ML : Week 4

model selection (cv)

Bias - Var decomp

Motivation

- f is it good?
  - → generalization error
- Model selection.
  - LS $\min_{w \in \mathbb{R}^D} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$

    (λ: hyper. para.)

  - $\{x\} \to w_1$ sol =

    $x_1, x_2 \dots x_k \to w_1 \dots w_k$

- which λ should I use?
- Augmenting the features space = (x_1, x_2)

  → $(x_1, x_2, x_1^2 + x_2^2)$

  $(x_1, x_2, x_1^3 + 5x_2, \dots)$

- Neural Network (Architecture, weight, depth...)

1st thing: consider Data model

samples $S = \{(x_i, y_i)\}_{i=1}^{N} \sim D_{\text{unknown}}$

and i.i.d

Learning Algorithm

$A(S) = f_S$

S: input, f_S: output

Question: How good is f_S?

→ True error:

$$L_D(f) = \mathbb{E}_{(x,y)\sim D}[\ell(y, f(x))]$$

where ℓ is the loss function

Loss functions =

- LS: $l(y,y') = \frac{1}{2}(y-y')^2$
- Logistic function
- Cross entropy

Names in the literature

true risk
expected risk
loss

Problem: D is unknown!

Solution: we have samples $S = \{(x_i, y_i)\}_{i=1}^{N}$ $iid \sim D$

Empirical risk:

$$L_S(f) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, f(x))$$

$$\mathbb{E}_{S\sim D}[L_S(f)] = \mathbb{E}_{S\sim D}\left[\frac{1}{N}\sum_i \ell(y_i, f(x_i))\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{S\sim D}[\ell(y_i, f(x_i))]$$

$$\rightarrow L_D(f)$$

$$\mathbb{E}_{S\sim D}[L_S(f)] = L_D(f) \Rightarrow \text{(Unbiased)}$$

Thus the empirical risk is an unbiased estimator!

But we can have fluctuations

In Application:

Given $S \sim D$

$f_S = alg(S)$

then $L_S(f_S) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f_S(x_i))$

↑ Training error

In general, $L_S(f_S) \neq L_D(f_S)$

Basic idea:

Split data S:

$S = S_{train} \cup S_{test}$

- learn $f_{S_{train}}$ on S_{train} (using training data)
- validate it using S_{test}: $L_{Test}(f_{S_{train}})$

$$\Rightarrow L_{Test}(f_{S_{Train}}) = \frac{1}{|S_{test}|} \sum_{i=1}^{|S_{test}|} \ell(y_i, f_{S_{train}}(x_i))$$

Test/validation risk (error)

$$\mathbb{E}_{S_{Test} \sim D}\left[L_{S_{Test}}(f_{S_{train}})\right] = L_D(f_{S_{train}})$$

- **Fluctuations:**

. Fixed funct. f

S samples (S was not used to learn f)

$$|L_D(f) - L_S(f)| \leq \varepsilon$$

→ Can we bound this?

→ $L_D(f) \leq L_S(f) + \varepsilon$

($L_S(f)$: we can compute)

Assumptions:

. $l \in [0,1]$

. $S \sim D$ iid $|S|$

. f

Claim:

$$\mathbb{P}\left[|L_D(f) - L_S(f)| \geq \sqrt{\frac{\log 2/\delta}{2|S|}}\right] \leq \delta$$

Proof we reduce it to the standard inequality:

$\theta_n \sim iid \in [a,b]$

$\tilde{\theta} = \mathbb{E}[\theta_n]$

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\theta_i - \mathbb{E}[\theta_i]\right| \geq \varepsilon\right] \leq 2 \cdot e^{-\frac{2N\varepsilon^2}{(a-b)^2}}$$

Concentration inequality.

Equivalence: $\theta_i = \ell(y_i, f(x_i))$

$$\varepsilon = \sqrt{\frac{\log 2/\delta}{2N}}$$

How can we use it?

- f given, S given, Is f good?

$$\mathbb{P}\left(L_D(f) \geq L_S(f) + \sqrt{\frac{(a-b)^2 \log 2/\delta}{2|S|}}\right) \leq \delta$$

- S

We split $S = S_{Train} \cup S_{Test}$

$A(S_{Train}) = f_{S_{Train}}$

$$P\left[L_D(f_{S_{Train}}) \geq L_{S_{Test}}(f_{S_{Train}}) + \sqrt{\frac{(a-b)^2 \log 2/\delta}{2|S_{Test}|}}\right] \leq \delta$$

→ bound on the generalization error

Generalization error = $L_D - L_S$

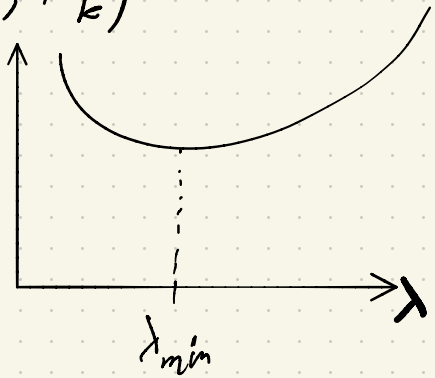
2nd

Model Selection

We have K hypo. $(\lambda_1, \lambda_2 \dots \lambda_k)$

$S = S_{Train} \cup S_{Test}$

- $f_{S_{Train}, \lambda_k} = A(S_{Train}, \lambda_k)$
- $L_{S_{Test}}(f_{S_{Train}, \lambda_k})$



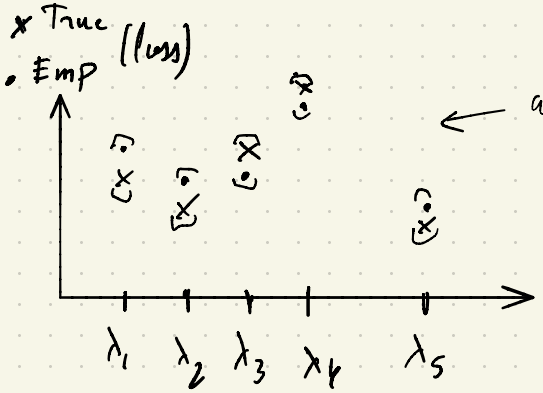
Union bound: $f_k = f_{S_{Train}, \lambda_k}$ ∃ Take This λ!

$$\mathbb{P}\left[\max_k \left| L_D(f_k) - L_{S_{Test}}(f_k) \right| \geq \varepsilon\right] \leq \delta$$

$$\leq \sum_k \mathbb{P}\left[\left| L_D(f_k) - L_{S_{Test}}(f_k) \right| \geq \varepsilon\right]$$

$$\leq 2K e^{-\frac{2N\varepsilon^2}{(a-b)^2}}$$

$$\mathbb{P}\left[\max_k \left| L_D(f_k) - L_{S_{Test}}(f_k) \right| \geq \sqrt{\frac{(b-a)^2 \log\frac{2k}{\delta}}{2|S_{Test}|}}\right] \leq \delta$$



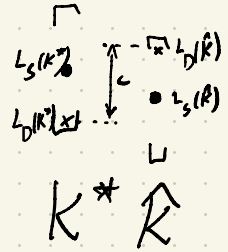
← all the width are the same

- $\hat{K} = \arg\min_k L_{S_{Test}}(f_{S_{Train}}, \lambda_k)$

$$K^* = \arg\min_k L_D(f_{S_{Train}}, \lambda_k)$$

$$P\left[|L_D(f_{K^*}) - L_D(f_{\hat{K}})| \geq 2\sqrt{\frac{(b-a)^2 \log \frac{2K}{\delta}}{2|S_{Test}|}} \right] \leq \delta$$

Worst case:



$$\epsilon \leq 2\sqrt{\frac{(b-a)^2 \log \frac{2K}{\delta}}{2|S_{Test}|}}$$

- Extension to infinitely many models ($K = \infty$) (Lecture notes)

Proof of Chernoff bound:

$a < \Theta < b$

Without loss of generality; assume $\mathbb{E}[\Theta] = 0$

we will prove that $\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N}\Theta_i \geq \varepsilon\right] \leq e^{-\frac{2N\varepsilon^2}{(b-a)^2}}$

together with $\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N}\Theta_i \leq -\varepsilon\right] \leq e^{-\frac{2N\varepsilon^2}{(b-a)^2}}$

$$\rightarrow \mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\Theta_i\right| \geq \varepsilon\right] \leq 2e^{-\frac{2N\varepsilon^2}{(b-a)^2}}$$

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N}\Theta_i \geq \varepsilon\right] \overset{s>0}{=} \mathbb{P}\left[\frac{s}{N}\sum\Theta_i \geq s\varepsilon\right]$$

$$= \mathbb{P}\left[e^{\frac{s}{N}\sum\Theta_i} \geq e^{s\varepsilon}\right]$$

$$\text{(markov)} \leq \min_s \mathbb{E}\left[e^{\frac{s}{N}\sum\Theta_i}\right] \cdot e^{-s\varepsilon}$$

Hoeffding lemma: $\mathbb{E}[e^{sx}] \leq e^{\frac{1}{8}s^2(b-a)^2}$

$$\Rightarrow \leq \min_s \underbrace{e^{s^2(b-a)^2/8N} \cdot e^{-s\varepsilon}}_{e(s)}$$

$$\leq e^{-\frac{2N\varepsilon^2}{(b-a)^2}} \quad \text{for } s = \frac{4N\varepsilon}{(b-a)^2}$$

Proof of Hoeffding lemma: by convexity:

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

$\mathbb{E}[x] = 0$

$$\mathbb{E}[e^{sx}] \leq \frac{be^{sa} - ae^{sb}}{b-a} \leq \dots \leq e^{s^2\frac{(b-a)^2}{8}}$$