Lecture 22

..CV)

Regularization

K-fold CV

1. Partition $\mathcal{D} = \mathcal{D}_1 \cup \dots \cup \mathcal{D}_K$ disjoint, $|\mathcal{D}_1| \approx \frac{n}{K}$

2. for $k = 1:K$
   - train $f_{1:m}^{(k)}$ on $\mathcal{D}_{-k}$
   - compute $L(f_m^{(k)}; \mathcal{D}_k) = L_m^{(k)}$

$\mathcal{D}_{-k} = \mathcal{D} \setminus \mathcal{D}_k$

$\mathcal{D}_{-k}$: train — $\mathcal{D}_k$: validate

$$\bar{L}_m = \frac{1}{K} \sum_k L_m^{(k)}$$

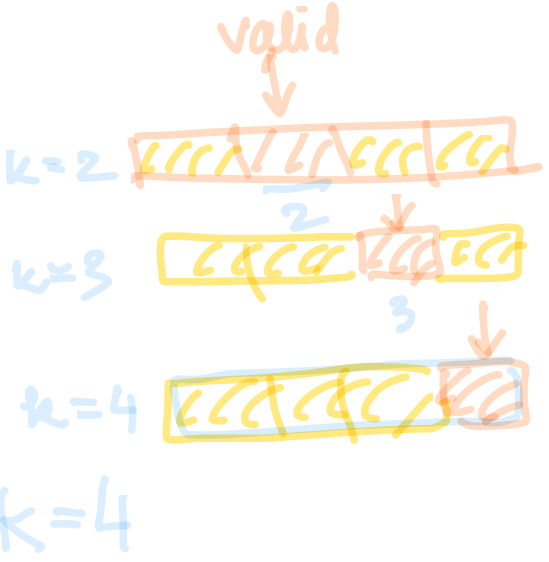
avg reduces variance

$$m^* = \operatorname*{argmin}_m \bar{L}_m$$

→ final predictor = model m^* trained on all $\mathcal{D}$

"Training set" $|\mathcal{D}_{-k}| = \frac{K-1}{K} n$

validation set $|\mathcal{D}_k| = \frac{1}{K} n$



$K \nearrow$ training time $\sim |\mathcal{D}_{-k}| \cdot K = \frac{K-1}{K} K \cdot n \propto K n$

needed only when $n \searrow$

$K = n$ Leave-One-Out CV for $n < 100$

large n

small n : K-fold

Regularization – a bias for simple models

- weight decay = L2 Regularization
- Lasso = L1 Regularization
- ~~data augmentation~~
- early stopping
- dropout
- AE : denoising auto encodus, sparse autoencoders
- Bagging DT → Random Forest (RF) = $DT_1 : M$ trained independently $M \simeq 100$

TRAIN perturb — data – e.g. subsample of $\mathcal{D}$ for $n'=n$ with replacement
algorithm – randomize splits

PREDICT $f(x) = \frac{1}{M}\sum_{k=1}^{m} DT_k(x)$ avg. over forest : Var ↘

single DT_k → low bias, high var

Model $y = \beta^T x + \varepsilon$ $\quad x, \beta \in \mathbb{R}^d$

TRAIN $\min_\beta L_{LS} + R(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y^i - \beta^T x^i)^2 + \lambda \cdot$ { $\|\beta\|_2^2$ L2 ← closed form ; $\|\beta\|_1$ L1 ← CONVEX, sparse ; 0 no regul ariz. ← closed form ; $\#\{\beta_j \neq 0, j=1:d\} = S$ ← sparse reg ✓ NOT convex (L0) }

regularization term

PREDICTION Want L_{LS} low

TRain $\hat\beta = (X^TX + 2\lambda I)^{-1}X^Ty$ not sparse

$\hat\beta = (X^TX)^{-1}X^Ty$, not closed form

$\lambda \uparrow$ bias ↑, Var ↘
$\lambda \downarrow$ bias ↘, Var ↗
$\lambda \geq 0$
λ chosen by CV

sparse Regression

Assumption : $\beta_{1:s} \neq 0$, $\beta_{s+1:d} = 0 \Leftrightarrow \beta$ sparse
$s \ll d$

$(x_{s+1:d}$ = irrelevant)

$\Downarrow$

$y = \beta_1 x_1 + \dots + \beta_s x_s + \text{noise}$

relevant variables

sparse Regression

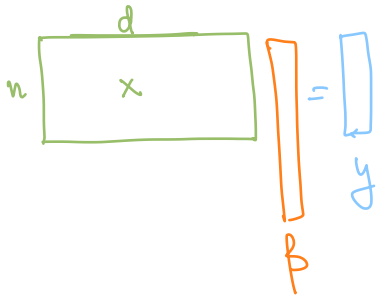
Assumption: $\beta_{1:s} \neq 0$, $\beta_{s+1:d} = 0 \Leftrightarrow$ $\beta \in \mathbb{R}^d$ sparse

$s \ll d$

$\Updownarrow$

$y = \beta_1 x_1 + \dots + \beta_s x_s + \text{noise}$ ($x_{s+1:d}$ = irrelevant)

↑ relevant variables



Convex Optimization for L1 regularization

$$\min_\beta \frac{1}{n} \sum_i (y^i - \beta^T x^i)^2 + \lambda \|\beta\|_1$$

convex fction of β → No local optima
→ algorithms ∃
→ —"— tractable
→ can be analyzed ⇒ $\hat\beta$ sparse

Theorem $y = X^T \beta^* + \varepsilon$, $\|\varepsilon\|$ small, $\varepsilon \in \mathbb{R}^n$

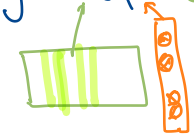
β^* s sparse

X "good", $n \sim s \log_2 \frac{d}{s}$

$\Rightarrow \operatorname{supp} \hat\beta \subseteq \operatorname{supp} \beta^* = S$ → L.S. Regression Re-estimate $\hat\beta_S$ from

$\operatorname{supp} \beta^* = \{j,\ \beta^*_j \neq 0\}$

$y = X_S \beta_S + \varepsilon$



Why $s \log_2 \frac{d}{s}$?

Ph = "find S" + estimate s β_j's ~s data points

#bits ?

$N \in [0 : 2^k - 1] \Rightarrow$ need k bits ($\lceil \log_2 N \rceil$)

S = one of $\binom{d}{s}$ sets (N)

$\Downarrow$

need $\log_2 \binom{d}{s} = \log_2 \frac{d(d-1)\cdots(d-s+1)}{s!} \approx \log_2 \frac{d^s}{s^s} = s \log_2 \frac{d}{s}$ bits