let the K parts be $C_1, C_2, \ldots C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k: if N is a multiple of K, then $n_k = n/K$.

- Compute

$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

  where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.

- Setting $K = n$ yields n-fold or *leave-one out cross-validation* (LOOCV).

A NICE SPECIAL CASE!

- With least-squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit! The following formula holds:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

  where $\hat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is the leverage (diagonal of the "hat" matrix; see book for details.) This is like the ordinary MSE, except the ith residual is divided by $1 - h_i$.

- LOOCV sometimes useful, but typically doesn't *shake up* the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.

- a better choice is $K = 5$ or 10.

AUTO DATA REVISITED

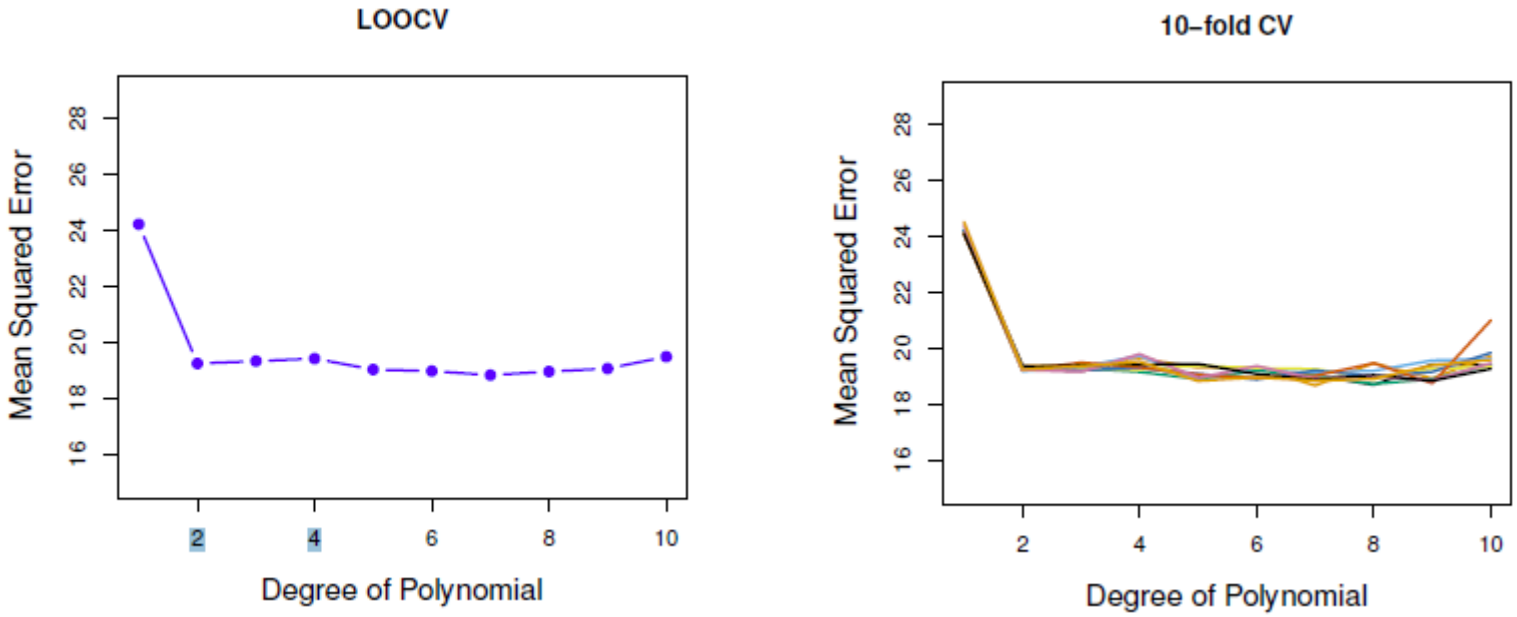
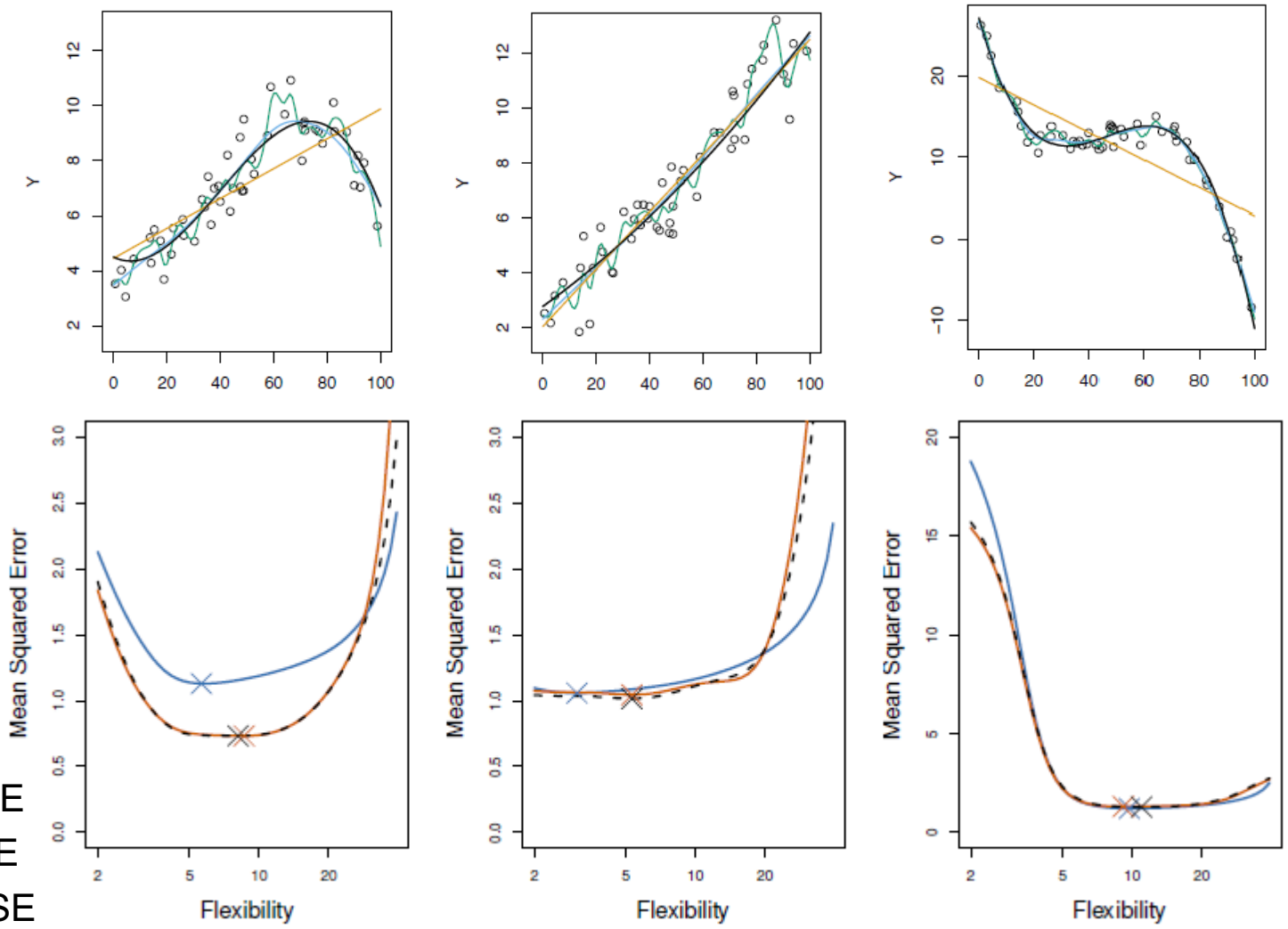


FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *The LOOCV error curve.* Right: 10-*fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

LOGISTIC REGRESSION WITH SEVERAL VARIABLES



- Blue: True Test MSE
- Black: LOOCV MSE
- Orange: 10-fold MSE
- Refer to chapter 2 for the top graphs, Fig 2.9, 2.10, and 2.11

LOOCV vs. the Validation Set Approach

- LOOCV has less bias
  - We repeatedly fit the statistical learning method using training data that contains n-1 obs., i.e. almost all the data set is used
- LOOCV produces a less variable MSE
  - The validation approach produces different MSE when applied repeatedly due to randomness in the splitting process, while performing LOOCV multiple times will always yield the same results, because we split based on 1 obs. each time
- LOOCV is computationally intensive (disadvantage)
  - We fit the each model n times!

Bias-Variance Trade-off for k-fold CV

- Putting aside that LOOCV is more computationally intensive than k-fold CV…
  Which is better LOOCV or K-fold CV?
  - LOOCV is less bias than k-fold CV (when $k < n$)
  - But, LOOCV has higher variance than k-fold CV (when $k < n$)
  - Thus, there is a trade-off between what to use

- Conclusion:
  - We tend to use k-fold CV with ($K = 5$ and $K = 10$)
  - These are the magical K's ☺
  - It has been empirically shown that they yield test error rate estimates that suffer neither from excessively high bias, nor from very high variance

Cross-Validation for Classification Problems

- We divide the data into K roughly equal-sized parts $C_1, C_2, \ldots C_K$. C_k denotes the indices of the observations in part k. There are n_k observations in part k: if n is a multiple of K, then $n_k = n/K$.

- Compute

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{Err}_k$$

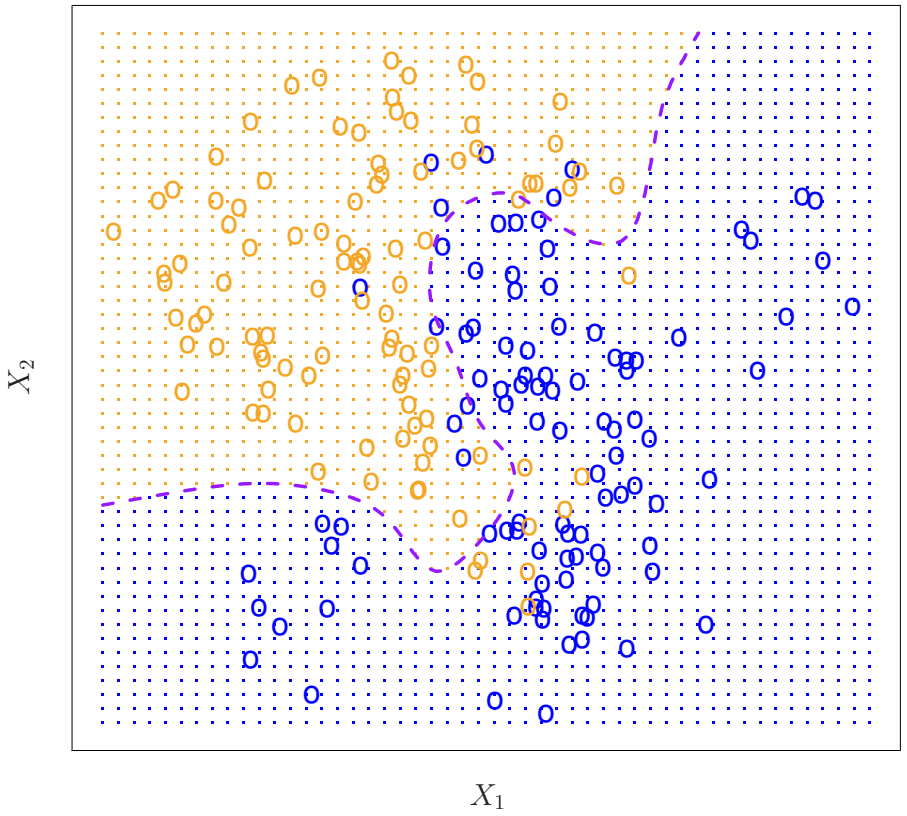
  where $\mathrm{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

- The estimated standard deviation of CV_K is

$$\widehat{\mathrm{SE}}(\mathrm{CV}_K) = \sqrt{\sum_{k=1}^{K} (\mathrm{Err}_k - \overline{\mathrm{Err}_k})^2/(K-1)}$$

CV TO CHOOSE: ORDER OF POLYNOMIAL

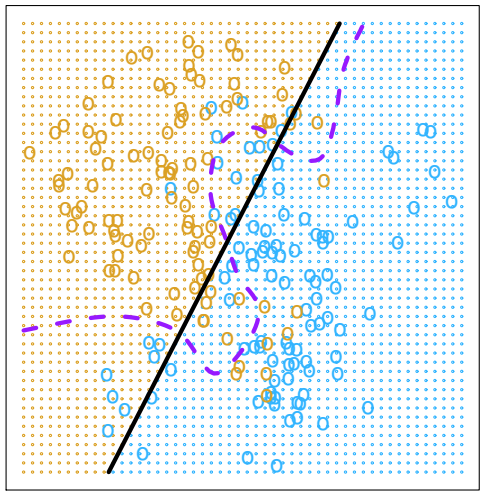
- The data set used is simulated (refer to Fig 2.13)
- The purple dashed line is the Bayes' boundary



Bayes' Error Rate: 0.133

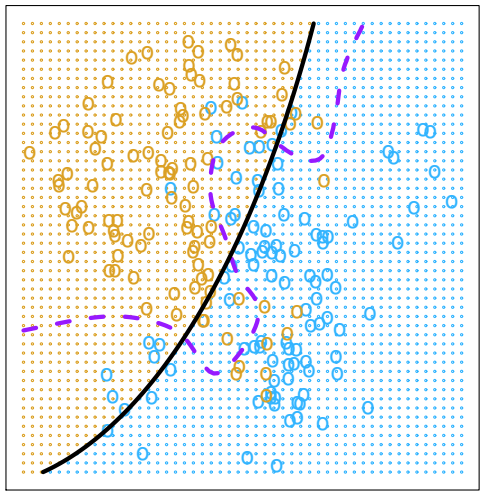
CV to Choose: Order of Polynomial Cont…

Degree=1



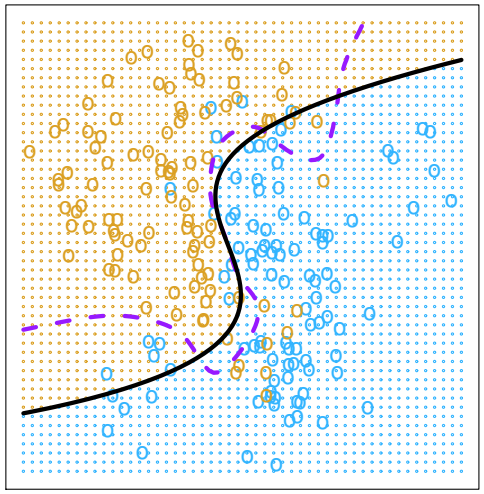
Error Rate: 0.201

Degree=2



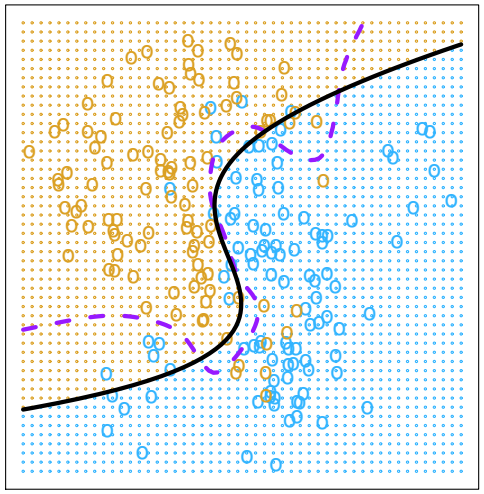
Error Rate: 0.197

Degree=3



Error Rate: 0.160

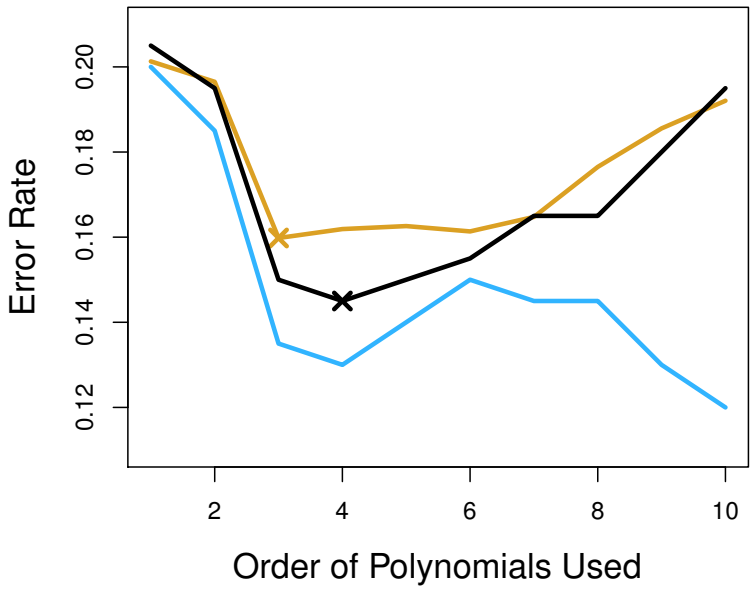
Degree=4

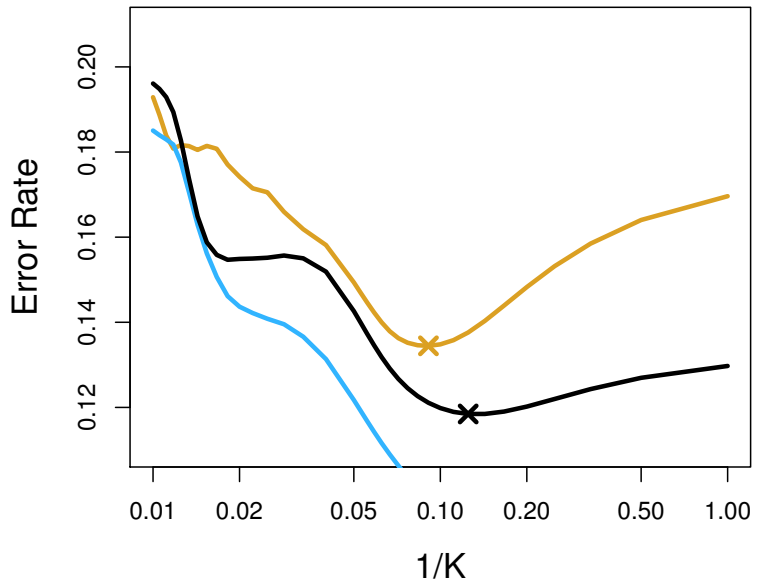


Error Rate: 0.162

MARQUETTE UNIVERSITY
Be The Difference.

CV TO CHOOSE: ORDER OF POLYNOMIAL CONT…





- Brown: Test Error
- Blue: Training Error
- Black: 10-fold CV Error

CROSS-VALIDATION: RIGHT AND WRONG

- Consider a simple classifier applied to some two-class data:
  1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
  2. We then apply a classifier such as logistic regression, using only these 100 predictors.

  How do we estimate the test set performance of this classifier?

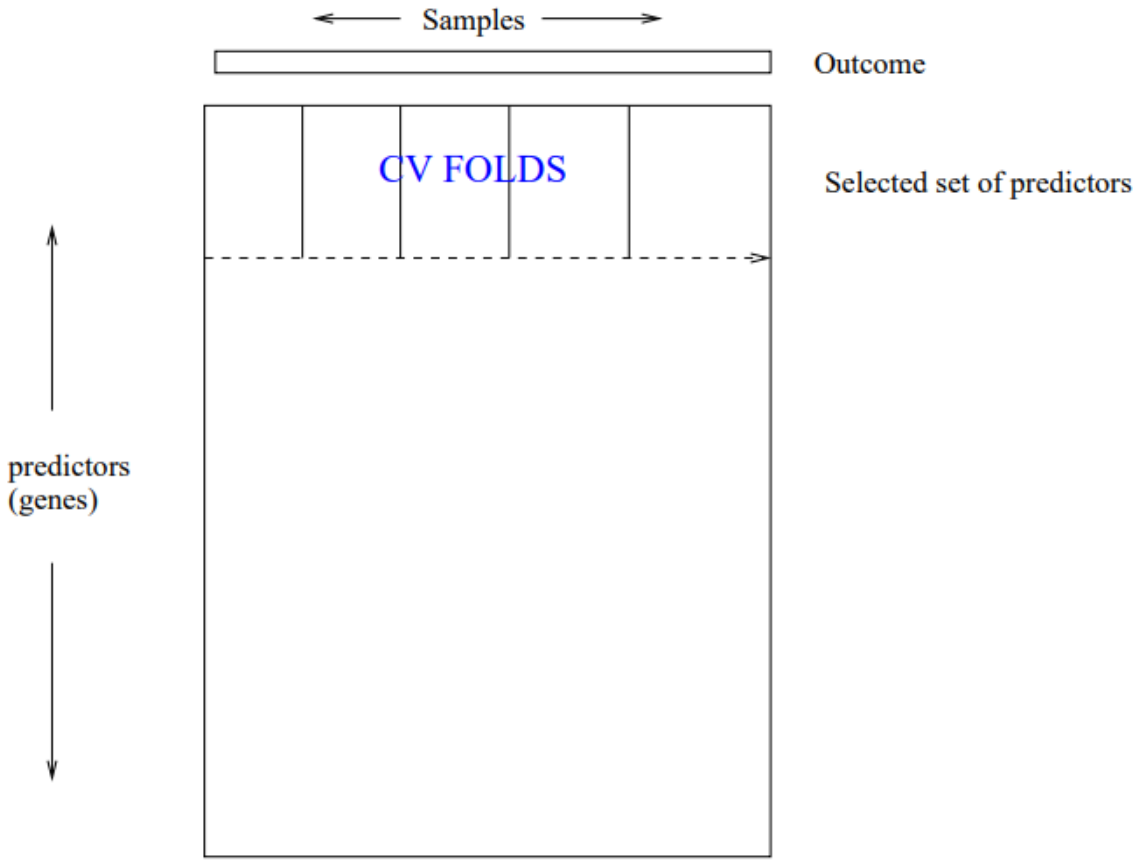
Can we apply cross-validation in step 2, forgetting about step 1?

- This would ignore the fact that in Step 1, the procedure *has already seen the labels of the training data*, and made use of them. This is a form of training and must be included in the validation process.

The Wrong and Right Way

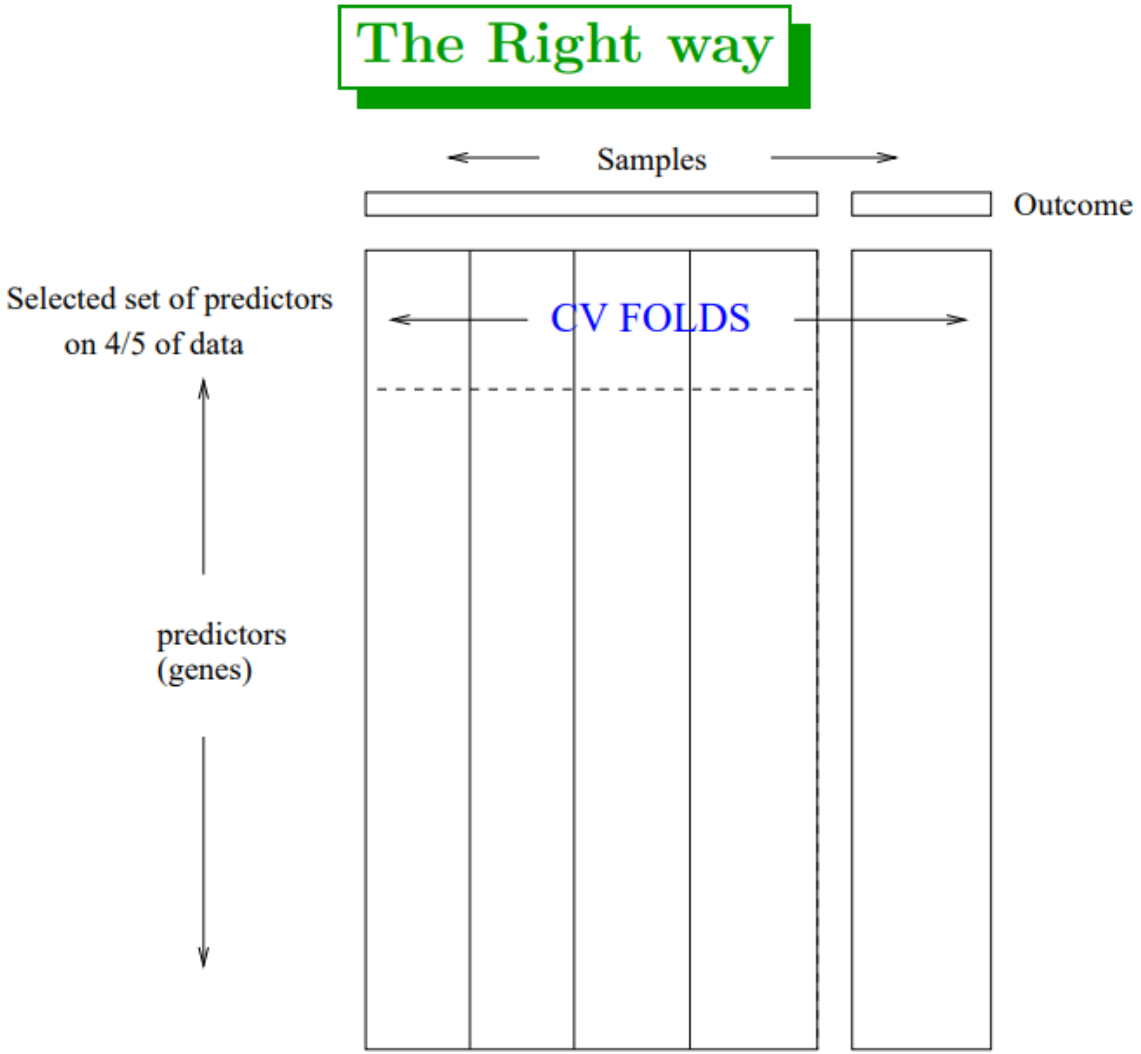
- *Wrong:* Apply cross-validation in step 2.

A little cheating goes a long way



The Wrong and Right Way

- *Right:* Apply cross-validation to steps 1 and 2.



QUESTIONS?

- ANY QUESTION?