Bias-Var

Slides taken from

LEARNING FROM DATA

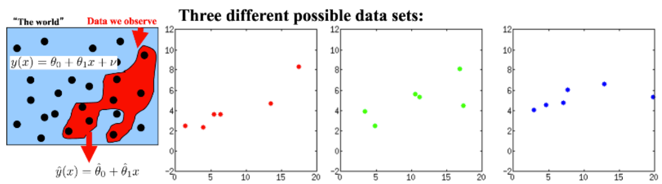
A SHORT COURSE

Yaser S. Abu-Mostafa
California Institute of Technology

Malik Magdon-Ismail
Rensselaer Polytechnic Institute

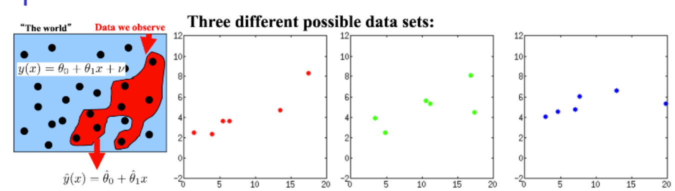
Hsuan-Tien Lin
National Taiwan University

Bias & variance

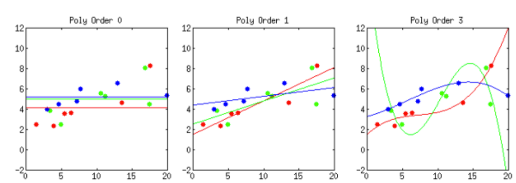


Slide by Alexander Ihler

Bias & variance



Each would give different predictors for any polynomial degree:



Slide by Alexander Ihler

Start with E_{out}

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]\end{aligned}$$

Now, let us focus on:

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

The average hypothesis

To evaluate $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$

we define the 'average' hypothesis $\bar{g}(\mathbf{x})$:

$$\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]$$

Imagine **many** data sets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$

$$\bar{g}(\mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\mathbf{x})$$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Using $\bar{g}(\mathbf{x})$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] &= \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \\ &= \mathbb{E}_{\mathcal{D}}\Big[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2 \\ &\quad + 2\ \left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\ \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)\Big] \\ &= \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right] + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\end{aligned}$$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Bias and variance

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\textbf{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\textbf{bias}(\mathbf{x})}$$

Therefore, $\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$

$$\begin{aligned}&= \mathbb{E}_{\mathbf{x}}[\textbf{bias}(\mathbf{x}) + \textbf{var}(\mathbf{x})] \\ &= \quad \textbf{bias} \quad + \quad \textbf{var}\end{aligned}$$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

The tradeoff

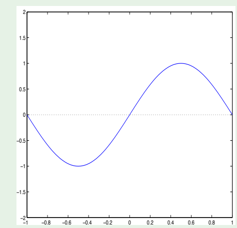
bias $= \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$ var $= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$

$\downarrow$ $\mathcal{H}\uparrow$ $\uparrow$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Example: sine target

f



$f: [-1,1] \to \mathbb{R}$ $f(x) = \sin(\pi x)$

Only two training examples! $N = 2$

Two models used for learning:

$\mathcal{H}_0$: $h(x) = b$

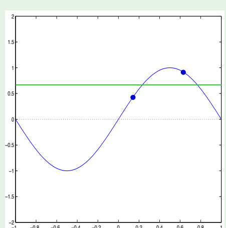
$\mathcal{H}_1$: $h(x) = ax + b$

Which is better, $\mathcal{H}_0$ or $\mathcal{H}_1$?

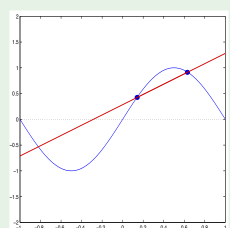
© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Learning - $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$

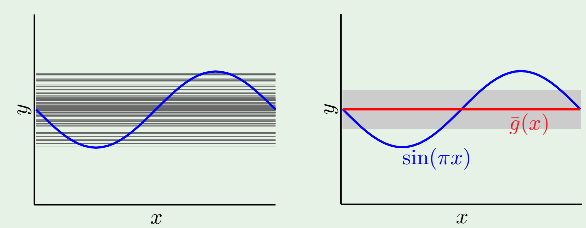


$\mathcal{H}_1$



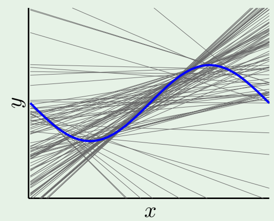
© Creator: Yaser Abu-Mostafa - LFD Lecture 8

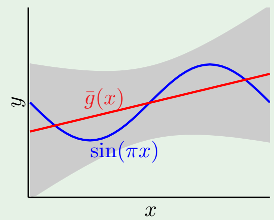
Bias and variance - $\mathcal{H}_0$



© Creator: Yaser Abu-Mostafa - LFD Lecture 8

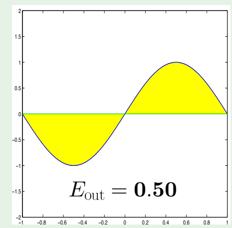
Bias and variance - $\mathcal{H}_1$



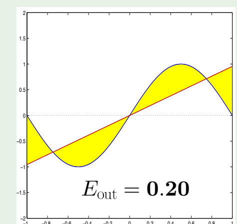


Approximation - $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$

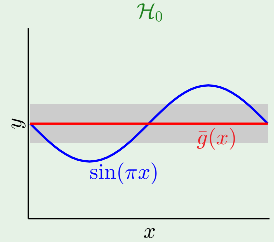


$\mathcal{H}_1$



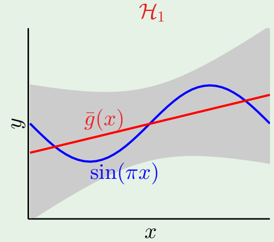
and the winner is . . .

$\mathcal{H}_0$



bias $= \mathbf{0.50}$ var $= \mathbf{0.25}$

$\mathcal{H}_1$



bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

Lesson learned

Match the **'model complexity'**

to the **data resources**, not to the **target complexity**

Outline

- Bias and Variance
- Learning Curves

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Expected E_{out} and E_{in}

Data set $\mathcal{D}$ of size N

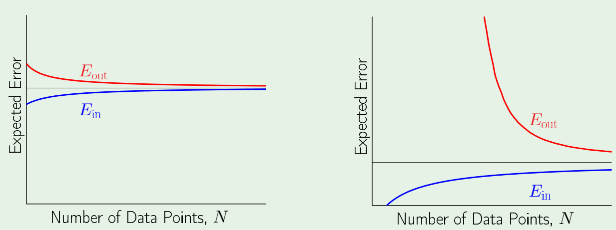
Expected out-of-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})]$

Expected in-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{(\mathcal{D})})]$

How do they vary with N?

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

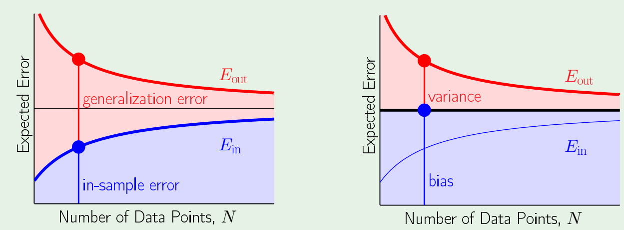
The curves



Simple Model | Complex Model

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

VC versus bias-variance



VC analysis | bias-variance

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Linear regression case

Noisy target $y = \mathbf{w}^{*\mathsf{T}}\mathbf{x}$ + noise

Data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

Linear regression solution: $\mathbf{w} = (\mathrm{X}^{\mathsf{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathsf{T}}\mathbf{y}$

In-sample error vector $= \mathrm{X}\mathbf{w} - \mathbf{y}$

'Out-of-sample' error vector $= \mathrm{X}\mathbf{w} - \mathbf{y}'$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

Linear regression case

Noisy target $y = \mathbf{w}^{*\mathsf{T}}\mathbf{x}$ + noise

Data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

Linear regression solution: $\mathbf{w} = (\mathrm{X}^{\mathsf{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathsf{T}}\mathbf{y}$

In-sample error vector $= \mathrm{X}\mathbf{w} - \mathbf{y}$

'Out-of-sample' error vector $= \mathrm{X}\mathbf{w} - \mathbf{y}'$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

21/22

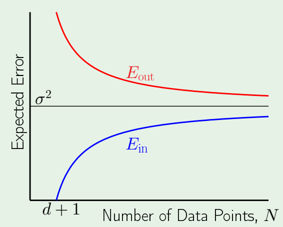
Learning curves for linear regression

Best approximation error $= \sigma^2$

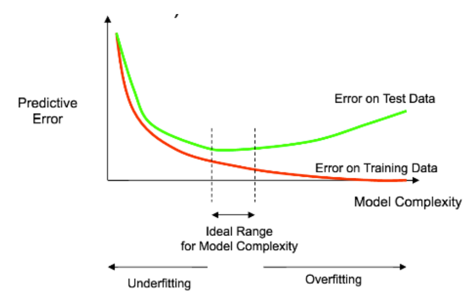
Expected in-sample error $= \sigma^2\left(1 - \frac{d+1}{N}\right)$

Expected out-of-sample error $= \sigma^2\left(1 + \frac{d+1}{N}\right)$

Expected generalization error $= 2\sigma^2\left(\frac{d+1}{N}\right)$



© Creator: Yaser Abu-Mostafa - LFD Lecture 8

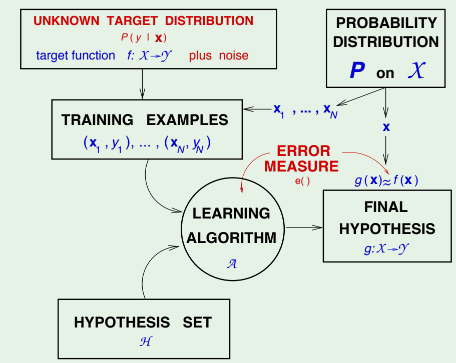


Andrew Ng

Debugging a learning algorithm:
Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors in its prediction. What should you try next?

- Get more training examples → fixes high variance
- Try smaller sets of features → fixes high variance
- Try getting additional features → fixes high bias
- Try adding polynomial features $(x_1^2, x_2^2, x_1x_2, \text{etc})$ → fixes high bias.
- Try decreasing λ → fixes high bias
- Try increasing λ → fixes high variance

The learning diagram - including noisy target



© Creator: Yaser Abu-Mostafa - LFD Lecture 4