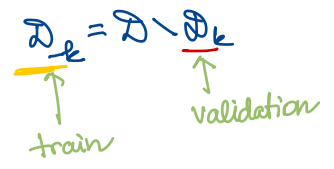


Lecture 22

... $c \nu$
Regularization

If $n \gg 1000$ k -fold CV (disjoint)

1. Partition $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_k$
 $|\mathcal{D}_k| \approx \frac{n}{k}$



2. for $k = 1:k$
train $f_{1:m}$ on $\underline{\mathcal{D}_{-k}} \Rightarrow f_m^{(k)}$
calculate loss $L_m^{(k)} = \text{Loss}(f_m^{(k)}; \underline{\mathcal{D}_k})$

• $\bar{L}^v(f_m) = \frac{1}{k} \sum_{k=1}^k L_m^{(k)}$

• $m^* = \underset{m=1:M}{\text{argmin}} \bar{L}_m^v$

• retrain f_{m^*} on \mathcal{D}

$k \uparrow$ $|\mathcal{D}_{-k}| = \frac{k-1}{k} n$ more data for training

more computation

very small $n \ll 100 \rightarrow \underline{k=n}$

Leave-One-Out CV

$n \downarrow \Rightarrow k \uparrow$

Regularization - form of bias \Rightarrow

towards LOW COMPLEXITY

Examples of regularization

• L1 $\|\beta\|_1$

• L2 $\|\beta\|_2$

• dropout

• weight decay $\|W\|_2^2$

• sparse AE, denoising AE

• Bagging \leftarrow DTrees \rightarrow Random Forests

• early stopping

$$\begin{bmatrix} y = \beta^T x + \epsilon \\ \text{Model} \\ \text{loss LLE} \end{bmatrix}$$

TRAIN $\min_{\beta} \frac{1}{n} \sum_{i=1}^n (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_2^2 \leftarrow \text{L2}$

PREDICT $\hat{y}(x) = \hat{\beta}^T x$
want LLE small

$\|\beta\|_1 \leftarrow \text{L1}$
 $\|\beta\|_4 \leftarrow \text{L4}$

TRAIN
m trees $DT_1, \dots, DT_m \rightarrow$ fit to \mathcal{D} with perturbations of \mathcal{D} of Algo

random subset of $x_{1:d}$
of \mathcal{D} (bootstrap)

PREDICT $f(x) = \frac{1}{m} \sum_{k=1}^m DT_k(x)$ average

$m \sim 100$

$$\|\beta\|_2 = \sqrt{\beta_1^2 + \dots + \beta_d^2}$$

$$\|\beta\|_1 = |\beta_1| + \dots + |\beta_d|$$

TRAIN $\min_{\beta} \frac{1}{n} \sum_{i=1}^n (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_1$ ← ||w|| closed form

$\hat{\beta} = (X^T X + 2\lambda I)^{-1} X^T y$ ← Optimization algo = Lasso

$X = \begin{bmatrix} (x^1)^T \\ \vdots \\ (x^n)^T \end{bmatrix} \in \mathbb{R}^{n \times d}$

$\|\beta\|_1$ ← Convex plo, no closed form
 # $\beta_j \neq 0, j=1:m$ ← l_1 regularization ← no closed form, NOT convex
 0 no regularization ← closed form $\hat{\beta} = (X^T X)^{-1} X^T y$

"TRAIN" = solve optimization

$(X^T X + 2\lambda I) - (X^T X) \geq 0$
Positive definite

Convex optimization - global optimum
- tractable
- efficient algorithm

Why L1? Sparse solutions to $X\beta = y$

Overparametrized

Assume $y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_s x_s + \Sigma$

$\beta_{s+1} = \beta_{s+2} = \dots = \beta_d = 0$ ← $d-s$ irrelevant

$s \ll d$

Need to solve $y = X_s \beta_s$ by LS

Thm X "good", $\|\Sigma\|_2$ small, $n = C s \log_2 \frac{d}{s} \implies \text{supp } \hat{\beta} = \text{supp } \beta^* = S^*$

$y = X\beta^* + \Sigma$, $\beta^* = s$ -sparse, L1-regul

$\{j \mid \beta_j \neq 0\}$

Information limits

#bits = #estimating s parameters + #finding s = $O(s \log_2 d/s)$

$\sim s C_1$

$N = \binom{d}{s} \text{ sets} \Rightarrow \log_2 \binom{d}{s} = \log_2 \frac{d^s}{s^s} = s \log_2 \frac{d}{s}$

$N \in \mathbb{O} = 2^k - 1$ need k bits = $\lceil \log_2 N \rceil$

$$\frac{d(d-1) \cdot (d-s)}{s!}$$

$\Rightarrow m \sim s \log_2 \frac{d}{s}$

Choice of $\lambda \leftarrow CV$
 $\lambda \uparrow$ sparser $\hat{\beta}$

