Lecture 4 - theory basics

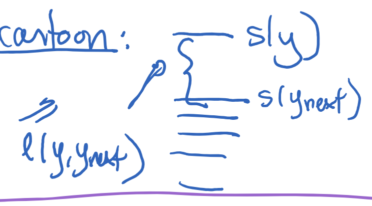
Thursday, January 26, 2023 1:28 PM

today: · finish surrogate losses
· theory basics

III) <u>structured hinge loss</u>

$$f^{SVM}(x,y;w) = \underbrace{\max_{\tilde{y}\in\mathcal{Y}(x)} \left[s(x,\tilde{y};w) + \ell(y,\tilde{y}) \right]}_{\text{loss-augmented decoding}} - s(x,y;w)$$

a) <u>cartoon</u>:



$\Rightarrow f^{SVM}(x,y;w) \approx 0$

b) $\boxed{f^{SVM}(x,y;w) \geq \ell(y, h_w(x))}$

why? $f(x,y;w) = \max_{\tilde{y}} [s(\tilde{y}) + \ell(\tilde{y})] - s(y)$ let $\hat{y} = \arg\max_{\tilde{y}} s(\tilde{y})$

$\geq s(\hat{y}) + \ell(\hat{y}) - s(y)$

if $y \in \mathcal{Y}(x)$ $\Rightarrow$ $s(\hat{y}) \geq s(y)$

$\geq \ell(\hat{y}) = \ell(y, \underbrace{h_w(x)}_{\hat{y}})$ //

<u>binary case</u>: structured hinge loss reduces to binary SVM loss

$y \in \{-1,+1\}$ $w = \begin{pmatrix} w_+ \\ w_- \end{pmatrix}$ $\varphi(x,+1) = \begin{pmatrix} \varphi(x) \\ 0 \end{pmatrix}$

$h_w(x) = \arg\max \{ \langle w_+, \varphi(x)\rangle, \langle w_-, \varphi(x)\rangle \}$

predict $+1$ if $\langle w_+, \varphi(x)\rangle \geq \langle w_-, \varphi(x)\rangle$

$\Leftrightarrow \langle \underbrace{w_+ - w_-}, \varphi(x)\rangle \geq 0$

$\tilde{w} \triangleq w_+ - w_-$ $h_w(x) = \mathrm{sgn}(\langle \tilde{w}, \varphi(x)\rangle)$

(see notes to show)

(recopied from 2020 notes -- derivations for binary SVM:)

structured hinge loss

$$\tilde{w} = w_+ - w_- \qquad h_w(x) = \mathrm{sgn}(\langle \tilde{w}, x\rangle)$$

$$\ell^{SVM}(x,y;w) \triangleq \max\{\langle w_+, x\rangle + \underbrace{\ell(y,+)}_{\mathbb{1}\{y\neq +1\}},\ \langle w_-, x\rangle + \underbrace{\ell(y,-)}_{1-\mathbb{1}\{y\neq +1\}}\} - \langle w_y, x\rangle$$

$w_+ = \tilde{w} + w_-$

$$= \max\{\langle \tilde{w}, x\rangle + \langle w_-, x\rangle + \mathbb{1}\{y\neq +1\},\ \langle w_-, x\rangle + 1 - \mathbb{1}\{y\neq +1\}\} - \langle w_y, x\rangle$$

$$= \max\{\langle \tilde{w}, x\rangle + \mathbb{1},\ 1-\mathbb{1}\} + \langle w_-, x\rangle - \langle w_y, x\rangle$$

case $y=+1$: $\max\{\langle \tilde{w}, x\rangle, 1\} - \langle \tilde{w}, x\rangle = [1 - \overset{+1}{y}\langle \tilde{w}, x\rangle]_+$

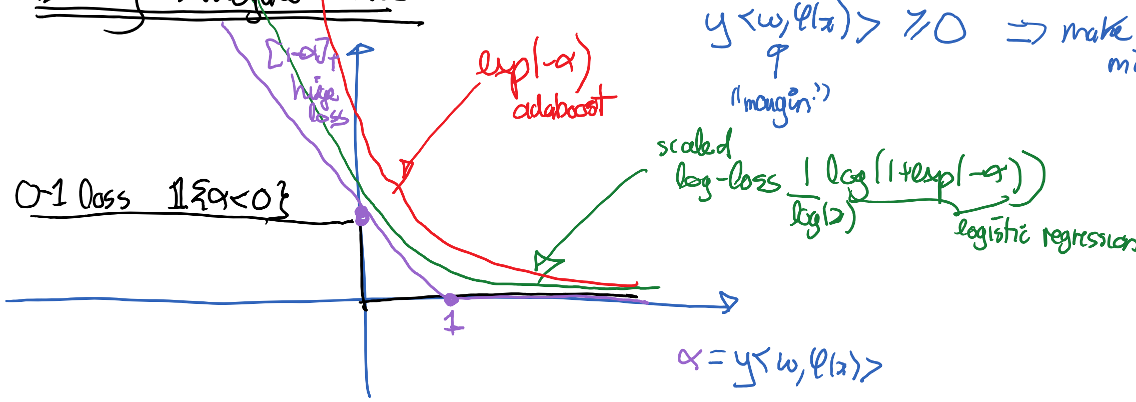
case $y=-1$: $\max\{\langle \tilde{w}, x\rangle + 1, 0\} + 0 = [1 - y\langle \tilde{w}, x\rangle]_+$

overall: $[1 - y\langle \tilde{w}, x\rangle]_+$

$$\ell^{SVM}(x,y;w) = [1 - y\langle \tilde{w}, \varphi(x)\rangle]_+ \quad \text{where } \tilde{w} \triangleq w_+ - w_-$$

ie. structured hinge loss reduces to binary SVM hinge loss when using $\ell(y,y') = \mathbb{1}\{y\neq y'\}$ and $\mathcal{Y} = \{-1,+1\}$

Binary surrogate losses



$y\langle w, \varphi(x)\rangle \geq 0 \Rightarrow$ make no mistake
("margin")

[Bartlett & al. 2006] $\rightarrow$ showed all these methods are consistent

theory basics

decision theory setup

estimate $h_w: \mathcal{X} \rightarrow \mathcal{Y}$

generalization error $= L_P(w) \triangleq \mathbb{E}_{(x,y)\sim P}[\ell(y, h_w(x))]$ (task loss)

ultimate goal is to find $w^* = \arg\min L_P(w)$

ultimate goal is to find $w^* = \arg\min_{w \in \mathcal{W}} L_P(w)$

problem: do not know P ("true" distribution on (x, y))

suppose $\underbrace{(x^{(i)}, y^{(i)})_{i=1}^n}_{\triangleq D_n \text{ training dataset}} \overset{iid}{\sim} P$ $\rightsquigarrow$ we could look at
$$\hat{L}_n(w) = \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, h_w(x^{(i)}))$$

from statistics / prob. theory

$\hat{L}_n(w) \xrightarrow[n\to\infty]{a.s.} L_P(w)$ for each fixed w (pointwise)

(LLN)

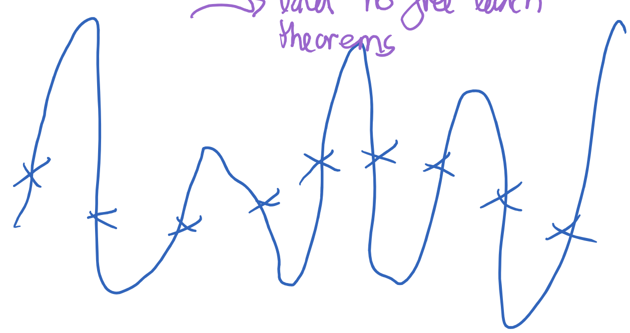
this is weaker than $\sup_w |\hat{L}_n(w) - L_P(w)| \xrightarrow{n\to\infty} 0$

note: minimizing the training error gives no guarantee on $L_P(\hat{w}_n)$ in general!

$\rightarrow$ later no free lunch theorems

e.g. polynomial regression

for n points, can get zero training error with poly of degree $n-1$



$\Rightarrow$ overfitting

in learning theory: study properties of learning alg. $A : D_n \to \mathcal{W}$

in particular, what can we say about $L_P(A(D_n))$

different approaches:

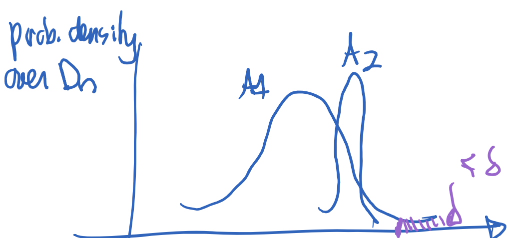
a) "frequentist risk" $R^F_{P,n}(A) \triangleq \mathbb{E}_{D_n \sim P^{\otimes n}} [L_P(A(D_n))]$ — D_n is random

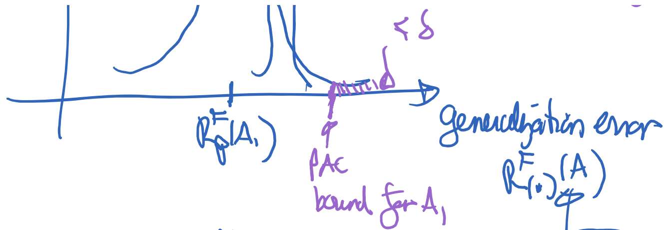
b) PAC framework ("probably approximately correct")

$P\{L_P(A(D_n)) > \text{some bound}\} \leq \delta$ "tail bound"

i.e. $L_P(A(D_n)) \leq$ some bound with prob. $\geq 1-\delta$

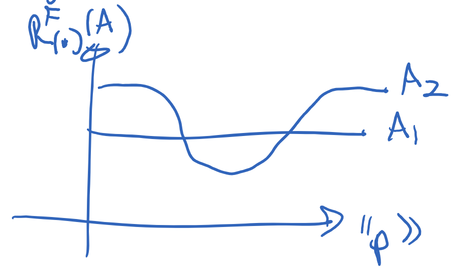
generalization error bound





issue with R^F_p → depends on P

"risk profiles"



weighted frequentist risk $\mathbb{E}_{\theta \sim \pi(\theta)}[R^F_{p_\theta}(A)]$

c) "Bayesian posterior risk"

$$R^{post}(\omega \mid D_n) \triangleq \mathbb{E}_{\theta \sim p(\theta \mid D_n)}[L_{p_\theta}(\omega)]$$

- prior $p(\theta)$ over dist.
- observation model $p(D_n \mid \theta)$ ⇒ posterior $p(\theta \mid D_n)$

Bayesian estimate $\hat{\omega}_n^{Bayes} = \arg\min_{\omega} R^{post}(\omega \mid D_n)$

$A^{Bayesian}$ is optimal for weighted frequentist risk using $\pi(\theta) = p(\theta)$

14h53

no free lunch?

frequentist risk analysis learning alg. A

let $\mathcal{P}$ be a set of distributions on $\mathcal{X} \times \mathcal{Y}$

sample complexity of A with respect to $\mathcal{P}$

is the smallest $n(\mathcal{P}, A, \varepsilon)$ s.t. $\forall n \geq n(\mathcal{P}, A, \varepsilon)$

we have $\sup_{p \in \mathcal{P}} [R^F_p(A;n) - L_p(h^*_p)] \leq \varepsilon$

"uniform result"

$h^*_p = \arg\min_{h: \mathcal{X} \to \mathcal{Y}} L_p(h)$

terminology:
- A is consistent for dist. p

  if $\lim_{n \to \infty} R^F_p(A;n) - L_p(h^*_p) = 0$
- A is uniformly consistent for a family $\mathcal{P}$

  if $\lim_{n \to \infty} \left[\sup_{p \in \mathcal{P}} [R^F_p(A;n) - L_p(h^*_p)] \right] = 0$

Binary classification $\mathcal{Y} = \{-1, +1\}$

I) if $\mathcal{X}$ is finite ; then the "voting procedure" (assign the most frequent label to an input x)

is uniformly and universally consistent

↳ i.e. $\mathcal{P}$ is all dist. on $\mathcal{X} \times \mathcal{Y}$

with (universal) sample complexity

$$n(\mathcal{P}, \varepsilon, A_{voting}) \lesssim \frac{|\mathcal{X}|}{\varepsilon^2}$$ (free lunch!!)

II) if $\mathcal{X}$ is infinite

no free lunch theorem I (for binary with & the 0-1 loss)

for any n and any learning alg. A

then $\sup\limits_{P \text{ all dist.}} \left[R_P^F(A; n) - L_P(h_P^*) \right] \geq \frac{1}{2}$

i.e. $\exists$ always a dist. $P_{A,n}$ s.t. A is worse than random predictor for $P_{A,n}$!

NFL II: [Thm. 7.2 in Devroye & al. 1996]

let ε_n be any non-increasing seq. converging to 0 + $(\varepsilon_1 \leq \frac{1}{16})$

for A, then $\exists P_A$ s.t.

$$\left[R_{P_A}^F(A; n) - L_{P_A}(h_{P_A}^*) \right] \geq \varepsilon_n \; \forall n$$

(ε_n could be arbitrarily slow e.g. $\frac{1}{\lg(\lg(\lg(\cdots)(n)))}$ 1023 ...)

consequence: we need assumptions on $\mathcal{P}$ to say anything useful

Occam's generalization error bound

- binary class. & 0-1 loss
- consider $\mathcal{W}$ to be a countable set

let's define a prior over $\mathcal{W}$: $\pi(w)$ i.e. $\sum\limits_{w \in \mathcal{W}} \pi(w) = 1$ $\pi(w) \geq 0 \; \forall w$

$|w|_\pi$ = "description length" of $w \triangleq \log_2 \frac{1}{\pi(w)}$ $\gg$ $\sum\limits_w 2^{-|w|_\pi} \leq 1$

"Kraft's inequality"

Occam's bound

$\pi(w)$ "Kraft's inequality"

Occam's bound

for any fixed P; with prob. $\geq 1-\delta$ over training set $D_n \sim P^{\otimes n}$

$$\forall w \in W \quad L_P(w) \leq \hat{L}_P(w) + \frac{1}{\sqrt{2n}}\, \Omega_\pi(w;\delta)$$

where $\Omega_\pi(w;\delta) \triangleq \sqrt{(\ln 2)\underbrace{|w|_\pi}_{\text{complexity measure}} + \ln\frac{1}{\delta}}$