Lecture 22

...cv

Regularization

If $n \gg 1000$ k-fold CV

1. Partition $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_K$ (disjoint)

$$|\mathcal{D}_k| \approx \frac{n}{K}$$

$\mathcal{D}_{-k} = \mathcal{D} \setminus \mathcal{D}_k$ — $\mathcal{D}_{-k}$: train, $\mathcal{D}_k$: validation

2. for $k = 1:K$

   train $f_{1:M}$ on $\mathcal{D}_{-k} \Rightarrow f_m^{(k)}$

   calculate loss $L_m^{(k)} = \text{Loss}(f_m^{(k)}; \mathcal{D}_k)$

- $\bar{L}^v(f_m) = \frac{1}{K} \sum_{k=1}^{K} L_m^{(k)}$
- $m^* = \underset{m=1:M}{\text{argmin}}\ \bar{L}_m^v$
- retrain f_{m^*} on $\mathcal{D}$

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

more computation

very small n $\lesssim 100$ $\rightarrow$ $K = n$ Leave-One-Out CV

$n \downarrow \Rightarrow K \uparrow$

Regularization – form of bias ⇒ towards LOW COMPLEXITY

Examples of regularization

- L_1 $\|\beta\|_1$
- L_2 $\|\beta\|_2$
- dropout
- weight decay $\|w\|^2$
- sparse AE, denoising AE
- Bagging ← DTrees → Random Forests
- early stopping

$$\left[\begin{array}{l} y = \beta^T x + \varepsilon \\ \text{Model} \end{array}\right]$$

loss L_{LE}

TRAIN $\min_\beta \frac{1}{n}\sum_{i=1}^{n}(y^i - \beta^T x^i)^2 + \lambda_1\|\beta\|^2$ ← ridge

L_{LS} $\|\beta\|_1$ ← lasso

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

want L_{LE} small

m trees $DT_1, \dots DT_m$ → fit to $\mathcal{D}$ with perturbations – of $\mathcal{D}$ / of Algo

TRAIN

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

PREDICT

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

$m \sim 100$

$\|\beta\| \equiv \|\beta\|_2^2 = \beta_1^2 + \dots \beta_d^2$

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

TRAIN $\min_\beta \frac{1}{n}\sum_{i=1}^{n}(y^i-\beta^T x^i)^2+\lambda\|\beta\|^2$ ← ridge

LLS

closed form $\hat\beta=(X^TX+2\lambda I)^{-1}X^Ty$

"TRAIN" = solve optimization

$\|\beta\|_1$ ← Convex pb, no closed form ← optimization algo, ≡ Lasso

$\#\{\beta_j\neq 0, j=1:m\}$ "L0 regularization" ← no closed form, NOT convex

0 no regularization

closed form $\hat\beta=(X^TX)^{-1}X^Ty$

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

$(X^TX+2\lambda I)-(X^TX)\geq 0$

Positive definite

Convex optimization – global optimum
– tractable
– efficient algorithm

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

Overparametrized

Assume $y=\beta_1x_1+\beta_2x_2+..\beta_sx_s+\varepsilon$ → s relevant variables

$\Updownarrow$

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

$s\ll d$

Need to solve $y=X_S\beta_S$ by LS

Thm X "good", $|\varepsilon|_2$ small, $n=Cs\log_2\frac{d}{s}$ $\Rightarrow$ $\operatorname{supp}\hat\beta=\operatorname{supp}\beta^*=S^*$

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

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

Information limits

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

$\sim s C_1$

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

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

$N \in \{0 : 2^k - 1\}$ need k bits $= \lceil \log_2 N \rceil$

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

Choice of $\lambda \leftarrow$ CV

$\lambda \nearrow$ sparser $\hat{\beta}$