23. Model Selection and Validation

Tuesday, March 5, 2024 6:36 AM

The issues we want to solve:

- How to set the hyperparameters in your algorithm?
  (e.g. # rounds T of boosting?)
- How to choose which algorithm or hypothesis class $\mathcal{H}$?
- Go beyond theory (w/ unknown constants...) — we want something practical!
- [SS] presents 2 methods, we'll add in classical methods
  - Correct method depends on the data regime: lots of data? limited data?

Generic setups

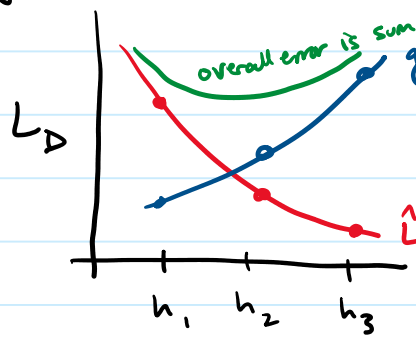
① $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3 \subseteq \dots$ eg. $\mathcal{H}_n = \{$ polynomials of degree n or less $\}$. Choose $n \in \mathbb{N}$

② algo is Tikhonov regularized least squares aka Ridge Regression $\min_w \frac{1}{2}\|X \cdot w - y\|^2 + \alpha \|w\|^2$. Choose $\alpha \geq 0$.

Technique #1: Model Selection via SRM (= Structured Risk Minimization) [SS 11.1]

mostly of theoretical interest.

Idea:



eg. $h_i \in ERM_{\mathcal{H}_i}$

Similarly:

error decomp.: (complexity →)

$$L_D(h) = \left(L_D(h) - \min_{h' \in \mathcal{H}} L_D(h')\right) + \left(\min_{h' \in \mathcal{H}} L_D(h')\right)$$

estimation error or variance — approx. error or bias

SRM based on a (confusing) weight fn

obviously this method only works well if theoretical bound captures right behaviour

For theoretical bound, use something like:

- VC dimension
- Minimum Description Length (MDL) (we'll revisit)
- Rademacher Complexity $L_D(h) \leq \hat{L}_S(h) + \underbrace{R_m(\mathcal{H})}_{\text{use in SRM}} + \sqrt{\frac{\log(1/\delta)}{m}}$

23a. Model Selection and Validation

Tuesday, March 5, 2024 1:35 PM

Technique #2: Validation set i.e. train / validation / test data split

Method of choice (simple, rigorous) if data is plentiful

Let D be any distribution, $S = (z_i)_{i=1}^{m}$ training set, $h = h_S$,

$V = (z_i)_{i=1}^{m_v}$ a validation or test set, $V \cap S = \emptyset$, all iid, then

$\forall\, h$ (either deterministic or independent of V)

$$\left| \hat{L}_V(h) - L_D(h) \right| \le \sqrt{\frac{\log(2/\delta)}{2 m_v}} \quad \text{w.p.} \ge 1-\delta$$

for loss function w/ range in $[0,1]$

proof: Hoeffding

aka $m_v = \frac{1}{2\varepsilon^2} \log(2/\delta)$

How to use? eg. model selection:

if we choose among $\{h_1, \dots, h_k\}$ then we have

(simultaneously) $\forall\, h \in \{h_1, \dots, h_k\}$ $\left| \hat{L}_V(h) - L_D(h) \right| \le \sqrt{\frac{\log(2k/\delta)}{2 m_v}}$

k too large is a bit like p-hacking ... hence always do test error also.

via union bound.

i.e. Select $h \in \underset{i \in [k]}{\operatorname{argmin}} \hat{L}_{V_1}(h_i)$. V_1 = validation

then ... re-estimate L_D using $\hat{L}_{V_2}$ V_2 = testing data

to get unbiased estimate (otherwise we're returning an order statistic)

23b. Model Selection and Validation

Tuesday, March 5, 2024 1:56 PM

Techniques not in [SS]

Technique #3 Mallow's C_p / UPRE

Unbiased Predictive Risk Estimate

cf. Vogel's '02 "Comp. Methods for Inverse Problems" and §7.5 in Hastie, Tibshirani, Friedman's "Elements of Stat. Learning"

Model: $y = X \cdot w_{true} + \eta$ ← noise (i.e. agnostic case)

$m > d$

we'll estimate w_{true} via $\hat{w}$

For a given $\hat{w}$, the predictive error is $p = X \cdot (\hat{w} - w_{true})$

and the predictive risk is $\frac{1}{m}\|p\|^2$

Restrict our attention to estimates $\hat{w}$ that are linear functions of the data y, and parameterized by α, i.e. $\hat{w} = w_\alpha := R_\alpha \cdot y$ ← matrix (≠ Rademacher complexity!)

Main ex.: ridge regression $w_\alpha := \operatorname{argmin} \frac{1}{2}\|X \cdot w - y\|^2 + \alpha \|w\|^2$ (variant: $\|w - w_0\|^2$)

$= \underbrace{(X^T X + \alpha I)^{-1} X^T}_{R_\alpha} y$

always unique sol'n if $\alpha > 0$

normal eq'n (useful theoretically even if we do other tricks numerically)

Question: what value of α to use?

α too small and we might overfit ($m \gg d$) or no unique sol'n ($m < d$)

α too large and we bias $\hat{w}$ too much toward 0. (more bias, less variance)

UPRE looks at $\mathbb{E}\frac{1}{m}\|p_\alpha\|^2$,

$p_\alpha := X \cdot (w_\alpha - w_{true})$

$= r_\alpha + \eta$

observable ↑ unobservable

$r_\alpha = X \cdot w_\alpha - y$ // residual

$= X \cdot w_\alpha - (X w_{true} + \eta)$

so $\mathbb{E}\frac{1}{m}\|p_\alpha\|^2 = \frac{1}{m}\mathbb{E}\|r_\alpha\|^2 + \underbrace{\frac{2}{m}\mathbb{E}\langle r_\alpha, \eta\rangle}_{(A)} + \underbrace{\frac{1}{m}\mathbb{E}\|\eta\|^2}_{=\sigma^2}$

Assume $\mathbb{E}[\eta] = 0$

$\mathrm{Var}(\eta) = \sigma^2 I_{m\times m}$ ("white" noise)

$r_\alpha := X w_\alpha - y$

$= (X R_\alpha - I) y$

$= (X R_\alpha - I)(X w_{true} + \eta)$

let $P := X \cdot R_\alpha$ (uppercase, ≠ p)

$= (P - I)\eta + (P - I) \cdot X \cdot w_{true}$

(A) $= \frac{2}{m}\mathbb{E}\langle (P-I)\eta, \eta\rangle + \frac{2}{m}\mathbb{E}\langle (P-I) X \cdot w_{true}, \eta\rangle$ (second term $= 0$ since $\mathbb{E}[\eta] = 0$)

$= \frac{2}{m}\left(\mathbb{E}\eta^T P \eta - \underbrace{\mathbb{E}\eta^T \eta}_{m\cdot\sigma^2}\right) = \frac{2}{m}\mathbb{E}\eta^T P \eta - 2$

23c. Model Selection and Validation

Tuesday, March 5, 2024 7:25 PM

what is $\mathbb{E}\, \eta^T P \eta$? ($P = P_\alpha$) (assuming $\mathbb{E}[\eta] = 0$, $\mathbb{E}[\eta\eta^T] = \mathrm{Var}[\eta] = \sigma^2 \cdot I$ uncorrelated)

$$\mathbb{E}\, \eta^T P \eta = \mathbb{E}\, \mathrm{tr}(\eta^T P \eta) \overset{\text{cyclic property}}{=} \mathbb{E}\, \mathrm{tr}(P \cdot \eta\eta^T) \overset{\text{linearity}}{=} \mathrm{tr}(P\, \mathbb{E}\, \eta\eta^T)$$

$$= \mathrm{tr}(P \sigma \cdot I) = \sigma\, \mathrm{tr}(P)$$

2 different ways to see it

trace of a 1×1 matrix

$$= \mathbb{E}\left[\sum_i \eta_i^2 P_{ii} + \sum_i \sum_{j \neq i} \eta_i \eta_j P_{ij} \right]$$ ($\sum_{j \neq i} = 0$)

Cyclic property of trace

$\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$

So...

$$\mathbb{E}\, \frac{1}{m} \| P_\alpha \|^2 = \frac{1}{m} \mathbb{E}\, \| r_\alpha \|^2 + \sigma^2 \left(\frac{2}{m} \mathrm{tr}(P_\alpha) - 1 \right)$$

a realization is an unbiased estimate.

r_α is observable

so the C_p statistic is $U(\alpha) = \underbrace{\frac{1}{m} \| r_\alpha \|^2}_{\text{empirical risk}} + \underbrace{\frac{2\sigma^2}{m} \mathrm{tr}(P_\alpha)}_{\text{penalizes complexity}}$

and UPRE chooses the α that minimizes this statistic.

Hastie calls $\frac{1}{m} P_\alpha$ the "effective number of parameters" d

Ex: ridge regression / Tikhonov $P_\alpha = X \cdot R_\alpha = X (X^T X + \alpha I)^{-1} X^T$

$$\mathrm{tr}(P_\alpha) = \mathrm{tr}\left((X^T X + \alpha I)^{-1} X^T X \right)$$

$$= \mathrm{tr}\left(\underbrace{(X^T X + \alpha I)^{-1} (X^T X + \alpha I)}_{= I} - \alpha I \right)$$

$$= \mathrm{tr}(I_{d \times d}) - \alpha\, \mathrm{tr}\left((X^T X + \alpha I)^{-1} \right)$$

$$= d - \sum_{i=1}^{d} \frac{\alpha}{\alpha + \lambda_i(X^T X)}$$

$\lambda_i(X^T X)$ = eigenvalues of $X^T X$

As $\alpha \to \infty$, $\mathrm{tr}(P_\alpha) = 0$ but empirical risk will be large

$\alpha \to 0$, $\mathrm{tr}(P_\alpha) = d$ so "effective dimension" = "actual dimension"

Stein's Unbiased Risk Estimate "SURE" (Stein '81)

generalizes this to nonlinear estimates, $w_\alpha = R_\alpha(y)$ (nonlinear)

and replaces trace w/ divergence of a Jacobian.

ie. $\mathrm{tr}(P_\alpha) \to \mathrm{div}(R_\alpha(y)) := \sum_{i=1}^{m} \frac{\partial r_i}{\partial y_i}$ ←?