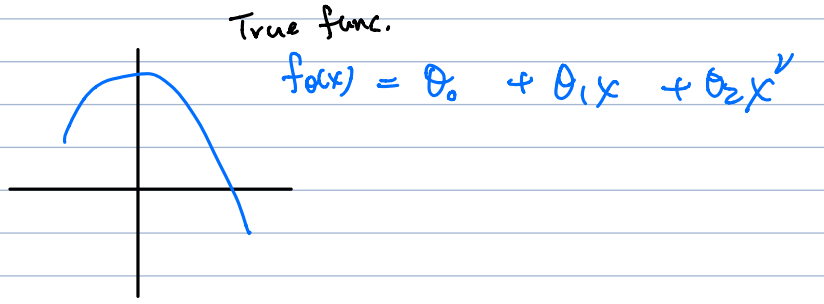
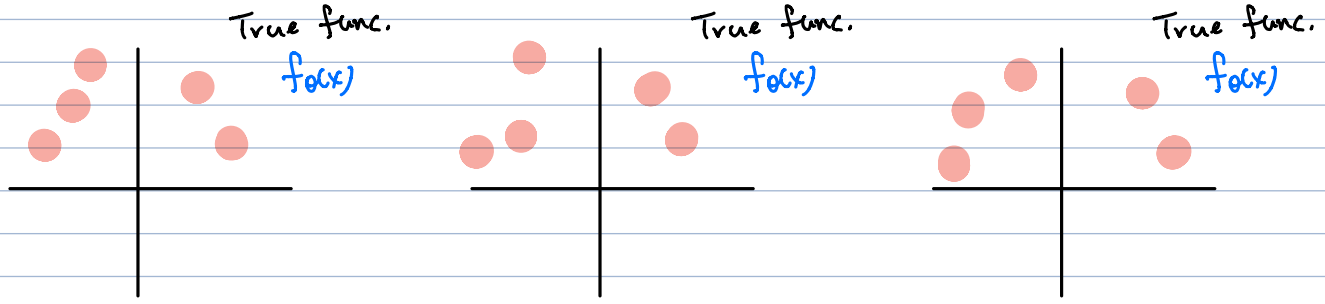
Overview

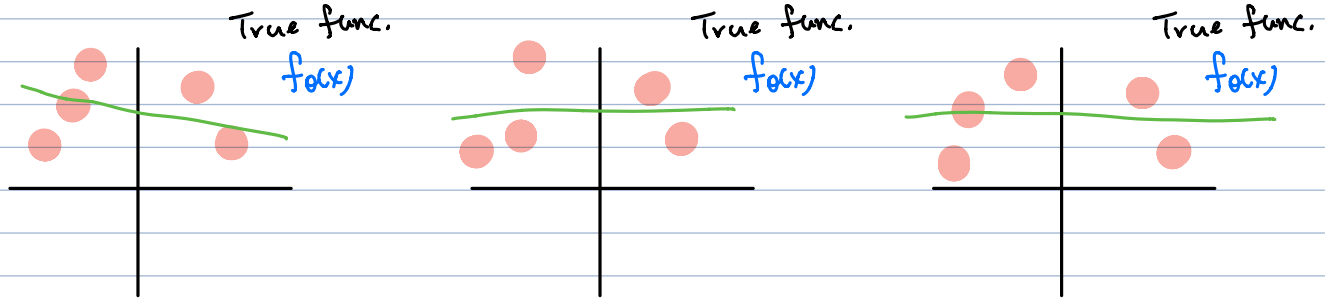
- Bias and Variance



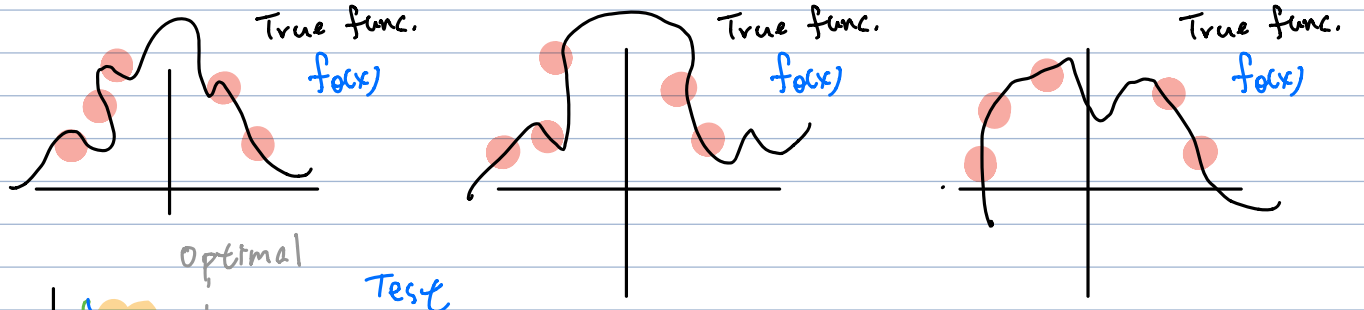
Samples

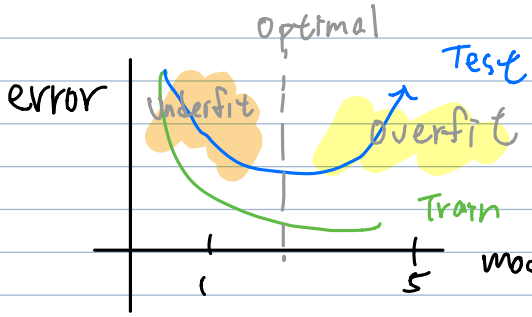


Fit a linear line Bias, Underfit



Fit a degree 5 polynomial High Variance, overfit





Classical Bias-Variance

Bias - Variance

$y = x\theta + \epsilon$ ($x \in \mathbb{R}^d$ features, θ model $\in \mathbb{R}^d$, $y \in \mathbb{R}$) $\quad \epsilon \sim N(0, \sigma^2)$

Procedure

1. Draw n points $(x^{(1)}, y^{(1)}) \dots (x^{(n)}, y^{(n)})$ as set S

$$y^{(i)} = x^{(i)}\theta_x + \epsilon^{(i)} \qquad \epsilon^{(i)} \sim N(0, \sigma^2)$$

2. Train a linear model on S : $h_S : \mathbb{R}^d \to \mathbb{R}$
3. Draw a test sample (x, y)

$$y = h_{\theta_x}(x) + \epsilon$$

4. Measure $\mathop{\mathbb{E}}_{\epsilon, S}\left[(h_S(x) - y)^2\right]$

Goal: Decomposition of Error

$$\mathop{\mathbb{E}}_{\epsilon, S}\left[\left(h_S(x) - (h_{\theta_x}(x) + \epsilon)\right)^2\right]$$

$\mathbb{E}[\epsilon] = 0$.

$$= \underbrace{\mathbb{E}[\epsilon^2]}_{\sigma^2} + \underbrace{\mathbb{E}\left[(h_S(x) - h_\theta(x))^2\right]}_{\text{Depend on } S \text{ (training set)}} + \mathbb{E}\left[\epsilon\,(h_S(x) - h_\theta(x))\right]$$

ϵ independent; $h_S(x) - h_\theta(x)$: 0

σ^2: Unavoidable error

$$h_{Avg}(x) \triangleq \mathop{\mathbb{E}}_{S}[h_S(x)]$$

"long run average training err on S" (select S many times)

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

$$= \underbrace{\mathbb{E}\left[(h_S(x) - h_{avg}(x))^2\right]}_{\text{Variance}} + \underbrace{\mathop{\mathbb{E}}_{S}\left[(h_{Avg}(x) - h_\theta(x))^2\right]}_{\text{Bias}^2} + \mathop{\mathbb{E}}_{S}\left[(h_S(x) - h_\theta(x))(h_{Avg}(x) - h_\theta(x))\right]$$

$(h_S(x) - h_\theta(x))$: 0; $(h_{Avg}(x) - h_\theta(x))$: 0; whole term : 0

Variance: $VAR(S)$

Bias^2: Does not depend on S

Examples	Bias	Variance
linear	high	low
degree 5	0	high

Reduce Variance that depends on the training set

Regularization

Most Classical case (linear)

Regularization Parameter

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} (x^{(i)}\theta - y^{(i)})^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

Penalty of complex

Solution $x \in \mathbb{R}^{n \times d}$

$$x^T x\theta - x^T y + \lambda\theta = 0$$

$$\Rightarrow \quad \theta = (x^T x + \lambda I)^{-1} x^T y \qquad \text{(Normal eq)}$$

What if $x^T x$ is not full rank? $n < d \Rightarrow \text{Rank}(x^T x) < d$

if we set $\lambda = 0$

$$x^T x\theta = x^T y \qquad v \in \text{Null}(x)$$

$$x^T x(\theta + v) = x^T x\theta = x^T y \qquad \text{(Not Unique)}$$

if $\lambda > 0$. $x^T x + \lambda I$ is full rank

eigenvalues $(x^T x) = \sigma_1^2 \geq \sigma_2^2 \geq \dots \geq 0$

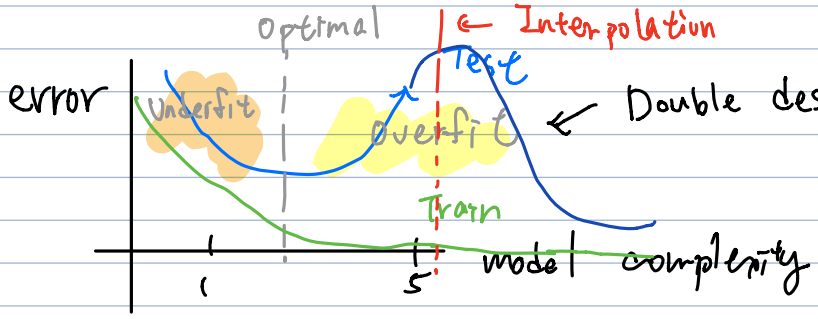
$$(x^T x + \lambda I) = \sigma_1^2 + \lambda, \sigma_2^2 + \lambda, \dots, \sigma_n^2 + \lambda \geq 0$$

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

$$\text{Var}(S) = \mathop{\mathbb{E}}_{S}\left[(\theta_\lambda x - \mathop{\mathbb{E}}_{S}[\theta_\lambda x])^2 \right]$$

$$\approx \mathbb{E}\left[(\theta_\lambda - \mathbb{E}[\theta_\lambda])^2 \right]$$

As λ increases, variance decreases



Picking hyperparameters

Three set of data

daily Train set → fit parameters

Dev set → tune parameters

Test set → blind

for degree $d \in \{1, \dots, m\}$

- Train model w/ parameters d on the train set
- Score on the dev set

⇒ Pick best score

How to pick λ? $\lambda \in \{0, 10^{-4}, 10^{-3}, \dots 10^{-1}\}$

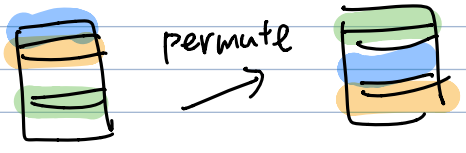
Improvements

Data Efficiency: k-fold cross validation

Compute Efficiency: Many hyperparameters

"modern ML"

Data k-fold Cross validation

1. 

2.

3. Combine Score (Average)

Train	Score
$S_1 S_2 S_3$	S_4
$S_1 S_3 S_4$	S_2
⋮	⋮

Computational

Motivation : Dropout, Regularization

More Advanced (Amenssin (15') * hyperband

Run all (5, 6, 7) ⇒ just for a few steps

score all of them

Prck $N/2$

then they run $N2^{ck}$ number of models