

21H-Q1

April 20th (

III

A hand-drawn graph illustrating the bias-variance tradeoff. The vertical axis is labeled "Error" and the horizontal axis is labeled "Bias". A curve labeled "variance" starts at the origin and increases as bias decreases. A curve labeled "Error" starts high on the y-axis and decreases as bias decreases. The two curves intersect at a point marked with a vertical line and labeled "underfitting" on the x-axis. Another point further to the right is marked with a vertical line and labeled "overfitting" on the x-axis. The area under the "Error" curve is shaded with diagonal lines.

I determine Training behaviour by using val set

At first, we split training 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,

5

| set A | set B |
|-------|-------|
| 100   | 100   |

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

$x_j(1)$  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]

$$D: \mathcal{X} \times \mathcal{Y} \quad l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$$

$$x, y \sim D$$

iid

(a) Define true risk  $R(h)$   
 $h: \mathcal{X} \rightarrow \mathcal{Y}$

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

(b) empirical risk  $\hat{R}(h, S)$   
 w.r.t  $N = \{(x_1, y_1), \dots, (x_n, y_n)\} \sim D$

$$\begin{aligned} \hat{R}(h, S) &= \mathbb{E}_{x, y \sim S} [l(h(x, y), y)] \\ &= \frac{1}{n} \sum_{i=1}^n l(h(x_i, y_i), y_i) \end{aligned}$$

[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  $R$

$\mathcal{D}^n : (x_i, y_i) \quad i=1, \dots, n$   
 $\hookrightarrow$  joint distribution

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^n}[\hat{R}(h)] &= \mathbb{E}_{\mathcal{D}^n} \left[ \frac{1}{n} \sum_{i=1}^n (\ell(h(x_i), y_i)) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}_{\mathcal{D}} [\ell(h(x_i), y_i)]}_{R(h)} \\ &= \frac{1}{n} \sum_{i=1}^n R(h) \\ &= R(h) \end{aligned}$$

2HL Q1

6

(d) Suppose  $n \sim D$   
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, [6] (c) is proof.

if it is independent

$$\lim_{n \rightarrow \infty} P_{D^n} (|\hat{R}(h) - R(h)| > \epsilon) = 0 \quad \forall \epsilon > 0$$

LLN (Law of Large Numbers)

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