CROSS VALIDATION

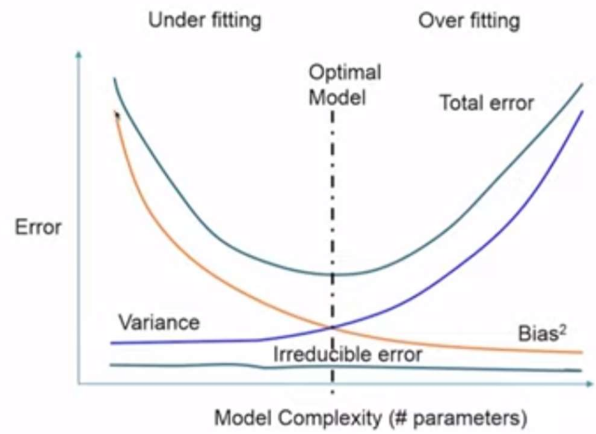
Motivation

- How to select the optimal number of meta or hyper-parameters of a model?
  - Number of principal components in principal components analysis
  - Number of clusters in K-means clustering
  - Number of terms 'n' in polynomial or nonlinear regression

    $$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \beta_n x^n$$

    (equivalent to multilinear regression by treating x, x^2, …x^n as different variables)
- MSE of training data set not useful as a measure
  - MSE will decrease with increasing number of parameters (can be reduced to zero)
- Use cross validation on a validation data set to determine optimal number of parameters

Bias-Variance trade-off on test data set



Training and Validation data sets

- For large data sets divide data set into training data set (~ 70% of the samples) and remaining validation/test data
  - Training set: $\{(\boldsymbol{x}_1, y_1);(\boldsymbol{x}_2, y_2);\ldots;(\boldsymbol{x}_n, y_n)\}$
  - Test set: $(\boldsymbol{x}_{0,i}, y_{0,i})$: i = 1...n_t observations
- Training error rate

$$MSE_{Training} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2$$

- Test error rates

$$\boxed{MSE_{Test} = \frac{1}{n_t}\sum_{i=1}^{n}(y_{0,i} - \mathbf{x}_{0,i}^T\hat{\boldsymbol{\beta}})^2}$$

Training and Validation data sets

- For large data sets divide data set into training data set (~ 70% of the samples) and remaining validation/test data
  - Training set: $\{(\boldsymbol{x}_1, y_1);(\boldsymbol{x}_2, y_2);\ldots;(\boldsymbol{x}_n, y_n)\}$
  - Test set: $(\boldsymbol{x}_{0,i}, y_{0,i})$: i = 1...n_t observations
- Training error rate

$$MSE_{Training} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2$$

Not of our interest for predictive ability of the model

- Test error rates

$$\boxed{MSE_{Test} = \frac{1}{n_t}\sum_{i=1}^{n}(y_{0,i} - \mathbf{x}_{0,i}^T\hat{\boldsymbol{\beta}})^2}$$

Of our interest

Data scarcity: Test data are not available

Validation Set Approach

- Enough data: (1) Training set, (2) Validation set, and (3) Test set
- Not enough data: Generate validation sets from a training set
- Validation set approach: Divides (often randomly) the training set into two parts

1 2 3 4 **n**

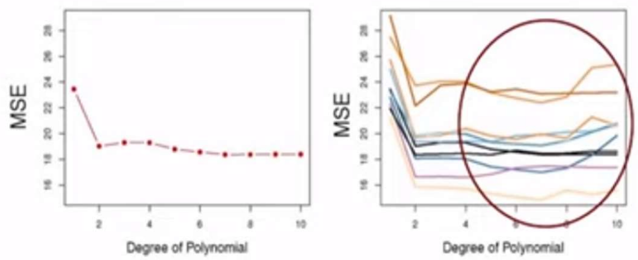
- A training set **1 2 3 4** **n_t**
- A validation set (or hold-out set) **1, 2 3 4** **n_v**

- Use training set, to fit the model
- Use validation set, to predict validation set errors

Provides an estimate of test error rates

Validation Set Approach: Example

- Example: mileage~ horsepower[1] (> 300 data points on horsepower of automobiles and mileage)
- Polynomial Model: mileage~*f*(horsepower)



High variability in estimates of test error

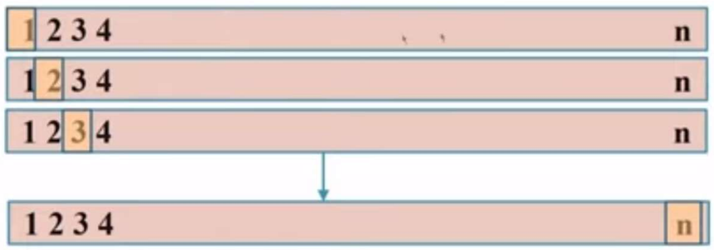
[1]Tibshirani et al (2013)

Sampling for small data sets

- Validation of models by repeatedly drawing random samples from a training set
  - Validation set (random sampling)
  - K-fold cross validation
  - Bootstrap
- Objective: Predict the performance of model(s) on the validation/test sets (drawn from training data)
- Resampling methods useful for data scarce situations

Leave-one-out-cross-validation (LOOCV)

- Build model using *(n-1)* samples and predict the response (y_i) for *the remaining sample*

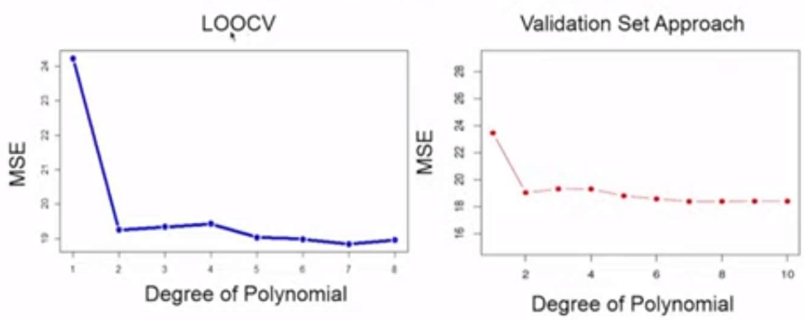


$$CV_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(1)})^2$$

GyanData Private Limited

LOOCV: Example

- Example: mileage~ horsepower[1]
- Nonlinear Model: mileage~*f*(horsepower)



[1]Tibshirani et al (2013)

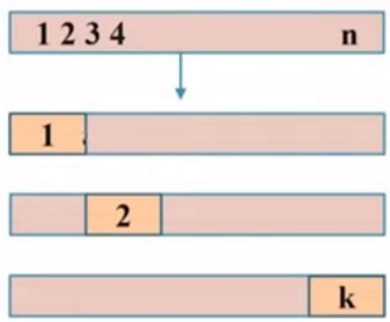
LOOCV

- Leave-one-out-cross-validation (LOOCV)
- Advantages
  - Far less bias comparison to the validation set approach
    Training set contains *(n-1)* observations each iteration
  - Yield the same results
    No randomness in the training/validation set splits
  - Does not overestimate the test error rate as much as the validation set approach
- Disadvantages
  - Expensive to implement due to fitting happens *n* times
  - It may select a model of excessive size (more variables) than the optimal model

k-Fold Cross Validation

- Training data into *k* disjoint samples of equal size,
  $Z_1, Z_2 \ldots, Z_k$
- For each validation sample Z_i
  - Use remaining data to fit the model
  - Predict the response for the validation sample Z_i and compute mean square error (MSE_i),
  - Repeat for all *k* samples
  - The k-fold CV

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

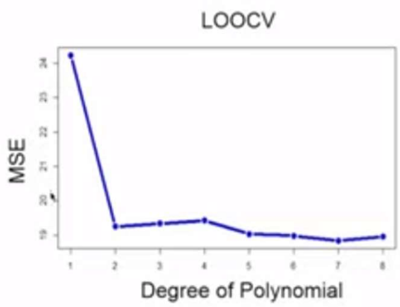


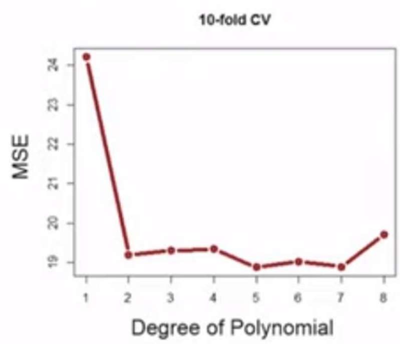
k-fold Validation

- For $k=n$, Leave-one-out-cross-validation (LOOCV)
- In practice, k=5 or 10 is taken,
- Less computation cost
- For computationally intensive learning methods
  - LOOCV fits the model n times
  - k-fold CV fits the model k times

k-fold CV: Example

- Example: mileage~ horsepower[1]
- Nonlinear Model: mileage~f(horsepower)





[1]Tibshirani et al (2013)