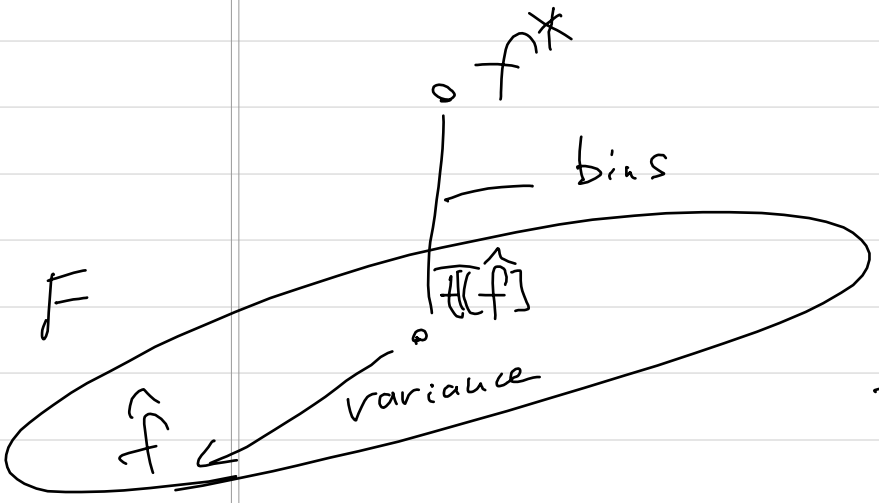
(a) Underfitting : condition that statistical model cannot capture feature and rules

Overfitting : condition that statistical model fits to not only feature but also noise

bias : gap between expectation of estimate and true model



$R(f)$ is risk function

$$f^* = \underset{f}{\operatorname{argmin}}\, R(f)$$

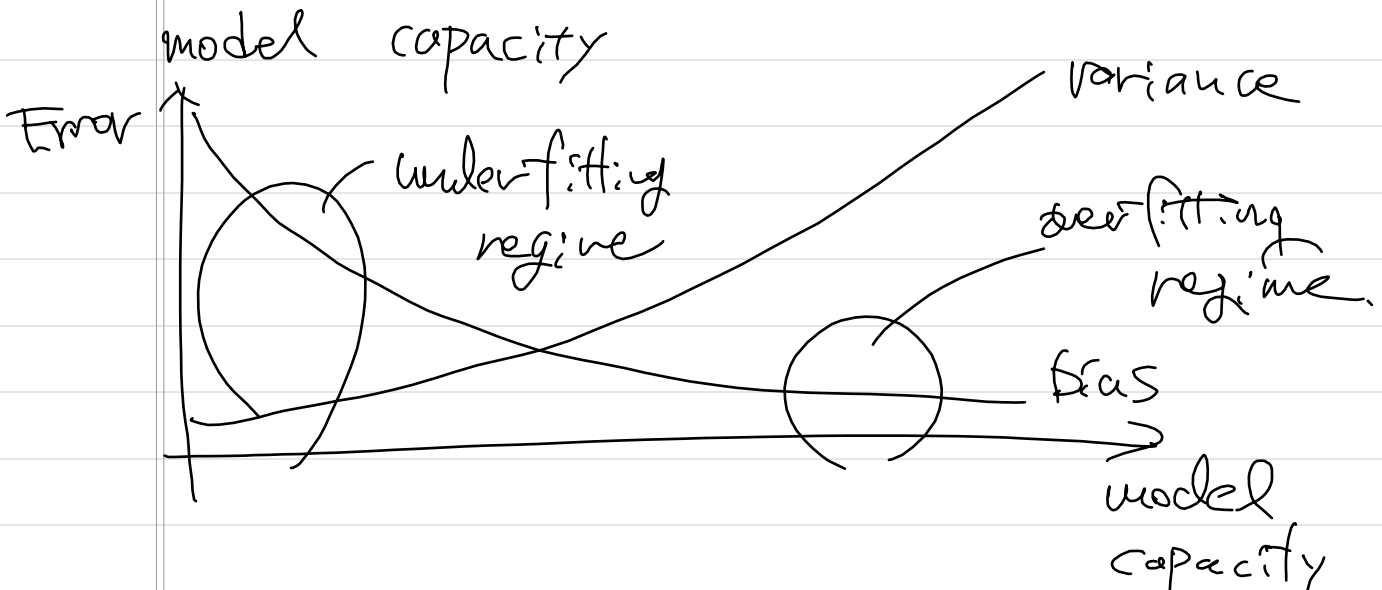
$$\hat{f} = \underset{f \in F}{\operatorname{argmin}}\, R(f)$$

$$\text{Bias}[\hat{f}] = \mathbb{E}[\hat{f}] - f^*$$

$$\text{Variance}[\hat{f}] = \mathbb{E}[\hat{f}^2] - (\mathbb{E}[\hat{f}])^2$$

Bias: an error from erroneous assumption in the learning algorithm, High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting)

Variance: an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting)



that control the bias-and-variance trade-off.

2IA_Q1

(c) #train sample ↗ variance ↘ ?

2 (a) Diff of param hyperparam

↑ updated by traing data set

↑ dose by evaluation on validation data as model selection

which determin training behaviour and efficient capacity

↓ param

↓ hyperparam

(b) (1) Linear Regression

$\theta = \{w, b\}$

$obj: l(f(x), y) = \frac{1}{2}\|f(x) - y\|^2 + \lambda \|w^2\|$

$f(x) = w^T x + b$

$w \in \mathbb{R}^{d \times d}$, $y \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, $b \in \mathbb{R}^d$.

L2-

λ : regularization

↓ loss func

$\theta_{t+1} = \theta_t - \eta \nabla_\theta l(\hat{y}, y)$

η = stepsize

(2) MLP

$\hat{y} = f(x; \theta)$

$\theta \in \mathbb{R}^d$ model param

$\theta \sim Beta(\alpha, \beta)$ ← hyperparam
$X_i \sim Bern(\theta)$ ← param

$y \sim N(x\beta, \sigma I)$

$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$

β, θ ← param
λ ← hyperparam

3

1. Predict likelihood
   input quiz score
   output likelihood.

   ① regression (logistic)
   ② MLP w/ sigmoid

2. Predict height

   ① linear regression
   regression tree

3. Unsupervised classification (text)

non-hierarchical
k-means (clustering)
Hierarchical clustering (Ward method)

4 Cross Entropy and MLE (relationship)

if we want to maximize log likelihood

$$\hat{\theta} = \arg\max_{\theta} p(y \mid x; \theta)$$

$$\Leftrightarrow \quad \hat{\theta} = \arg\max_{\theta} \log p(y \mid x; \theta)$$

$$= \arg\max_{\theta} L(\theta)$$

$$= \arg\max_{\theta} \sum_{k=1}^{K} y_k \log p_k$$

$$= \arg\min_{\theta} -\sum_{k=1}^{K} y_k \log p_k$$

↑ it is equivalent CLE (y_k, p_k)

$$= \prod_{k=1}^{K} p(y_i \mid x_i; \theta)^{y_i}$$

$$= \prod_{k=1}^{K}$$

- we focus
- one train sample

2(A-Q1

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

$S \subset \mathbb{R}^d \times \{-1, 1\}$

(a) | A | B | ← data set

step1 use A as train data

then evaluate $\hat{\theta}_A$ on data set B

step2 use B as vise-versa

estimated error $= \frac{1}{2}(l(B, \hat{\theta}_A) + l(A, \hat{\theta}_B))$

(b) CV error of learned classifier

an based estimate of its true error?