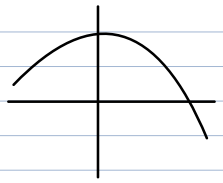
Bias-Variance Decomposition

- Classical & Modern Elements

Classical Theory

- Regularization
- Parameter selection
  - K-fold CV — Data Efficient
  - Successive Halving — Compute Efficient

Modern Theory (Bonus)



True function:

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

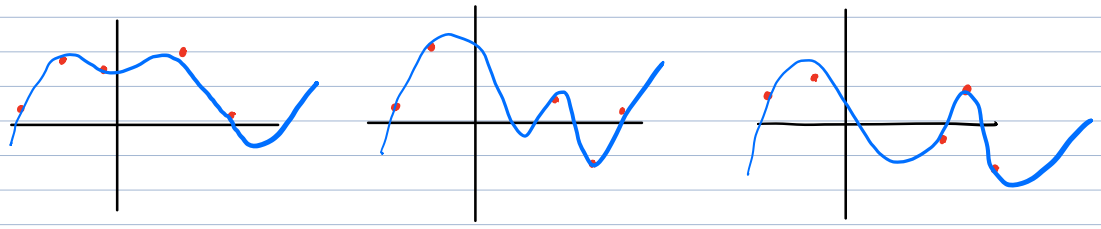
- don't observe h_θ
- only samples from it

$S^{(1)}$ $S^{(2)}$ $S^{(3)}$

Samples

What if we fit a line to samples?

Informally: we underfit the data [Bias]

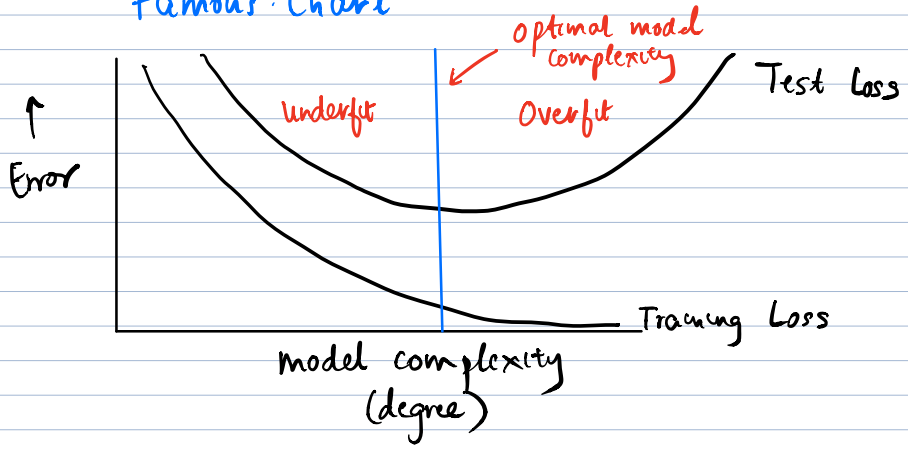


What if we fit high degree polynomial (degree 5)

Overfitting data [Variance]

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

Famous Chart



More Formally: Bias-Variance Tradeoff

$y = h_\theta(x) + \varepsilon$ — h_θ: True hypothesis ($h_\theta(x) = \theta \cdot x$); x: Features (data) $\in \mathbb{R}^d$

$\varepsilon \sim N(0, \sigma^2)$ error

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

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

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

$\theta, x \in \mathbb{R}^d$

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

We examine $\mathbb{E}_{S,\varepsilon}\left[(h_S(x)-y)^2\right]$

X, Y independent
$\mathbb{E}[XY] = \mathbb{E}[X]\cdot\mathbb{E}[Y]$

Goal: Decompose error

$$\mathbb{E}_{S,\varepsilon}\left[(h_S(x)-y)^2\right] = \mathbb{E}_{S,\varepsilon}\left[(h_S(x)-h_\theta(x)-\varepsilon)^2\right]$$

$$= \mathbb{E}_{S,\varepsilon}\left[\varepsilon^2\right] - 2\,\mathbb{E}\left[\varepsilon\,(h_S(x)-h_\theta(x))\right] + \mathbb{E}_S\left[(h_S(x)-h_\theta(x))^2\right]$$

ε indpt of S, $\mathbb{E}[\varepsilon]=0$

$$= \sigma^2 + 0 + \mathbb{E}_S\left[(h_S(x)-h_\theta(x))^2\right]$$

σ^2: unavoidable error

Define: $h_{avg}(x) \triangleq \mathbb{E}_S[h_S(x)]$

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

h_{avg}: average prediction (over S)

$$\mathbb{E}_S\left[(h_S(x)-h_\theta(x))^2\right] = \mathbb{E}_S\left[(h_S(x)-h_{avg}(x)+h_{avg}(x)-h_\theta(x))^2\right]$$

$$= \mathbb{E}_S\left[(h_S(x)-h_{avg}(x))^2\right] + \mathbb{E}_S\left[(h_{avg}(x)-h_\theta(x))^2\right]$$

$$+ 2\,\mathbb{E}_S\left[(h_S(x)-h_{avg}(x))(h_{avg}(x)-h_\theta(x))\right]$$

Since $\mathbb{E}_S[h_S(x)] = h_{avg}(x)$

$\Rightarrow \mathbb{E}_S[h_S(x)-h_{avg}(x)] = 0$

$$= \mathbb{E}_S\left[(h_S(x)-h_{avg}(x))^2\right] + \left[h_{avg}(x)-h_\theta(x)\right]^2$$

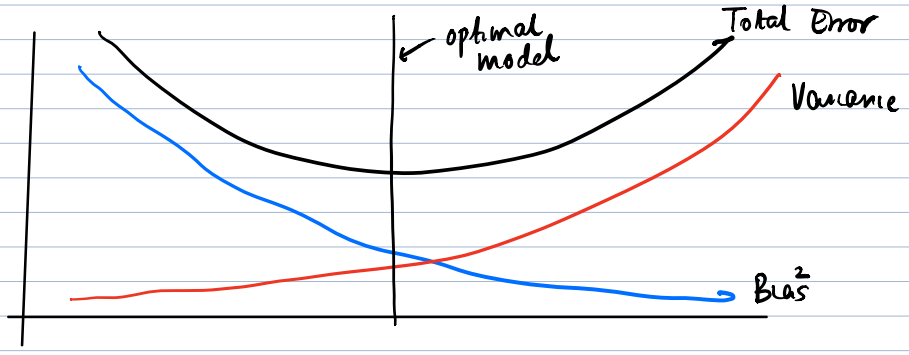
Variance of training procedure: $VAR_S(h)$

Bias^2 does not depend on S
$\Rightarrow$ error introduced by model family

Summary: $\mathbb{E}_{S,\varepsilon}\left[(h_S(x) - h_0(x))^2\right]$

$=$ unavoidable error σ^2 + Bias^2 + Variance

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



Regularization

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

→ Explicit (Change the model) Penalty terms

→ Implicit (by the algorithm)

Classical setting:

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

(λ: regularization parameter $\in \mathbb{R}_+$; $\|\theta\|_2^2$: pick a less complex model)

$\lambda = 0$ ordinary least squares

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

tradeoff λ: hyperparameter

Solution: fix $\lambda > 0$

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

$$\nabla_\theta L(\theta) = X^T(X\theta - y) + \lambda\theta$$

$$(X^TX + \lambda I)\theta = X^Ty \qquad \text{Normal equations}$$

Underdetermined (modern ML)

$\text{Rank}(X^TX) < d \qquad (\lambda = 0)$

there is no unique soln

$\lambda = 0 \qquad X^TX\theta_0 = X^Ty$

$\exists\, v \qquad (X^TX)v = 0$

$\theta_0 + v$ is a solution too!

$\lambda > 0 \qquad X^TX$ is PSD $\qquad \sigma_{max}^2 > \dots > \sigma_n^2 \geq 0$

$\sigma_n^2 = 0$

$X^TX + \lambda I \qquad \sigma_1^2 + \lambda \geq \dots \geq \sigma_n^2 + \lambda \geq 0$

$$\theta_\lambda = (X^TX + \lambda I)^{-1}X^Ty$$

↑ regularized solution (ridge regression)

Reduces variance!

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

increase λ, spectrum getting flatter

$\Rightarrow$ decreases variance!

Bonus: Implicitly regularize as well

thought expt: run gradient descent

$$\theta_{GD}: P_{\text{NULL}(X^TX)}(\theta_{GD}) + P_{\text{SPAN}(X)}(\theta_{GD})$$

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

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

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

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