Bias-Variance Trade-off in ML

Sargur Srihari
srihari@cedar.buffalo.edu

Bias-Variance Decomposition

- Choosing λ in maximum likelihood/least squares estimation
- Five part discussion:
  1. On-line regression demo
  2. Point estimate

     Chinese Emperor's Height
  3. Formulation for regression
  4. Example
  5. Choice of optimal λ

Bias-Variance in Regression

- Low degree polynomial has high bias (fits poorly) but has low variance with different data sets
- High degree polynomial has low bias (fits well) but has high variance with different data sets

Machine Learning

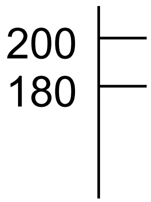
Bias-Variance in Point Estimate

True height of Chinese emperor: 200cm, about 6'6"

Poll a random American: ask "How tall is the emperor?"

We want to determine how wrong they are, on average

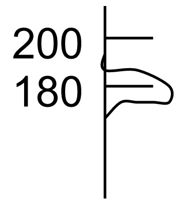
Each scenario has expected value of 180 (or bias error = 20), but increasing variance in estimate



Bias
No variance

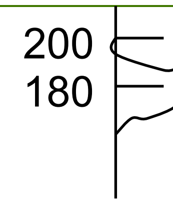
- Scenario 1
- Everyone believes it is 180 (variance=0)
- Answer is always 180
- The error is always -20
- Ave squared error is 400
- Average bias error is 20
- 400=400+0

Bias
Some variance



- Scenario 2
- Normally distributed beliefs with mean 180 and std dev 10 (variance 100)
- Poll two: One says 190, other 170
- Bias Errors are -10 and -30
  - Average bias error is -20
- Squared errors: 100 and 900
  - Ave squared error: 500
- 500 = 400 + 100

Bias
More variance



- Scenario 3
- Normally distributed beliefs with mean 180 and std dev 20 (variance=400)
- Poll two: One says 200 and other 160
- Errors: 0 and -40
  - Ave error is -20
- Sq. errors: 0 and 1600
  - Ave squared error: 800
- 800 = 400 + 400

Squared error = Square of bias error + Variance

As variance increases, error increases

Bias -Variance in Regression

- $y(\mathrm{x})$: estimate of the value of t for input x
- $h(\mathrm{x})$: optimal prediction

$$h(\mathrm{x}) = E[t \mid \mathrm{x}] = \int t p(t \mid \mathrm{x}) dt$$

- If we assume loss function $L(t,y(\mathrm{x}))=\{y(\mathrm{x})-t\}^2$
- $E[L]$ can be written as

expected loss = $(\text{bias})^2$ + variance + noise

- where

$$(\text{bias})^2 = \int \{E_D[y(\mathrm{x};D)] - h(\mathrm{x})\}^2 p(\mathrm{x}) dx$$

Difference between expected value and optimal

$$\text{variance} = \int E_D\left[\{y(\mathrm{x};D)] - E_D[y(\mathrm{x};D)]\}^2\right] p(\mathrm{x}) dx$$

$$\text{noise} = \int \{h(\mathrm{x}) - t\}^2 p(\mathrm{x},t) d\mathrm{x} dt$$

Dependence of Bias-Variance on Model Complexity

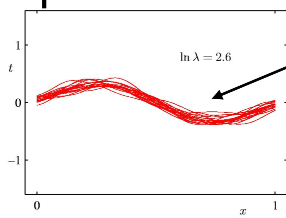
- *h(x)=sin(2px)*
- Regularization parameter *l*
- *L=100* data sets
- Each with *N=25*
- *24* Gaussian Basis functions
  - No of parameters *M=25*
- Total Error function:

$$\frac{1}{2}\sum_{n=1}^{N}\{\ t_n - \mathrm{w}^T\phi(x_n)\ \}^2 + \frac{\lambda}{2}\mathrm{w}^T\mathrm{w}$$

where f is a vector of basis functions

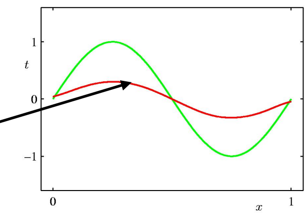
20 Fits for 25 data points each

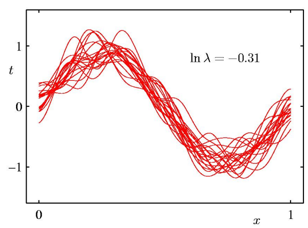
Red: Average of Fits
Green: Sinusoid from which data was generated

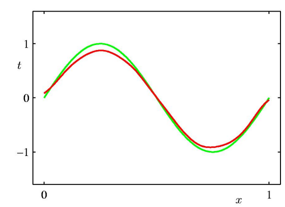


High λ

Low Variance
High bias

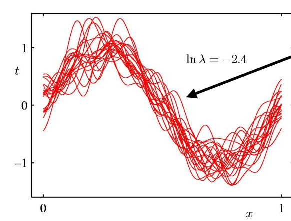






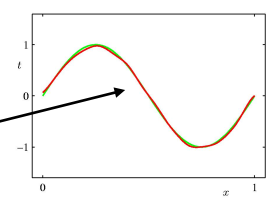
Result of averaging multiple solutions with complex model gives good fit

Weighted averaging of multiple solutions is at heart of Bayesian approach: not wrt multiple data sets but wrt posterior distribution of parameters



Low λ

High Variance
Low bias



Determining optimal λ

- Average Prediction

$$\bar{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

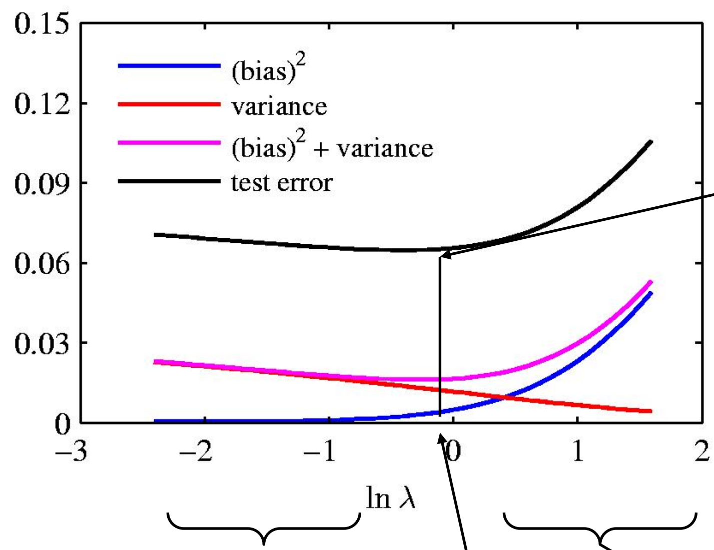
- Squared Bias

$$(bias)^2 = \frac{1}{N}\sum_{n=1}^{N}\left\{\bar{y}(x_n) - h(x_n)\right\}^2$$

- Variance

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \bar{y}(x_n)\right\}^2$$

Squared Bias and Variance vs λ



Test error minimum occurs close to minimum of (bias^2+variance)

ln λ=-0.31

Small values of λ allow model to become finely tuned to noise leading to large variance

Large values of λ pull weight parameters to zero leading to large bias

Bias-Variance vs Bayesian

- Bias-Variance decomposition provides insight into model complexity issue
- Limited practical value since it is based on ensembles of data sets
  - In practice there is only a single observed data set
  - If there are many training samples then combine them
    - which would reduce over-fitting for a given model complexity
- Bayesian approach gives useful insights into over-fitting and is also practical