Bias-Variance Trade-off in ML

Karan Nathwani

Bias-Variance Decomposition

1. Model Complexity in Linear Regression
2. Point estimate
   - Bias-Variance in Statistics
3. Bias-Variance in Regression
   - Choosing λ in maximum likelihood/least squares estimation
   - Formulation for regression
   - Example
   - Choice of optimal λ

Model Complexity in Linear Regression

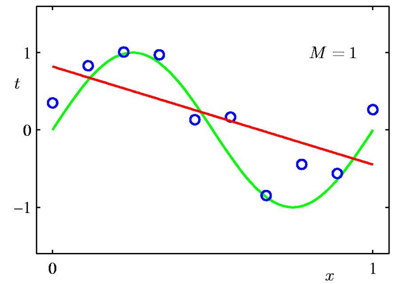
- We looked at linear regression where form of basis functions ϕ and their no. M are fixed
- Using maximum likelihood (equivalently least squares) leads to severe overfitting if complex models are trained with limited data
  - However limiting M to avoid overfitting has side effect of not capturing important trends in the data
- Regularization can control overfitting for models with many parameters
  - But seeking to minimize wrt both $\mathbf{w}$ and λ leads to unregularized solution with $\lambda=0$

Overfitting is a property of Max Likelihood

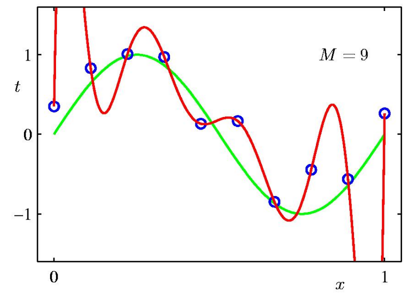
- Does not happen when we marginalize over parameters in a Bayesian setting
- Before considering Bayesian view, instructive to consider frequentist viewpoint of model complexity
- It is called *Bias-Variance* trade-off

Bias-Variance in Regression

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



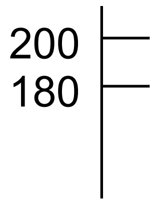
- High degree polynomial has low bias (fits well) but has high variance with different data sets



Bias-Variance in Point Estimate

True height of Chinese emperor: 200cm (6.5ft)
Poll question: "How tall is the emperor?"
Determine how wrong people are, on average

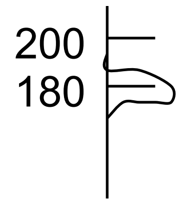
If all answer 200, ave. squared error is 0. Consider three datasets with mean=180 (or bias error = 20), but increasing variance (0, 10 and 20)



Bias
No variance

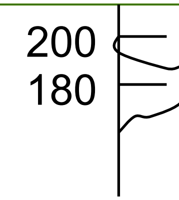
- Dataset 1
- Everyone believes it is 180 (variance=0)
- Answer is always 180
- The error is always -20
- <u>Ave squared error is 400</u>
- Average bias error is 20
- <u>400</u>=400+0

Bias
Some variance



- Dataset 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
  - <u>Ave squared error: 500</u>
- <u>500</u> = 400 + 100

Bias
More variance



- Dataset 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
  - <u>Ave squared error: 800</u>
- <u>800</u> = 400 + 400

Average Squared Error = (Bias error)^2 + Variance

As variance increases, error increases

Prediction in Linear Regression

- Need to apply *Decision Theory*: choose a specific estimate $y(\mathrm{x})$ of the value of t for a given x
- In doing so if we incur incur loss $L(t,y(\mathrm{x}))$
- Then the expected loss is

$$E[L] = \iint L(t,y(\mathrm{x}))p(\mathrm{x},t)d\mathrm{x}dt$$

- Using squared loss function $L(t,y(x))=\{y(x)-t\}^2$

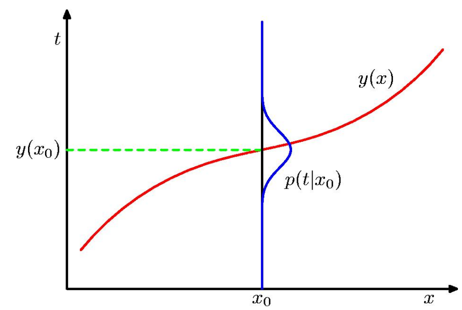
$$E[L] = \int\int (y(\mathrm{x}) - t)^2 p(\mathrm{x},t)\, d\mathrm{x}\, dt$$

- Taking derivative of E wrt $y(\mathrm{x})$, using calculus of variations,

$$\frac{\delta E[L]}{\delta y(\mathrm{x})} = 2\int (y(\mathrm{x}) - t)p(\mathrm{x},t)\, dt$$

- Setting equal to zero, solving for $y(\mathrm{x})$ and using sum and product rules

$$y(\mathrm{x}) = \frac{\int tp(\mathrm{x},t)\, dt}{p(\mathrm{x})} = \int tp(t \mid \mathrm{x})\, dt = E_t[t \mid \mathrm{x}]$$



Regression function $y(x)$, which minimizes the expected squared loss, is given by the mean of the conditional distribution $p(t|x)$

Alternative Derivation

- We can show that the optimal prediction is equal to the conditional mean in another way
- First we have

$$\{y(\mathrm{x}) - t\}^2 = \left\{y(\mathrm{x}) - E[t \mid \mathrm{x}] + E[t \mid \mathrm{x}] - t\right\}^2$$

- Substituting into the loss function, we obtain expression for the loss function as

$$E[L] = \int \left\{y(\mathrm{x}) - E[t \mid \mathrm{x}]\right\}^2 p(\mathrm{x}) d\mathrm{x} + \int \mathrm{var}(t \mid \mathrm{x}) p(\mathrm{x})\, d\mathrm{x}$$

- The function $y(\mathrm{x})$ we seek to determine enters only in the first term, which will be minimum when $y(\mathrm{x}) = E[t \mid \mathrm{x}]$

Bias -Variance in Regression

- $y(\mathrm{x})$: regression function using some method
- $h(\mathrm{x})$: optimal prediction (using squared loss)

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

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

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

- where

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

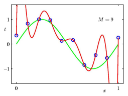
Difference between expected value and optimal

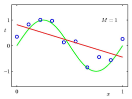
$$\text{variance} = \int E_D\left[\{y(\mathrm{x};D)] - E_D[y(\mathrm{x};D)]\}^2\right] p(\mathrm{x}) dx$$

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

Goal: Minimize Expected Loss

- We have decomposed expected loss into sum of (squared) bias, a variance and a constant noise term
- There is a trade-off between bias and variance
  - Very flexible models have low bias and high variance
  - Rigid models have high bias and low variance
  - Optimal model has has the best balance





Dependence of Bias-Variance on Model Complexity

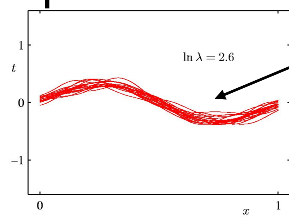
- *h(x)=sin(2πx)*
- Regularization parameter λ
- $L=100$ data sets
- Each with $N=25$
- *24* Gaussian Basis functions
  - No of parameters $M=25$
- Total Error function:

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

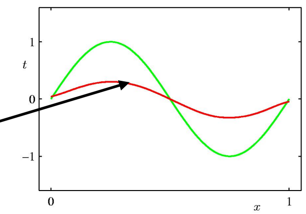
Where ϕ is a vector of basis functions

20 Fits for 25 data points each

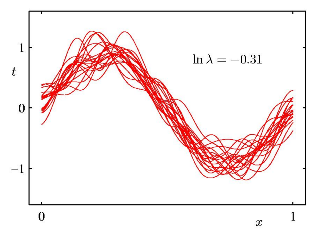
Red: Average of Fits
Green: Sinusoid from which data was generated

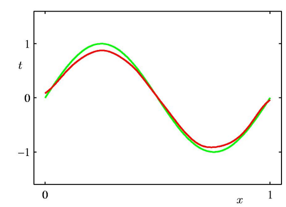


Low Variance High bias



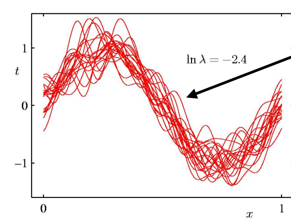
High λ



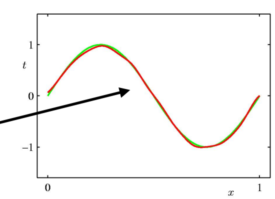


Result of averaging multiple solutions with complex model gives good fit

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



High Variance Low bias



Low λ

Determining optimal λ

- Average Prediction

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

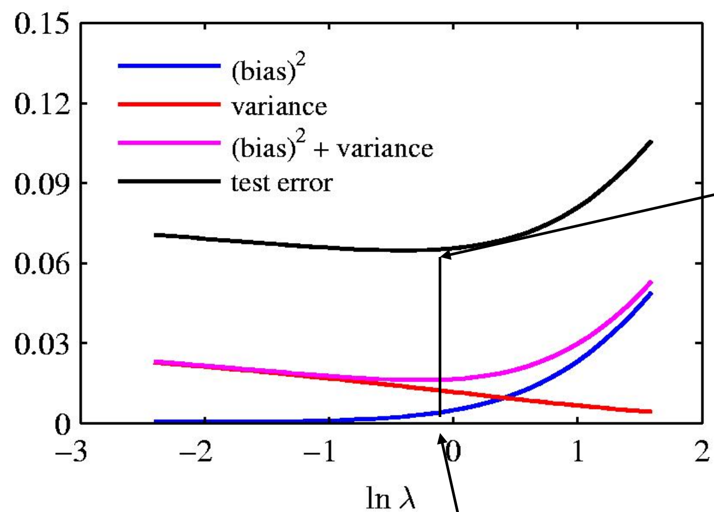
- Squared Bias

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

- 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)

$\ln\lambda$=-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