Mathematics of Machine Learning

1   Introduction

- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint distribution P_0

- classification setting $\mathcal{Y} \in \{-1, 1\}$

- regression setting $\mathcal{Y} = \mathbb{R}$

Assumption 1. $\mathcal{X} \in \mathbb{R}^p$

- hypothesis $h : \mathcal{X} \to \mathcal{Y}$

- loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

- **Classification setting**

- misclassification error $l(h(x), y) = \begin{cases} 1 & \text{if } h(x) = y \\ 0 & \text{otherwise} \end{cases}$

- classifier h

- **Regression setting**

- squared error $l(h(x), y) = (h(x) - y)^2$

- risk $R(h) = \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} l(h(x), y) dP_0(x, y)$

Fact. $R(h) = \mathbb{E}l(h(X), Y)$ *for deterministic h*

Setting 1. *l misclassification error, R risk*

- Bayes classifier h_0 —— minimises misclassification risk

- Bayes risk $R(h_0)$

- regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$

Proposition 1.1. *Bayes classifier h_0, then $h_0(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$*

> *Proof.* $R(h) = \frac{1}{4}\mathbb{E}(Y - h(X))^2 = \frac{1}{4}\mathbb{E}(Y - \mathbb{E}(Y|X))^2 + \frac{1}{4}\mathbb{E}(\mathbb{E}(Y|X) - h(X))^2$ $\qquad\qquad\square$

Setting 2. P_0 *unknown*

- training data (X_i, Y_i) —— i.i.d. of (X, Y)

- $R(\hat{h})$

Fact. $R(\hat{h}) = \mathbb{E}(l(h(X), Y) \mid X_1, Y_1, \ldots, X_n, Y_n)$

- class $\mathcal{H}$ of hypotheses

Example. *(i)* $\mathcal{H} = \{h : h(x) = \operatorname{sgn}(\mu + x^\top \beta)\}$

(ii) $\mathcal{H} = \{h : h(x) = \operatorname{sgn}(\mu + \sum \phi_j(x)\beta_j)\}$ *with dictionary* $\phi_i : \mathcal{X} \to \mathbb{R}$

Setting 3. $\operatorname{sgn}(0) = -1$

- conditional expectation $\mathbb{E}(Z \mid W)$

> **Proposition 1.2.**
>
> *(i)* ***Role of independence*** $\mathbb{E}(Z|W) = \mathbb{E}Z$
>
> *(ii)* ***Tower property*** $\mathbb{E}[\mathbb{E}(Z|W) \mid f(W)] = \mathbb{E}[Z \mid f(W)]$
>
> *(iii)* ***Taking out what is known*** $\mathbb{E}(f(W)Z|W) = f(W)\mathbb{E}(Z|W)$
>
> *(iv)* ***Conditional Jensen*** $\mathbb{E}(f(Z)|W) \geq f(\mathbb{E}(Z|W))$ —— f *convex,* $f(Z)$ *integrable*

- empirical risk / training error $\hat{R}(h) = \frac{1}{n}\sum l(h(X_i), Y_i)$

- empirical risk minimiser (ERM) $\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{R}(h)$ (multiple minimiser)

- generalisation error $R(\hat{h})$

- $h^* \in \arg\min_{h \in \mathcal{H}} R(h)$

- stochastic error / excess risk $R(\hat{h}) - R(h^*)$ —— increase with complexity of $\mathcal{H}$

- approximation error $R(h^*) - R(h_0)$ —— decrease with complexity of $\mathcal{H}$

Fact. $R(\hat{h}) - R(h_0) = $ *excess risk + approximation error*

2 Statistical learning theory

Fact. $R(\hat{h}) - R(h^*) = \left(R(\hat{h}) - \hat{R}(\hat{h})\right) + \left(\hat{R}(\hat{h}) - \hat{R}(h^*)\right) + \left(\hat{R}(h^*) - R(h^*)\right)$

– concentration inequalities

Fact (Markov's inequality)**.** W *non-negative,* ϕ *strictly increasing, then* $\mathbb{P}(W \geq t) \leq \frac{\mathbb{E}\phi(W)}{\phi(t)}$

Fact (Chernoff bound)**.** $\phi(t) = e^{\alpha t}$, $\alpha > 0$, *then* $\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t}\mathbb{E}e^{\alpha W}$

– sub-Gaussian with parameter σ —— $\mathbb{E}e^{\alpha(W - EW)} \leq e^{\frac{\alpha^2 \sigma^2}{2}}$

Proposition 2.1. W *sub-Gaussian with* σ*, then* $\mathbb{P}(W \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$

Proof. Chernoff bound $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Fact. W *sub-Gaussian with* σ*, then*

(i) W *sub-Gaussian with* σ' *for all* $\sigma' \geq \sigma$

(ii) $-W$ *sub-Gaussian with* σ

Fact. $\mathbb{P}(|W - \mathbb{E}W| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$

Proposition 2.2. W_i *independent, sub-Gaussian with* σ_i, ~~*mean* μ_i~~*,* *then* $\gamma^\top W$ *sub-Gaussian with* $\sqrt{\sum_i \gamma_i \sigma_i}$

Proof. expand $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Fact. *same setting, pick* $\gamma = (1, \ldots, 1)$, *then* $\mathbb{P}\left(\sum_i (W_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_i \sigma_i^2}\right)$

Proposition 2.3. W_i *mean 0, sub-Gaussian with* σ *(non necessarily independent),* *then* $\mathbb{E}\max_j W_j \leq \sigma\sqrt{2\log(d)}$

Proof. $\exp(\alpha\mathbb{E}\max W_j) \leq \mathbb{E}\exp(\alpha\max W_j) \leq \sum \exp(\alpha W_j) \leq de^{\frac{\alpha^2\sigma^2}{2}}$, *then maximise over* α $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

– Rademacher r.v. ϵ —— take $\{-1, 1\}$ with equal prob

Fact. *Rademacher* ϵ *sub-Gaussian with* $\sigma = 1$

Lemma 2.4 (Hoeffding's lemma)**.** *W mean 0, take values in $[a,b]$, then W sub-Gaussian with $\sigma = \frac{b-a}{2}$*

Proof. weaker result $\sigma = b - a$: consider independent W', conditional Jensen, Rademacher sub-Gaussian, $\mathbb{E}e^{\alpha W} \leq \mathbb{E}e^{\alpha\epsilon(W-W')} \leq \mathbb{E}e^{\alpha^2(W-W')^2/2} \leq \mathbb{E}e^{\alpha^2(b-a)^2/2}$ □

– symmetrisation argument

Fact (Hoeffding's inequality)**.** *W_i independent, mean 0, $a_i \leq W_i \leq b_i$ a.s. , then $\mathbb{P}(\frac{1}{n}\sum_i W_i \geq t) \leq \exp\left(-\frac{2n^2t^2}{\sum_i(b_i-a_i)^2}\right)$*

Theorem 2.5. *$\mathcal{H}$ finite, l take values in $[0, M]$, then with probability at least $1 - \delta$, $R(\hat{h}) - R(h^*) \leq M\sqrt{\frac{2(\log|\mathcal{H}|+\log\frac{1}{\delta})}{n}}$*

Proof. decomposition $R(\hat{h}) - R(h^*)$, then Hoeffding's inequality □

– $G(X_1, Y_1, \ldots, X_n, Y_n) = \sup_{h\in\mathcal{H}} R(h) - \hat{R}(h)$

Fact. *l takes values $[0, M]$, then $G(x_1, y_1, \ldots, x_n, y_n) - G(x'_1, y'_1, x_2, y_2, \ldots, x_n, y_n) \leq \frac{M}{n}$*

– $a_{j:k}$ —— subsequence $a_j, \ldots, a_k$

– bound differences property:
  $f(w_1, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_n) - f(w_1, \ldots, w_{i-1}, w'_i, w_{i+1}, \ldots, w_n) \leq L_i$

Theorem 2.6 (Bounded differences inequality)**.** *f bound differences property, W_i independent, then $\mathbb{P}(f(W_{1:n}) - \mathbb{E}f(W_{1:n}) \geq t) \leq \exp\left(-\frac{2t^2}{\sum_i L_i^2}\right)$*

Proof. (D_i) martingale difference wrt Doob martingale, $F_i(w_{1:i}) = \mathbb{E}(f(W_{1:n}|W_{1:i} = w_{1:i}))$
$\begin{cases} A_i = \inf_{w_i} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W_{1:n}|W_{1:i-1})) \\ B_i = \sup_{w_i} F_i(W_{1:(i-1)}, w_i) - \mathbb{E}(f(W_{1:n}|W_{1:i-1})) \end{cases}$, then use $W_{(i+1:n)}$ independent to W_i, then
Azuma-Hoeffding □

– martingale sequence $(Z_i)_{i\geq 0}$ wrt $(W_i)_{i\geq 0}$ ——

  (i) $\mathbb{E}|Z_i| < \infty$
  (ii) Z_i $\sigma(W_{0:i})$-measurable
  (iii) $\mathbb{E}(Z_i|W_{0:(i-1)}) = Z_{i-1}$

– martingale difference sequence $D_i = Z_i - Z_{i-1}$

– Doob martingale $Z_i = \mathbb{E}f(W_{1:n})|W_{1:i}$ —— martingale provided $\mathbb{E}|f(W_{1:n})| < \infty$

Lemma 2.7. (D_i) *martingale difference sequence wrt* (W_i), $\mathbb{E}(e^{\alpha D_i}|W_{0:i-1}) \le e^{\frac{\alpha^2 \sigma_i^2}{2}}$, *then* $\gamma^\top D$ *sub-Gaussian with* $\sqrt{\sum \gamma_i^2 \sigma_i^2}$

Proof. Tower property with $\sigma(W_{1:i})$ for $i = n-1, n-2, \ldots, 1$ □

Theorem 2.8 (Azuma-Hoeffding). (D_i) *martingale difference sequence wrt* (W_i), $\exists \sigma(W_{0:(i-1)})$-*measurable* A_i, B_i, *constant* L_i *st*

 (i) $A_i \le D_i \le B_i$

 (ii) $B_i - A_i \le L_i$

, *then* $\mathbb{P}\left(\sum_i D_i \ge t\right) \le \exp\left(-\frac{2t^2}{\sum L_i^2}\right)$

Proof. Hoeffding's Lemma conditionally on $W_{0:(i-1)}$, then lemma, then Gaussian tail bound □

Setting 4. $\mathcal{H}$ *(possibly infinite) hypothesis class,* l *takes values in* $[0, M]$

Fact. $R(\hat{h}) - R(h^*) \le (G - \mathbb{E}G) + \mathbb{E}G + \hat{R}(h^*) - R(h^*)$

 – $Z_i = (X_i, Y_i)$

 – $\mathcal{F} = \{(x, y) \mapsto -l(h(x), y) : h \in \mathcal{H}\}$

Fact. $G = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum (f(Z_i) - \mathbb{E}f(Z_i))$

 – $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i f(Z_i)\right)$ —— ϵ_i i.i.d. Rademacher independent of $Z_{1:n}$

Intuition. *capture how closely* $f(Z_i)$ *align with random label* ϵ_i *(dot product)*

Theorem 2.9. $\mathcal{F}$ *class of real functions,* Z_i *i.i.d.* , *then* $\mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum (f(Z_i) - \mathbb{E}f(Z_i))\right) \le 2\mathcal{R}_n(\mathcal{F})$

Proof. Z_i' i.i.d. copy of Z_i, symmetrisation technique:
$\sup \frac{1}{n} \sum f(Z_i) - \mathbb{E}f(Z_i) \le \mathbb{E}\left(\sup \frac{1}{n} \sum f(Z_i) - f(Z_i')|Z_{1:n}\right)$ □

 – $\mathcal{F}(z_{1:n}) = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$

 – empirical Rademacher complexity $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i f(z_i)\right)$

 – $\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n})) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \epsilon_i f(Z_i) \mid Z_{1:n}\right)$

Fact. $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\hat{\mathcal{R}}(\mathcal{F}(Z_{1:n}))$

Theorem 2.10 (Generalisation bound based on Rademacher complexity).
$\mathcal{F} = \{(x,y) \mapsto l(h(x),y)\}$, l *takes values in* $[0, M]$,

then with probability at least $1 - \delta$, $R(\hat{h}) - R(h^*) \leq 2\mathcal{R}_n(\mathcal{F}) + M\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}$

Proof. decomposition: $R(\hat{h}) - R(h^*) \leq (G - \mathbb{E}G) + \mathbb{E}G + \hat{R}(h^*) - R(h^*)$

Bounded differences inequality: $\mathbb{P}\left(G - \mathbb{E}G \geq \frac{t}{2}\right) \leq \exp\left(-\frac{t^2 n}{2M^2}\right)$,

Hoeffding's inequality: $\mathbb{P}\left(\hat{R}(h^*) - R(h^*) \geq \frac{t}{2}\right) \leq \exp\left(-\frac{t^2 n}{2M^2}\right)$

$\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(-\mathcal{F})$, so $\mathbb{E}G \leq 2\mathcal{R}_n(\mathcal{F})$, then $t = M\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$ $\qquad \square$

Setting 5. *classification setting, misclassification loss,* $\mathcal{F} = \{(x,y) \mapsto l(h(x),y) : h \in \mathcal{H}\}$

Fact. $|\mathcal{F}(z_{1:n})| = |\mathcal{H}(x_{1:n})|$

Lemma 2.11. $\hat{R}(\mathcal{F}(z_{1:n})) \leq \sqrt{\frac{2\log|\mathcal{F}(z_{1:n})|}{n}} = \sqrt{\frac{2\log|\mathcal{H}(x_{1:n})|}{n}}$

Proof. $\mathcal{F}' = \{f_1, \dots, f_d\}$ st $\mathcal{F}'(z_{1:n}) = \mathcal{F}(z_{1:n})$, $W_j = \frac{1}{n}\sum \epsilon_i f_j(z_i)$, then W_j sub-Gaussian with $\sigma = \frac{1}{\sqrt{n}}$, then apply max bound $\qquad \square$

Setting 6. $\mathcal{F}$ *class of functions* $f : \mathcal{X} \mapsto \{a, b\}$, $\mathcal{F} \geq 2$

- $\mathcal{F}$ shatters $x_{1:n}$ —— $|\mathcal{F}(x_{1:n})| = 2^n$

- shattering coefficient $s(\mathcal{F}, n) = \max_{x_{1:n}} |\mathcal{F}(x_{1:n})|$

- VC dimension $VC(\mathcal{F}) = \sup\{n : s(\mathcal{F}, n) = 2^n\}$

Lemma 2.12 (Sauer-Shelah). $VC(\mathcal{F}) = d$, *then* $s(\mathcal{F}, n) \leq \sum_0^d \binom{n}{i} \leq (n+1)^d$

Proof. non-empty $Q \subset [n]$, stronger statement: at least $|\mathcal{F}(x_{1:n})| - 1$ non-empty Q st $\mathcal{F}$ shatters x_Q, then induction on $|\mathcal{F}(x_{1:n})|$ $\qquad \square$

Fact. $\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2VC(\mathcal{F})\log(n+1)}{n}}$

Setting 7. $\mathcal{F}$ *vector space of functions,* $\mathcal{H} = \{h : h(x) = \text{sgn}(f(x)), f \in \mathcal{F}\}$

Example. $\mathcal{X} = \mathbb{R}^p$, $\mathcal{F} = \{x \mapsto x^\top \beta : \beta \in \mathbb{R}^p\}$