

Bias - Variance Decomposition

- Classical & Modern Elements

Classical Theory

- Regularization

- Parameter selection

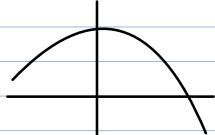
K-fold CV

Data Efficient

Successive Halving

Compute Efficient

Modern Theory (Bonus)

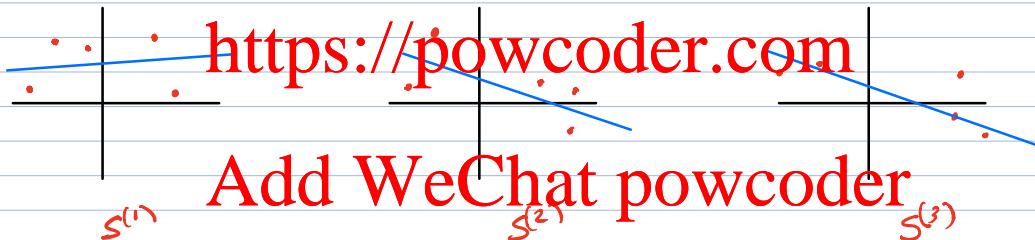


True function:

$$h_\theta(x) = \theta_2 x^2 + \theta_1 x + \theta_0$$

- don't observe h_θ
only sample from it

Assignment Project Exam Help

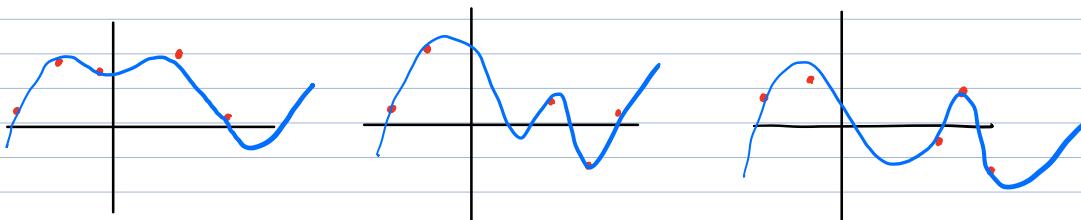


Samples

What if we fit a line to samples?

Informally: we underfit the data

[Bias]



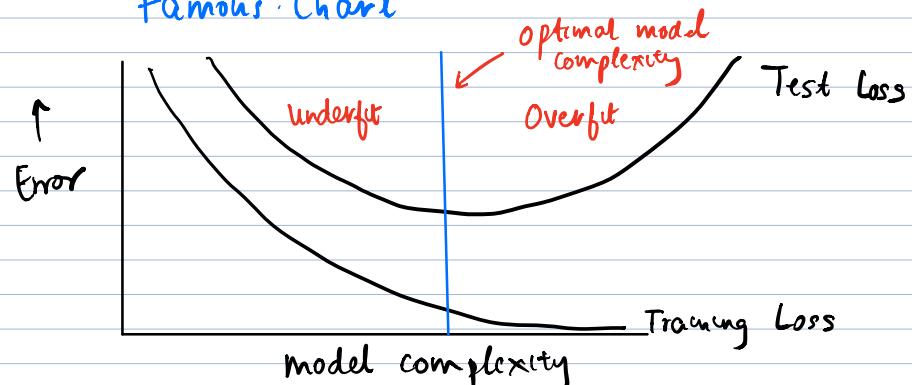
What if we fit high degree polynomial (degree 5)

Overfitting data

[Variance]

- Hope: If we picked quadratics (since h_0 quadratic)
- low bias & variance
 - not expect zero error (inherent noise)

Famous Chart



Assignment Project Exam Help

More Formally: Bias-Variance Tradeoff
 ↗ True hypothesis ($h_0(x) = \theta \cdot x$)

$$y = h_{\theta}(x) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \text{ error}$$

Add WeChat powcoder

output observed $y \in \mathbb{R}$

Features (data) $x \in \mathbb{R}^d$

$\mathbb{E}[\varepsilon] = 0$

Variance σ^2 $\mathbb{E}[\varepsilon^2] = \sigma^2$

Procedure:

1. Draw n labeled points $(x^{(1)}, y^{(1)}) \dots (x^{(n)}, y^{(n)})$
call this S

$$y^{(i)} = h_{\theta}(x^{(i)}) + \varepsilon^{(i)}$$

2. We train model on S

Call it $h_S: \mathbb{R}^d \rightarrow \mathbb{R}$

3. Pick $x \in \mathbb{R}^d$ (test point) $y = h_{\theta}(x) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$

4. Measure $(h_S(x) - y)^2$ (Risk)

We examine $\underset{s, \varepsilon}{\mathbb{E}} [(h_s(x) - y)^2]$

$$\begin{aligned} X, Y &\text{ independent} \\ \mathbb{E}[XY] \\ &= \mathbb{E}[X] \cdot \mathbb{E}[Y] \end{aligned}$$

Goal: Decompose error

$$\begin{aligned} \underset{s, \varepsilon}{\mathbb{E}} [(h_s(x) - y)^2] &= \underset{s, \varepsilon}{\mathbb{E}} [(h_s(x) - h_\theta(x) - \varepsilon)^2] \\ &= \underset{s, \varepsilon}{\mathbb{E}} [\varepsilon^2] - 2 \underset{s, \varepsilon}{\mathbb{E}} [\varepsilon (h_s(x) - h_\theta(x))] + \underset{s}{\mathbb{E}} [(h_s(x) - h_\theta(x))^2] \\ &\quad \uparrow \text{indep of } S, \mathbb{E}[\varepsilon] = 0 \\ &= \sigma^2 + 0 + \underset{s}{\mathbb{E}} [(h_s(x) - h_\theta(x))^2] \\ &\quad \text{unavoidable error} \end{aligned}$$

Define: $h_{avg}(x) \triangleq \mathbb{E}_{s, \varepsilon}[h_s(x)]$

Assignment Project Exam Help

randomly select S , train to fit h_s , evaluate $h_s(x)$

<https://powcoder.com>

$$\underset{s}{\mathbb{E}} [(h_s(x) - h_{avg}(x))^2] = \underset{s}{\mathbb{E}} [(h_s(x) - h_{avg}(x) + h_{avg}(x) - h_\theta(x))^2]$$

$$= \underset{s}{\mathbb{E}} [(h_s(x) - h_{avg}(x))^2] + \underset{s}{\mathbb{E}} [(h_{avg}(x) - h_\theta(x))^2]$$

$$+ 2 \underset{s}{\mathbb{E}} [(h_s(x) - h_{avg}(x))(h_{avg}(x) - h_\theta(x))]$$

$$\text{Since } \underset{s}{\mathbb{E}}[h_s(x)] = h_{avg}(x)$$

$$\Rightarrow \underset{s}{\mathbb{E}}[h_s(x) - h_{avg}(x)] = 0$$

$$= \underset{s}{\mathbb{E}} [(h_s(x) - h_{avg}(x))^2] + [h_{avg}(x) - h_\theta(x)]^2$$

Variance of branching procedure

$\text{VAR}_s(h)$

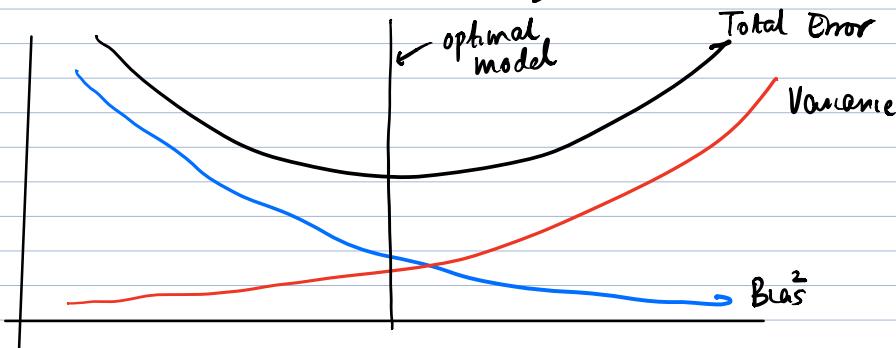
Bias² does not depend on S

\Rightarrow error introduced by
Model family

$$\text{Summary: } \underset{s, \epsilon}{\mathbb{E}} \left[(h_s(x) - h_0(x))^2 \right]$$

= Unavoidable error + Bias² + Variance

$$= \sigma^2 + (h_{avg}(x) - h_0(x))^2 + \underset{s}{\mathbb{E}} \left[(h_s(x) - h_{avg}(x))^2 \right]$$



Assignment Project Exam Help

- Reduce Variance to obtain more robust model
(to training set variation)

<https://powcoder.com>

→ Explicit (Change the model) Penalty terms

→ Implicit (by the algorithm)

Classical setting:

$$\underset{\theta \in \mathbb{R}^d}{\operatorname{Argmin}} \frac{1}{2} \sum_{i=1}^n [x^{(i)} \cdot \theta - y^{(i)}]^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

↑ regularization parameter $\in \mathbb{R}_+$

↑ pick a less complex model

$\lambda = 0$ ordinary least squares

$\lambda = 100^{100}$ $\theta \sim 0$ probably good

tradeoff λ : hyperparameter

Solution: fix $\lambda > 0$

$$L(\theta) = \frac{1}{2} (x\theta - y)^T (x\theta - y) + \frac{\lambda}{2} \theta^2$$

$$\nabla_{\theta} L(\theta) = x^T (x\theta - y) + \lambda \theta$$

$$(x^T x + \lambda I)\theta = x^T y \quad \text{Normal equation}$$

Undetermined (modern ML)

$$\text{Rank}(x^T x) < d \quad (\lambda = 0)$$

there is no unique soln

$$\lambda > 0 \quad x^T x \theta_0 = x^T y$$

$$\forall v \quad (x^T x)v = 0$$

$\theta_0 + v$ is a solution too!

$\lambda > 0 \quad x^T x \text{ is PSD} \quad \sigma_{\max}^2 > \dots > \sigma_n^2 > 0$

$$\lambda > 0 \quad x^T x + \lambda I \quad \sigma_1^2 + \lambda > \dots > \sigma_n^2 + \lambda \geq 0$$

Add WeChat powcoder

↗ regularized solution
(ridge regression)

Reduces variance!

$$\text{Var}_s(h) : \mathbb{E} \left[[h_s(x) - h_0(x)]^2 \right]$$

increase λ , spectrum getting flatter
⇒ decreases variance!

Bonus: Implicitly regularize as well

thought expt: run gradient descent

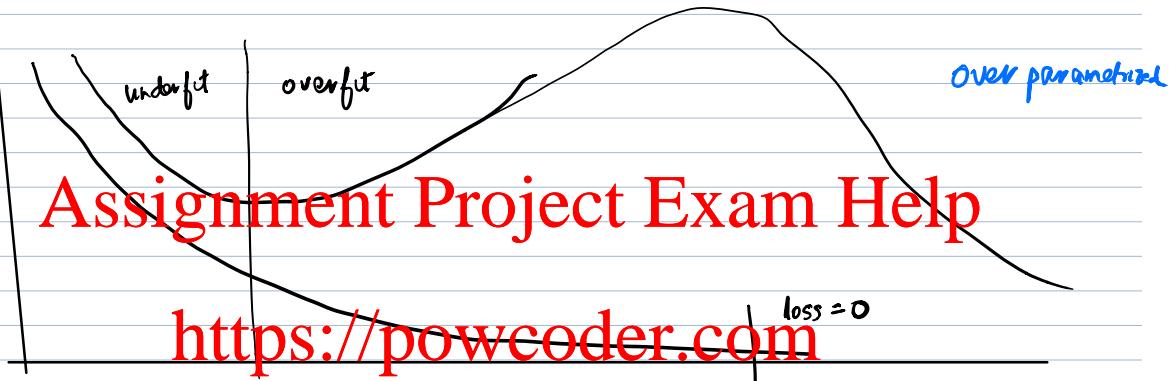
$$\theta_{GD} : P_{\text{NULL}(x^T x)}(\theta_{GD}) + P_{\text{SPAN}(x)}(\theta_{GD})$$

Claim: $P_{\text{NULL}}(\theta_{GD}) = P_{\text{NULL}}(\theta^{(0)})$

$$\theta^{(t+1)} = \theta^{(t)} - \alpha X^T (X\theta^{(t)} - y) \quad \leftarrow \text{always in } \text{span}(X)$$

$\theta^{(0)} = 0 \Rightarrow \text{min norm soln just by using gradient descent}$

Extra: Belkin, Hsu, Ma, Mandel 2018 "Double Descent"



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder