Supervised learning: data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$

$\ell(y, f(x))$ loss function

↳ prediction $f(x) \in \mathcal{Y}$

goal: $\min \mathbb{E}\, \ell(y, f(x)) = R(f)$ -

↳ testing distribution

- $f^*(x)$ Bayes predictor for square loss $\mathbb{E}(y/x)$

for 0-1 loss: $\operatorname{argmax} p(y/x)$

- local averaging (k-nn)
- empirical risk minimization

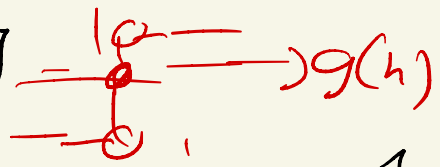
- $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$

empirical risk

× dealing with $\mathcal{Y}$ discrete

× Estimation error

Binary classification: $y = \{-1, 1\}$



$f: X \to \{-1, 1\}$

$g: X \to \mathbb{R}$ with $f(x) = \text{sign}(g(x))$ $\begin{cases} = 1 & \text{if } g(x) > 0 \\ = 0 & \text{if } g(x) = 0 \\ = -1 & \text{if } g(x) < 0 \end{cases}$

0-1 error:

$$\ell(y, f(x)) = \underbrace{1_{y \neq f(x)}}_{1_{yg(x) \leq 0}} = \ell_{0-1}(yg(x))$$

square loss

$(1-u)^2$ $\quad (1-yg(x))^2 = (y-g(x))^2$

because $y \in \{+1, -1\}$

convex surrogate

logistic loss

$\log(1+e^{-u})$

$\max(1-u, 0) = (1-u)_+$

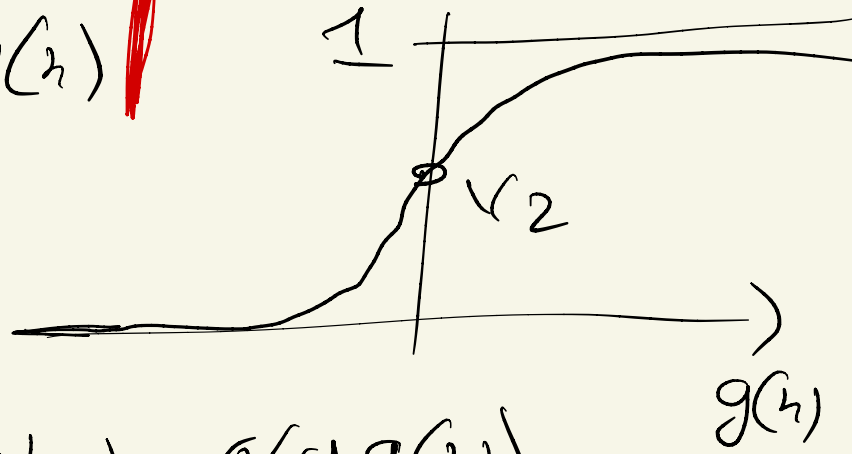
hinge loss $\Rightarrow$ SVM

logistic loss: $\log(1+e^{-u})$

probabilistic interpretation with conditional probabilistic model

$$p(y=1|x) = \sigma(g(x)) = \frac{1}{1+e^{-g(x)}}$$

sigmoid



$$p(y=-1|x) = 1 - \sigma(g(x))$$
$$= \sigma(-g(x)) \Rightarrow p(y|x) = \sigma(y\,g(x))$$

Maximum likelihood with independent data

$$\max \sum_i \log p(y_i|x_i) = \max \sum_i \log \sigma(y_i g(x_i))$$
$$= \max \sum_i -\log(1+e^{-y_i g(x_i)})$$

"cross-entropy loss"

Analysis of convex surrogates :

square, hinge, logistic, exp

0-1 error $R_{01}(g) = \mathbb{E}\, \phi_{01}(y g(h)) -$

$R_\phi(g) = \mathbb{E}\, \phi(y g(h)) -$

goal: if $R_{01}(g) - R^*_{01}(g)$ is small

if $R_\phi(g) - R^*_\phi(g)$ is small

Square loss ① check that optimal predictors are the same

② $R_{01}(g) - R^*_{01}(g) \leq \sqrt{R_{square}(g) - R^*_{square}(g)}$

optimal predictor:

$y(x) \in \{-1, 1\}$

0-1 loss: $f^*(x) = 1$ if $P(y=1|x) > 1/2 = \text{sign}(\mathbb{E}(y|x))$

-1 if $P(y=-1|x) > 1/2$

$\mathbb{E}(y|x) = 1 \cdot P(y=1|x) + (-1) P(y=-1|x) = 2P(y=1|x) - 1 = 2\eta(x) - 1$

Lemma: $R_{0-1}(g) - R^*_{0-1}(g) \leq \mathbb{E}\left|2\eta(x) - 1 - g(x)\right|$

Consequences

$$\leq \mathbb{E}\left|g^*_{square}(x) - g(x)\right|$$

$$\leq \sqrt{\mathbb{E}\left(g^*_{square}(x) - g(x)\right)^2} \quad \text{(Jensen)}$$

$$= \sqrt{R_{square}(g) - R^*_{square}}$$

Proof: $R_{0-1}(g) - R^*_{0-1}(g)$

$$= \mathbb{E}\left[1_{yg(x)\leq 0} - 1_{yg^*(x)\leq 0}\right]$$

$$= \mathbb{E}\left(\mathbb{E}\left(1_{yg(x)\leq 0} - 1_{yg^*(x)\leq 0} \,\middle|\, x\right)\right)$$

Non zero only when $g(x)$ and $g^*(x)$ have different signs

two cases: ① $g(x) > 0$, $g^*(x) < 0$; ② $g(x) < 0$, $g^*(x) > 0$

$$\mathbb{E}\left(1_{yg(x)\leq 0} - 1_{yg^*\leq 0} \,\middle|\, x\right) = \underbrace{1_{y=-1}}_{1-\eta(x)} - \underbrace{1_{y=+1}}_{\eta(x)}$$

$$= P(y=-1 \mid x) - P(y=1 \mid x)$$

$$= 1 - 2\eta(x)$$

$$\leq 1 - 2\eta(x) + g(x)$$

$$= \left|-2\eta(x) + g(x)\right|$$

Decomposition of the risk: $\ell(y, f(x))$ where $f(x) \in \mathbb{R}$

implicitly = $\ell(y f(x))$ & logistic loss

$R(f) = E\,\ell(y, f(x))$ expected risk | if $y \in \{-1, 1\}$

"generalization error"

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

, class of models $\mathcal{F} : \{ f : \mathcal{X} \to \mathbb{R} \}$

Empirical risk minimizer

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}(f)$$

$R(f)$ "Bayes error"

$f \in \mathcal{F}$

goal: $R(\hat{f}) - R^* = \inf$ all functions f

$$= \underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)}_{\text{estimation error} \to 0 \text{ random } \downarrow n} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R^*}_{\text{approximation error} \to 0 \text{ deterministic } \downarrow \mathcal{F}} =$$

Estimation error :

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = R(\hat{f}) - R(f_{\mathcal{F}}^*)$$

$f_{\mathcal{F}}^* \in \operatorname{argmin}_{f \in \mathcal{F}} R(f)$

$$= R(\hat{f}) - \hat{R}(\hat{f}) + \underbrace{\hat{R}(\hat{f}) - \hat{R}(f_{\mathcal{F}}^*)}_{\leq 0} + \hat{R}(f_{\mathcal{F}}^*) - R(f_{\mathcal{F}}^*)$$

because $\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}(f)$

$$\leq \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) + \sup_{f \in \mathcal{F}} \hat{R}(f) - R(f)$$

estimation error $\leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$

(x_i, y_i) IID

If $f \in \mathcal{F}$ fixed : $\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$

Hoeffding

Central limit theorem

$\frac{1}{n} \sum Z_i \sim \text{Normal}(\mathbb{E}Z, \frac{\sigma^2}{n}) = \frac{1}{n} \sum_{i=1}^{n} Z_i \longrightarrow \mathbb{E}Z$

law of large #'s

Hoeffding's inequality: Z_i IID, $Z_i \in [0,1]$ almost surely

$$P\left(\left|\frac{1}{n}\sum_i Z_i - \mathbb{E}Z\right| \geq t\right) \leq 2e^{-2nt^2} = \delta$$

with probability greater than $1-\delta$

$$\left|\frac{1}{n}\sum_i Z_i - \mathbb{E}Z\right| \leq \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

$$nt^2 = \log\frac{2}{\delta}$$

$$t = \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Proof: back at chapter 1 S

$$P(S_n - t) \leq \frac{\mathbb{E}(e^{-\lambda S})}{e^{\lambda t}}$$

Markov / inequality Chernoff

$$2e^{-\frac{2nt^2}{(b-a)^2}}$$

$Z \in (a,b) \Rightarrow$

Finite number of models: $\mathcal{F} = \{f_1, \dots, f_m\}$, $|\mathcal{F}| = m$

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$$

$\forall f$, $P(|R(f) - \hat{R}(f)| \geq t) \leq 2\exp\left(-\frac{2nt^2}{4c_\infty^2}\right)$

if $R(f)$ almost surely $\in (a, b)$

$R(f) - \hat{R}(f) \in [-c_\infty, c_\infty)$

$$P\left(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \geq t\right) = P\left(\bigcup_{f \in \mathcal{F}} \{|R(f) - \hat{R}(f)| \geq t\}\right)$$

union bound

$$\leq \sum_{f \in \mathcal{F}} P(|R(f) - \hat{R}(f)| \geq t)$$

$$\leq 2e^{-nt^2/2c_\infty}$$

$$\leq 2|\mathcal{F}| \exp(-nt^2/2c_\infty) = \delta$$

With prob $1 - \delta$

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \leq \sqrt{\frac{\log 2|\mathcal{F}|/\delta \; 2c_\infty}{n}} = \sqrt{\frac{2c_\infty}{n}} \sqrt{\log|\mathcal{F}| + \log\frac{2}{\delta}}$$

Rademacher complexity:

$$\sup_{f \in \mathcal{F}} \hat{R}(f) - R(f)$$

definition

$$\mathcal{H} = \left\{ h : (x,y) \mapsto \ell(y, f(x)) \;\middle|\; f \in \mathcal{F} \right\}$$

$$= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \ell(y_i, f(x_i)) - \mathbb{E}\, \ell(y_i, f(x_i))$$

$$= \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i h(z_i) - \mathbb{E}\, h(z)$$

$(z = (x,y))$

goal: get a bound on

$$\mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i h(z_i) - \mathbb{E}\, h(z) \leq 2 R_n(\mathcal{H})$$

Def. Rademacher complexity of $\mathcal{H}$.

$$R_n(\mathcal{H}) = \mathbb{E}_{\varepsilon; z_1, \dots, z_n} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i)$$

with ε a vector of Rademacher random variables

$$P(\varepsilon_i = \pm 1) = 1/2 \qquad \forall i \in \{1, \dots, n\}$$

Symmetrization lemma:

Proof: $D = \{z_1, \dots z_n\}$, independent copy $D' = \{z'_1, \dots z'_n\}$

$$\mathop{E}_{D} \sup_{h \in H} \frac{1}{n} \sum_i h(z_i) - Eh(z) = \mathop{E}_{D,D'} \sup_{h \in H} \frac{1}{n} \sum_i \left[h(z_i) - E(h(z'_i) \mid D) \right]$$

$$= \mathop{E}_{D,D'} \sup_{h \in H} \frac{1}{n} \sum_i E(h(z_i) - h(z'_i) \mid D)$$

$$- E\left(\frac{1}{n} \sum_i h(z_i) - h(z'_i) \,\middle|\, D \right)$$

$$\leq \mathop{E}_{D,D'} E\left(\sup_{h \in H} \frac{1}{n} \sum_i h(z_i) - h(z'_i) \,\middle|\, D \right)$$

$\sup E \leq E \sup$

$E E(\cdot \mid T) = E \cdot$

$$= \mathop{E}_{D,D',\varepsilon} \left(\sup_{h \in H} \frac{1}{n} \sum_i \varepsilon_i \left(h(z_i) - h(z'_i) \right) \right)$$

$$\frac{1}{n} \sum_i \varepsilon_i h(z_i) - \frac{1}{n} \sum_i \varepsilon_i h(z'_i) \quad \sup A + B \leq \sup A + \sup B$$

$$\leq \mathop{E}_{D,D',\varepsilon} \sup_{h \in H} \frac{1}{n} \sum_i \varepsilon_i h(z_i) + \mathop{E}_{D,D',\varepsilon} \sup_{h} \frac{1}{n} \sum_i (-\varepsilon_i) h(z'_i)$$

$R_n(H) <$ $\qquad R_n(H) = 2 R_n(H)$

$h(z) = h(x,y) = \ell(y, f(x))$

$\sum \varepsilon_i \ell(y_i, f(x_i))$ différence

$= \sum \varepsilon_i \cdot \varphi_i(f(x_i))$

Lipschitz constantes with constant G

$|\varphi_i(a) - \varphi_i(s)| \leq G|a-s|$

see proof in book

Contraction principle

$$\text{Rad}: \quad E_\varepsilon \sup_f \sum_{i=1}^n \varepsilon_i \varphi_i(f(x_i)) \leq G \cdot E_\varepsilon \sup_f \sum_{i=1}^n \varepsilon_i f(x_i)$$

Conséquence : ℓ Nelos function one G-Lipschitz

$$R_n(H) \leq G \cdot R_n(F)$$

Exemple : linear functions : $F = \{ f : x \to \theta^T \varphi(x) \quad \|\theta\|_2 \leq D \}$

$$R_n(F) = E \sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum_i \varepsilon_i \theta^T \varphi(x_i)$$

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}\sup_{\|\theta\|_2 \le D} \left(\frac{1}{n}\sum_i \varepsilon_i \underbrace{\phi(x_i)^T}_{f(x_i)} \theta \right)$$

$$\sup_{\|\theta\|_2 \le D} \theta^T M = \|M\|_2 \cdot D$$

$$= \mathbb{E}\left\| \frac{1}{n}\sum_i \varepsilon_i \phi(x_i) \right\|_2 \times D$$

Jensen's inequality

$$\le \sqrt{\mathbb{E}\left\| \frac{1}{n}\sum_i \varepsilon_i \phi(x_i) \right\|_2^2 \times D}$$

$$= \sqrt{\frac{1}{n^2}\sum_i \mathbb{E}\,\underbrace{\varepsilon_i^2}_{1} \|\phi(x_i)\|^2 \times D}$$

because
$\mathbb{E}\varepsilon_i = 0$
+ iid

$$\hat{R}_n(\mathcal{F}) = \frac{1}{\sqrt{n}} D \times \mathbb{E}\|\phi(x_1)\|^2$$

Next week:
- summary chapter 4
- chapter 5: optimization