INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo
Dept. of Bio-industrial Mechatronics Engineering
National Taiwan University

Today:

- Overfitting
- Model attribute selection

Outline

- Goal of the lecture
- Overfitting
- Bias-variance tradeoff
- Cross-validation
- Information criteria

Goals

- After this, you should be able to:
  - Understand the risk of overfitting
  - Quantitatively assess the performance of different models
  - Calculate information criteria for regression models
  - Estimate model prediction error using cross-validation

Review – Sum of Squared Errors

- Sum of squared errors (SSE) is a measure of the discrepancy between the data and the model estimation

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- Note that it is calculated with the same set of data (i.e., training data) that are used to generate the model $\hat{y}$
- Also known as residual sum of squares (RSS)
- Typically when SSE decreases, R^2 increases

Response Surface Methodology

- The true relationship between the response variable y and the explanatory variables $\boldsymbol{x}$ are usually unknown
- The approximation of the response variable using a set of explanatory variable is called response surface methodology
- In many practical application, high-order polynomial models are employed, i.e.,

$$\hat{y} = f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

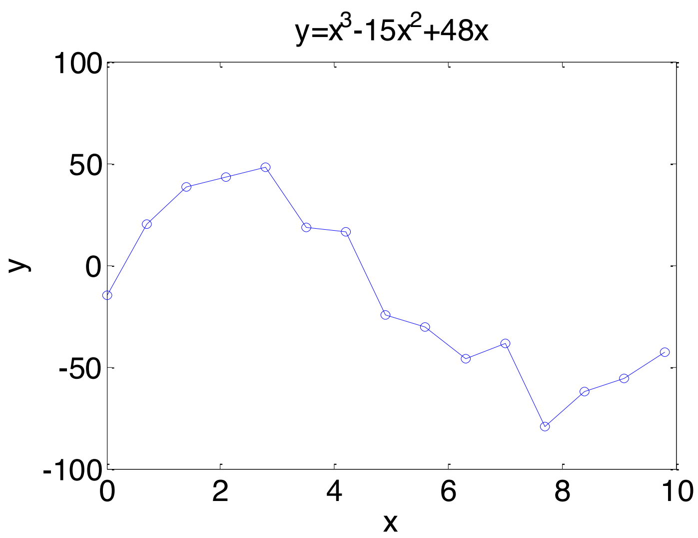
$$\hat{y} = f(x_1, x_2) = \beta_0 + \beta_{11} x_1 + \beta_{12} x_2 + \beta_{21} {x_1}^2 + \beta_{22} {x_2}^2 + \cdots$$

- High-order polynomial models usually gives larger R^2, but that means high-order models better???

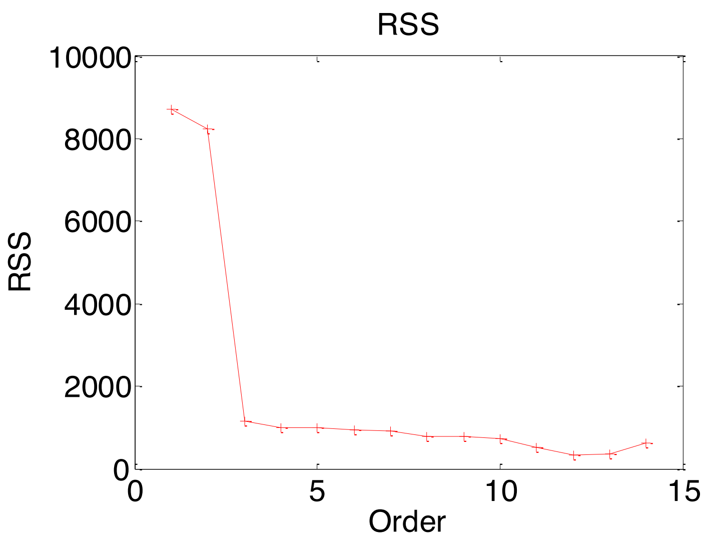
Example – Polynomial Curve Fitting

- Assume the true model is a 3rd-order polynomial with random noise
- RSS decreases when the order of the polynomial increases

3rd-order Polynomial with **Noise**



RSS of Models with Different Orders



Example MATLAB Code

```
clear; close all;
% generate data
x=(0:.7:10)';
y=x.^3-15*x.^2+48*x+10*randn(length(x),1);
plot(x,y, '-ob'); set( gcf, 'Color', 'w');
xlabel('x', 'FontSize', 16); ylabel('y', 'FontSize', 16);
set(gca,'FontSize', 16); title('y=x^3-15x^2+48x');

% regress y with different order of x
for i=1:length(x)-1
    X(:,i)=power(x,i);
    [b,bint,r]=regress(y,[ones(length(x),1) X(:,1:i)]);
    RSS(i)=r'*r;
end
figure; plot((1:14), RSS, '-+r'); set( gcf, 'Color', 'w');
set(gca,'FontSize', 16); xlabel('Order', 'FontSize', 16);
ylabel('RSS', 'FontSize', 16); title('RSS');
```

Model Complexity

- Polynomial curve fitting, order $M = 3$

$$\hat{y} = f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{j=1}^{q} \beta_{ij} x_i x_j + \sum_{i=1}^{q} \sum_{j=1}^{q} \sum_{k=1}^{q} \beta_{ijk} x_i x_j x_k \,,$$

where q is the number of variables

- Number of explanatory variables for the polynomial becomes unwieldy very quickly
- Intuitively, adding more attributes (or higher order terms) would improve the model, as more information never hurts, right?

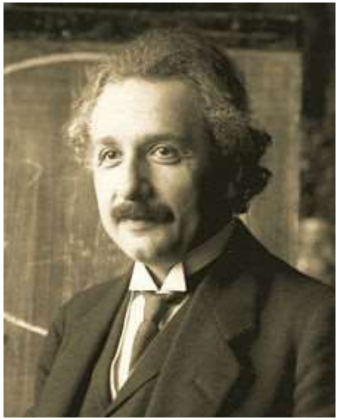
Occam's Razor

Occam's razor:

"*entia non sunt multiplicanda praeter necessitatem*"

Entities should not be multiplied beyond necessity

William of Occam

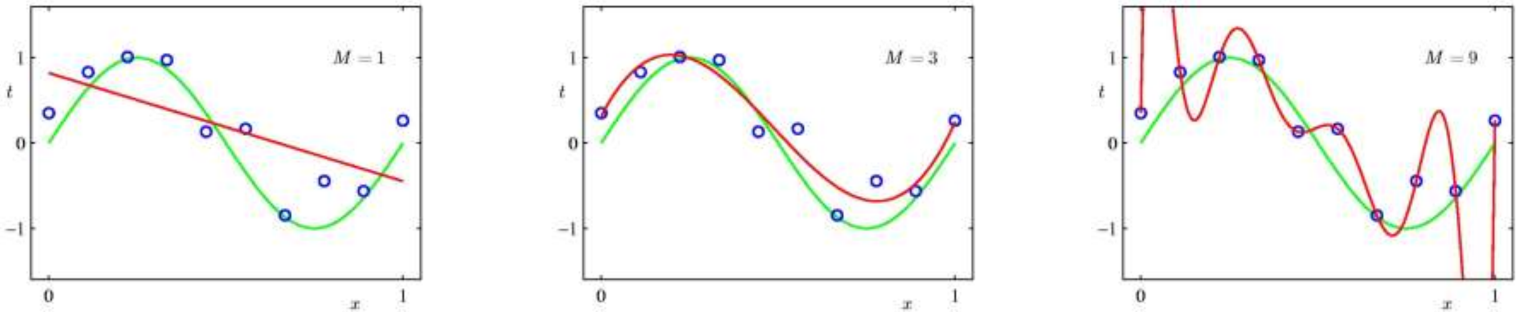


Everything should be made as simple as possible, but not simpler

Albert Einstein

Overfitting – the Problem of Complex Model

- The phenomenon in which a model well predicts the outcome with the data points used to develop the model, but subsequently fails to provide valid predictions in unseen cases
- Data drawn from a sinusoidal model $\sin(x/_{2\pi})$ with noise



Bishop. *Pattern Recognition and Machine Learning*

- Overfitting happens when a model is capturing idiosyncrasies of the data rather than generalities

Understanding Overfitting

- What is the source of overfitting?
- Why do some models overfit more than others?
- How does one tackle overfitting?

Model Performance Estimation

- A loss function is defined as the difference between the true and estimated target values, i.e.,

$$L\big(y, f(\boldsymbol{x})\big) = \begin{cases} \big(y - f(\boldsymbol{x})\big)^2 & \text{squared error} \\ |y - f(\boldsymbol{x})| & \text{absolute error} \end{cases},$$

where f is the model that estimates the target $y \in \Re$ from measurements $\boldsymbol{x} \in \Re^M$, i.e., $\hat{y} = f(\boldsymbol{x})$

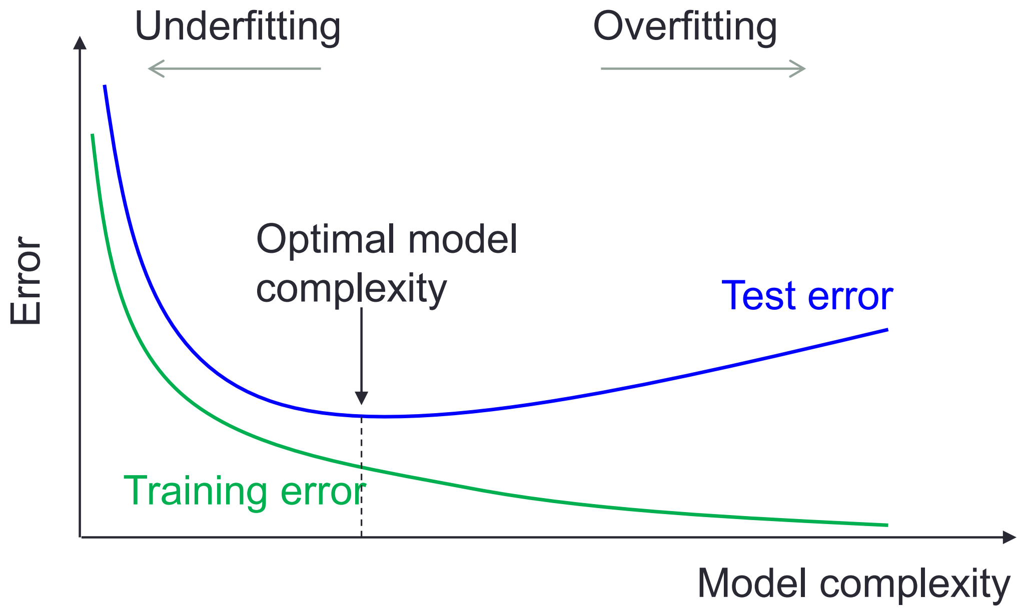
- The training error is $\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f_i(\boldsymbol{x})\big)$, where N is the number of samples
- The test (generalization) error is $err = E\big[L\big(y, f(\boldsymbol{x})\big)\big]$

Measuring Model Complexity

- The complexity of the model can be measured by the "degrees of freedom" (number of parameters)
- What is the complexity of this model?

$$\hat{y} = f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

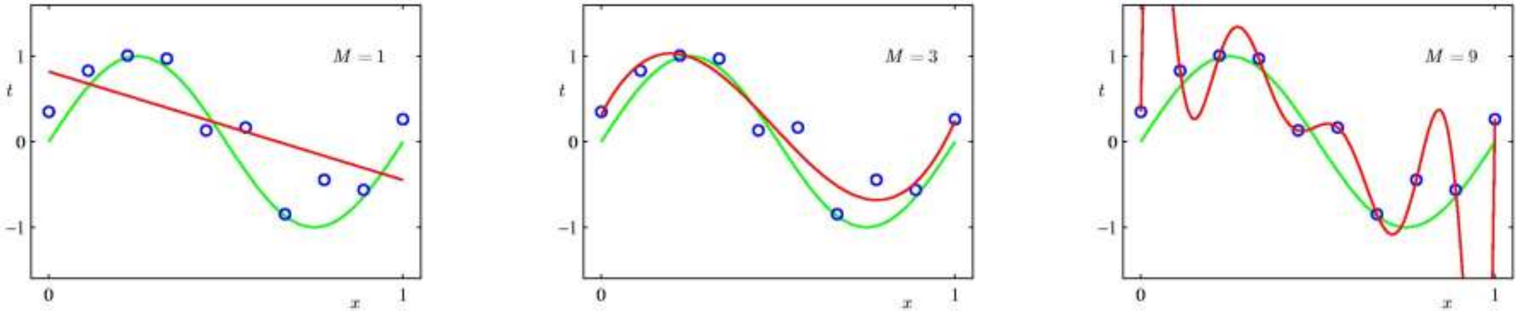
Optimal Model Complexity



- Training error usually monotonically decreasing with increase in model complexity
- Test error bounced back once the complexity increases

Example of Model Complexity vs. Errors

- Errors of the 1st-order, 3rd-order, and 9th-order models
- True model vs. Measured data vs. Regression model
- The distance between the true model and the regression model increases when the complexity of the model increases



Bishop. *Pattern Recognition and Machine Learning*

Bias and Variance of An Estimator $\hat{y}$

- An estimator $\hat{y}$ is a model for estimating a parameter y

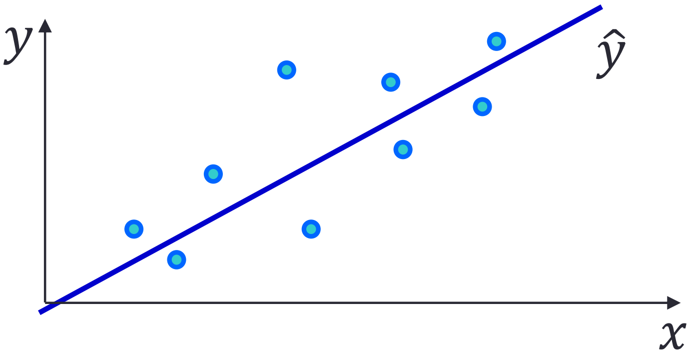
- Bias of $\hat{y}$ is the difference between the estimator's expected value $E[\hat{y}]$ and the true value y, i.e., $Bias(\hat{y}) = E[y - E[\hat{y}]]$

  (NOTE: true value y is usually unknown)

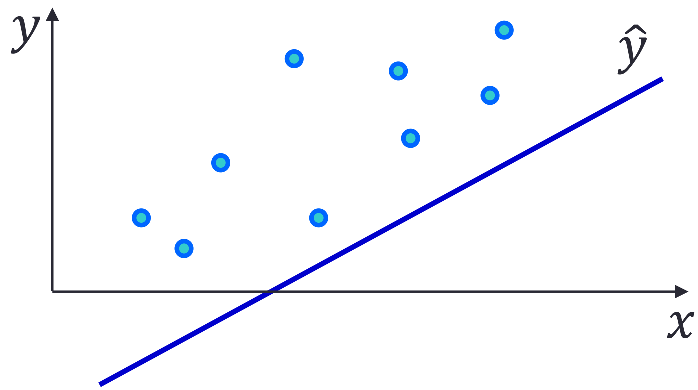
- Variance of $\hat{y}$ is a measure of how far a set of numbers are spread out from each other, i.e., the variance of an estimator $\hat{y}$ is $Var(\hat{y}) = E[(\hat{y} - E[\hat{y}])^2]$

Illustration of Bias and Variance

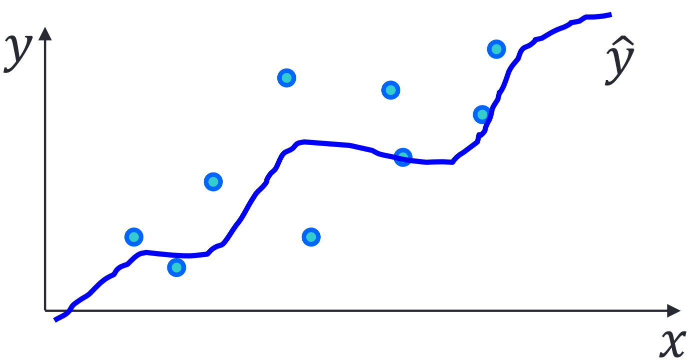
Medium bias and low variance



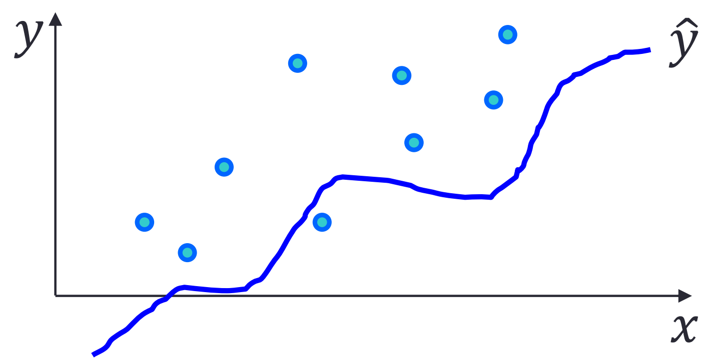
High bias and low variance



Low bias and high variance



High bias and high variance



Bias and Variance vs. Model Complexity

- $(Test\ error)^2 = Bias^2(\hat{y}) + Var(\hat{y})$
- The bias decreases when the model complexity increases
- The variance increases when the model complexity increases
- An estimator is said to be unbiased if its bias is equal to zero for all values of model parameter $\boldsymbol{\beta}$
- High-order ordinary least squares (OLS) estimators often have low bias but large variance (especially when the degree of model complexity is high)

Bias-variance Decomposition of the "Test Error"

- The expectation value of the "squared test error" can be decomposed to:

$$(Test\ error)^2 = E[(y-\hat{y})^2] = E[\{(y-E[\hat{y}]) + (E[\hat{y}]-\hat{y})\}^2]$$

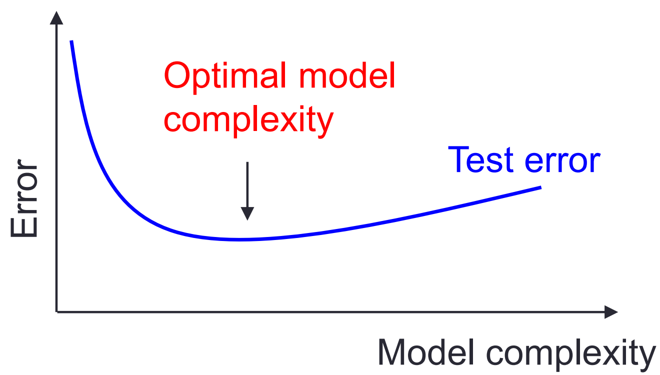
$$= E[(y-E[\hat{y}])^2] + E[(E[\hat{y}]-\hat{y})^2] + 2E[(y-E[\hat{y}])(E[\hat{y}]-\hat{y})\}]$$

$$= Bias^2(\hat{y}) + Var(\hat{y}) + 2E\left[yE[\hat{y}] - y\hat{y} - (E[\hat{y}])^2 + \hat{y}E[\hat{y}]\right]$$

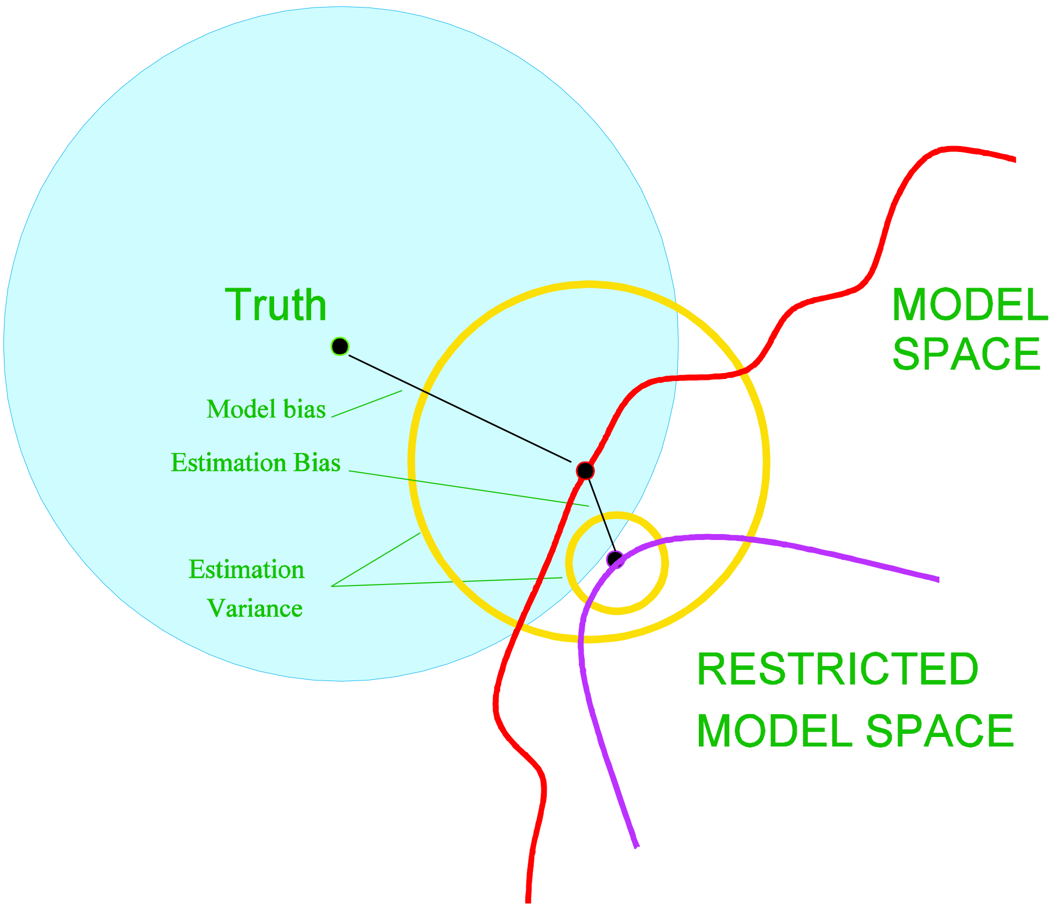
$$= Bias^2(\hat{y}) + Var(\hat{y}) + 2(\cancel{E[y]E[\hat{y}]} - \cancel{E[y\hat{y}]} - \cancel{(E[\hat{y}])^2} + \cancel{(E[\hat{y}])^2})$$

Model Complexity and Bias-variance Tradeoff

- Intuition for the bias-variance trade-off:
  - Complex model $\Rightarrow$ sensitive to data $\Rightarrow$ much affected by changes in $\boldsymbol{x}$ $\Rightarrow$ high variance, low bias
  - Simple model $\Rightarrow$ more rigid $\Rightarrow$ does not change as much with changes in $\boldsymbol{x}$ $\Rightarrow$ low variance, high bias
- **One of the most important goals in machine learning is to find a model with the optimal model complexity so it gives the least test error**



Accuracy of Model and Restricted Model



Summary of Overfitting

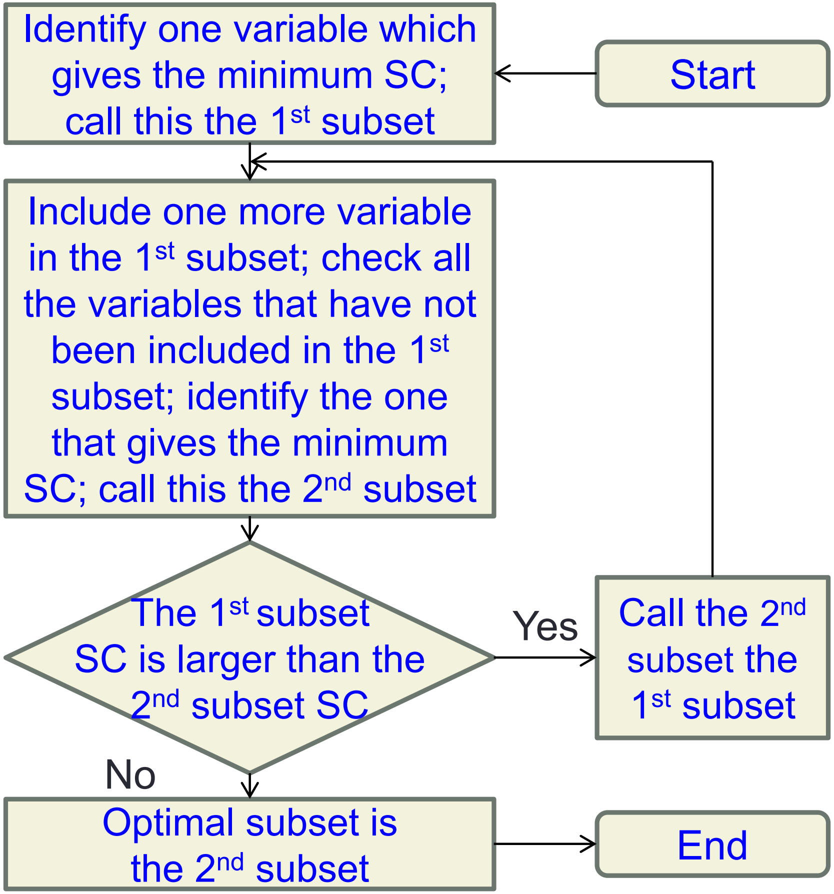
- Overfitting happens when the fitted model is too complex
- The degree of overfitting depends on model complexity and training data availability (why?)
- This means that overfitting is NOT a problem if there are infinitely many data points
- If a model overfits, it will be unstable – that means, removal part of the data will change the fit significantly

Tackle Overfitting – Variable Selection

- Variable selection – retain only the subset of variables that gives the “best fit“
- Direct greedy search of the best variable subset can take a long time of the number of variables is large
- Some typical variable selection search methods:
- Forward selections: starts with the intercept and add at each step the predictors that most improves the fit
- Backward elimination: starts with the full model and removes one by one the worst explanatory variable
- Stepwise selection: combines forward and backward to decide at each step which variable to remove and/or to add

Example: Forward Selection with Selection Criteria

- Choose a selection criteria (SC)
- The process stops when including further variables do not improve selection criteria



Variable Selection Criteria

- The variables are selected based on some methods:
  1. Cross-validation
  2. Information criteria
  3. Partial F-test

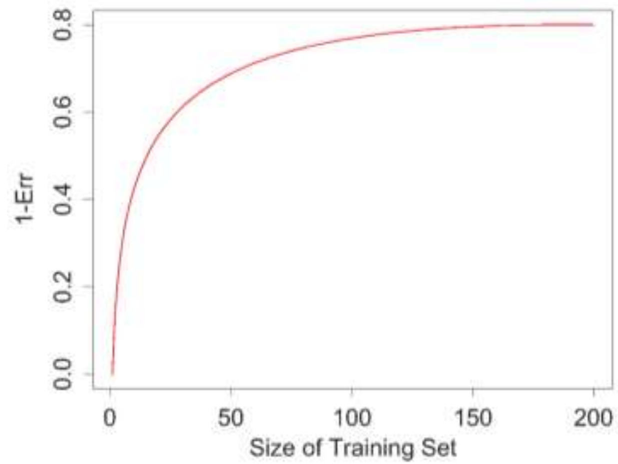
Method 1 – Cross-validation (CV)

- One of the most often used technique for model performance estimation
- k-fold cross-validation:
  1. Partition data into k roughly equal parts
  2. Train with all but jth part, test on the jth part
  3. The procedure is executed for a total of k times until all parts have been the test part
  4. The k error rates are averaged to yield an overall error estimate
- Leave-one-out cross-validation, i.e., $k = N$

Choice of k

- If $k = N$ then CV is approximately unbiased, but has high variance
- On the other hand, with $k = 5$, CV has low variance but more bias
- Typically $k = 10$ is chosen
- Increasing the number of k also increases the computational burden

Hypothetical Learning Curve with $k = 5$



Hastie, Tibshirani, and Friedman.
The Elements of Statistical Learning.

- How many samples do we need for cross validation?

Method 2 – Information Criteria

- The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are defined as:

$$AIC = N \cdot \ln\left(\frac{RSS}{N}\right) + 2 \cdot k$$

Accuracy Complexity

$$BIC = N \cdot \ln\left(\frac{RSS}{N}\right) + \ln(N) \cdot k,$$

Accuracy Complexity

where RSS is the residual sum of squares from regression, and k denote the number of model parameters, i.e., $k = \text{size}(\boldsymbol{\beta})$

Information Criteria – AIC and BIC

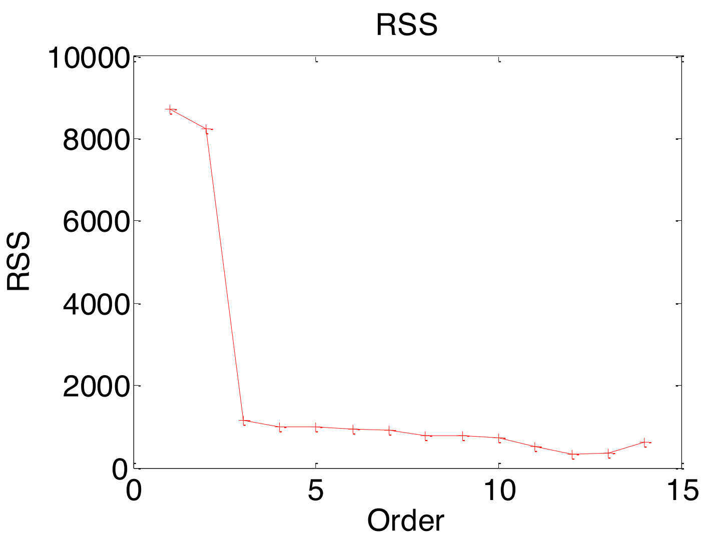
- A direct measure of training error with penalty of complexity
- A trade-off between model complexity and accuracy
- Information criteria tend to penalize complex models, giving preference to simpler models in selection
- A model associated with a smaller AIC or BIC value is preferred

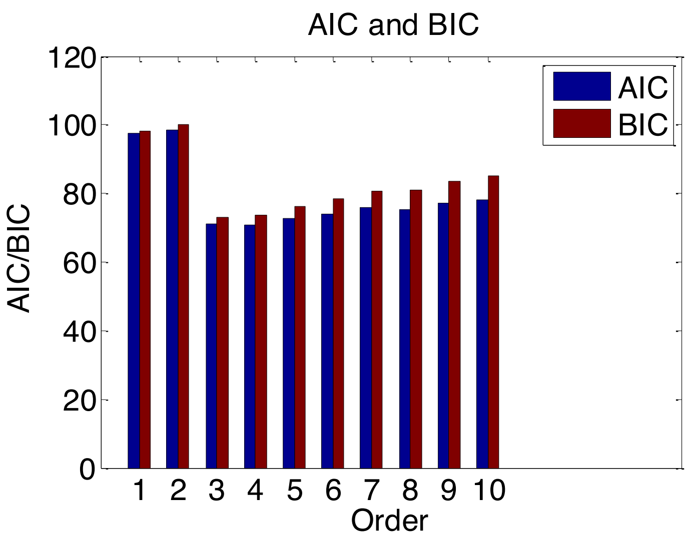
AIC or BIC?

- BIC is asymptotically consistent as a selection criterion – given a family of models including the true model, the probability that BIC will select the correct one approaches one as the sample size becomes large, i.e., $N \to \infty$
- AIC does not have the above property; instead, it tends to choose more complex models as $N \to \infty$
- For small or moderate samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity

AIC and BIC of A Previous Example

- Higher order models are strongly penalized by AIC and BIC





Example MATLAB Code

```
% Calculate and plot AIC and BIC
for i=1:length(x)-1
    X(:,i)=power(x,i);
    [b,bint,r]=regress(y,[ones(length(x),1) X(:,1:i)]);
    RSS(i)=r'*r;
    AIC(i)=length(y)*log(RSS(i)/length(y))+2*i;
    BIC(i)=length(y)*log(RSS(i)/length(y))+log(length(y))*i;
end

figure; bar( [ AIC(1:10); BIC(1:10)]', 1);
title('AIC and BIC', 'FontSize', 16);
xlabel('Order', 'FontSize', 16); set(gca,'FontSize', 16);
ylabel('AIC/BIC', 'FontSize', 16);
set( gcf, 'Color', 'w'); legend({'AIC', 'BIC'});
```

Information Criteria – Mallows' C_p

- Mallows‘ C_p is defined as:

$$C_p = \frac{RSS_k}{\frac{RSS_{FULL}}{N}} - (N - k),$$

  where RSS_k is the residual sum of squares for the model containing k explanatory variables, and RSS_{FULL} is the residual sum of squares for the model containing all the explanatory variables
- The model that has a C_p value <u>closest to</u> k is considered the best model

Criterion 3 – Partial F-test

- The partial F-test is used to test the significance of one or more variables in the presence of other variable(s) in the full model
- Suppose a $f$ull model: $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$, and a $r$educed model: $\hat{y} = \beta_0 + \beta_2 x^2 + \cdots + \beta_{M-1} x^{M-1}$
- The null hypothesis is the change in sum of squares is not due to changes in model complexity
- Hypothesis: $H_0: \beta_M = 0$, $H_1: \beta_M \neq 0$
- F statistic: $F_{(q,n-p-1)} = \frac{(SSE_r - SSE_f)/q}{MSE_f}$, where the full model has p variables, and the reduced model has q variables

References

- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Chapter 3 and 7
- C. M. Bishop, Pattern Recognition and Machine Learning