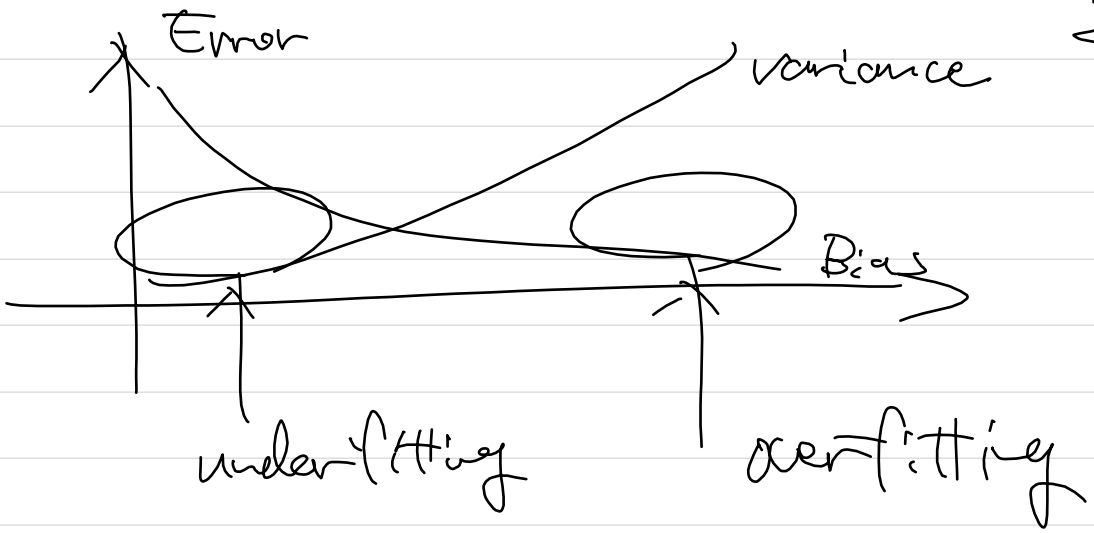
21H - Q1

April 20th (

2(H-Q1

[2]



[3] Diff of Hyperparam & Param

Hyperparam → use for model selection & determine training behaviour by using val set

Param → updated by training set

[4] How to choose hyperparam :

At first, we split traing data into train and val set

then Train model on train set and evaluate on val set

then choose the best hyperparam by

2|HQ1 3

choosing lowest error on val set.
↑ the model trained with

or using AIC and so on
↑ for regular model,

set A set B

| 800 | 1200 |

$x_1(1) > x_2(1) > \cdots x_j(1) \cdots > x_{1000}(1)$

$\forall j$ $x_j(1)$ is the first feature of j-th train sample

if first feature of each sample have strong correlation to the label

in this case, even for $x_j \sim D$ i.i.d

training on set A leads to overfit to first feature of set A.

So, it is better to do shuffle randomly.

6 $D: X \times Y \quad l: Y \times Y \to \mathbb{R}_+$

$$x, y \overset{iid}{\sim} D$$

(a) Define true risk $R(h)$

$h: X \to Y$

$$R(h) = \mathbb{E}_{x,y \sim D}\left[l\left(h(x,y), y\right) \right]$$

(b) empirical Risk $\hat{R}(h, S)$

wrt $S = \{(x_1, y_1) \cdots (x_n, y_n)\} \sim D$

$$\hat{R}(h, S) = \mathbb{E}_{x,y \sim S}\left[l(h(x,y), y) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} l(h(x_i, y_i), y_i)$$

[6] (c) Show $\hat{R}(h,S)$ is unbiased estimator of the true risk $R(h)$

$$\mathbb{E}[\hat{R}(h,S)] = R(h)$$

Then $\hat{R}$ is unbiased estimator of R

$D^n : (x_i, y_i)$ $i = 1 \dots n$
~ joint distribution

$$\mathbb{E}_{D^n}[\hat{R}(h)] = \mathbb{E}_{D^n}\left[\frac{1}{n}\sum_{i=1}^{n}(l(h(x_i), y_i))\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_D[l(h(x_i), y_i)]}_{R(h)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} R(h)$$

$$= R(h)$$

(d) Suppose $\sqrt{n}$ D
identically
but not independently

$\hat{R}(h, S)$ is still unbiased estimator of $R(h)$?

$\hat{R}(h, S)$ is still unbiased estimator.

6 (c) doesn't use property of independence;

So, b(c) is proof.

if it is independent

$$\lim_{n \to \infty} Pr_{D^n}\left(|\hat{R}(h) - R(h)| > \varepsilon\right) 0 \quad \forall \varepsilon > 0$$

LLN (Law of Large Numbers

$\hat{R}(h)$ convergence in probability $\to R(h)$