Lecture 3 - scribbles

Tuesday, January 30, 2018
14:36

today: - structured hinge loss
- theory binary classification

III) structured hinge loss

$$f(x,y;w) = \max_{\tilde{y} \in \mathcal{Y}(x)} [s(x,\tilde{y};w) + \ell(y,\tilde{y})] - s(x,y;w)$$

"loss-augmented decoding"

carbon $\geq \ell(y,\tilde{y}_{max})$ $s(y)$ $s(\tilde{y}_{next})$ $\Rightarrow f(x,y;w) = 0$

b) $f(x,y;w) \geq \ell(y, h_w(x))$

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

$\geq s(\hat{y}) + \ell(\hat{y}) - s(y)$ using $\hat{y} = \operatorname{argmax}_{\tilde{y} \in \mathcal{Y}(x)} s(\tilde{y}) = h_w(x)$

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

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

binary case:

$y \in \{-1, +1\}$ $\quad w = \begin{pmatrix} w_+ \\ w_- \end{pmatrix}$ $\quad$ prediction: $h_w(x) = \operatorname{argmax} \{\langle w_+, x\rangle, \langle w_-, x\rangle\}$

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

$\Leftrightarrow \langle w_+ - w_-, x\rangle \geq 0$

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

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

$$\delta^{SVM}(x, y; w) = \max\{\langle w_+, x\rangle + \ell(y, +), \langle w_-, x\rangle + \ell(y, -)\} - \langle w_y, x\rangle$$

$\ell(y,+) = \mathbb{1}\{y \neq +1\}$ $\quad$ $\ell(y,-) = \mathbb{1}\{y \neq -1\}$

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

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

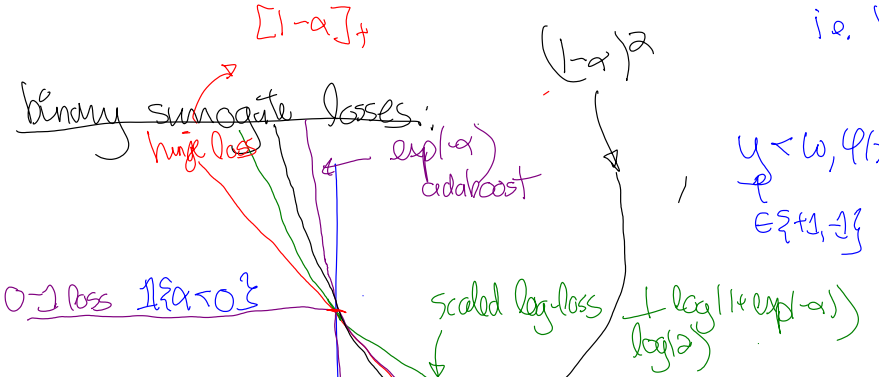
$y = +1$: $= \max\{\langle \tilde{w}, x\rangle, 1\} - \langle \tilde{w}, x\rangle = \max\{0, 1 - y\langle \tilde{w}, x\rangle\} = [1 - y\langle \tilde{w}, x\rangle]_+$

$y = -1$: $\max\{\langle \tilde{w}, x\rangle + 1, 0\} + 0 = \max\{0, 1 - y\langle \tilde{w}, x\rangle\} = \ldots$

$\delta^{\text{structured hinge}}(x, y; w) = [1 - y\langle \tilde{w}, x\rangle]_+$ $\Rightarrow$ that structure hinge loss reduces to standard binary SVM hinge loss when using $\ell(y, y') = \mathbb{1}\{y \neq y'\}$ and $\mathcal{Y} = \{-1, +1\}$

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

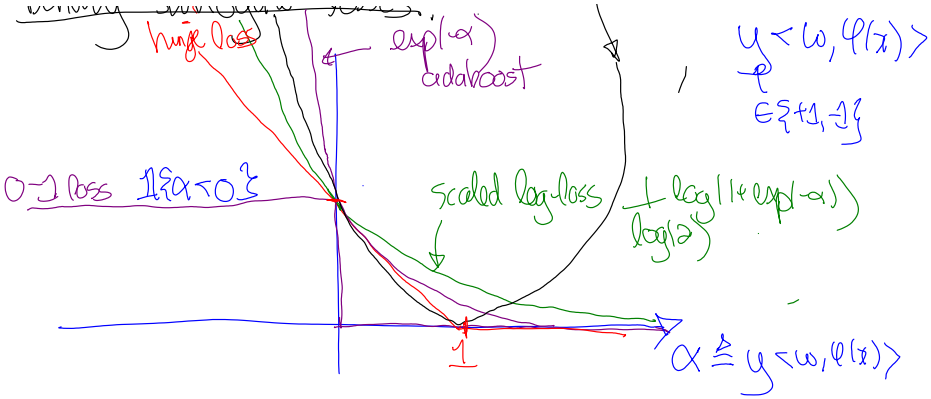
i.e. binary SVM is special case of structured SVM



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

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

log-loss: $\log(1 + \exp(-\alpha))$



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

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

log-loss: $\log(1+\exp(-\alpha))$

[logistic regression $\log(1+\exp(-y\langle w, \varphi(x)\rangle))$]

[Bartlett et al. 2006] → showed all these methods are consistent

Question on scaling the margin:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\nu - y_i\langle w, \varphi(x_i)\rangle\right]_+ + \lambda\frac{\|w\|^2}{2}$$

$$\nu\left[\frac{1}{n}\sum_{i=1}^{n}\left[1 - y_i\langle \underbrace{w}_{\tilde{w}}, \varphi(x_i)\rangle\right]_+ + \frac{\lambda}{\nu}\frac{\|w\|^2}{2}\right] + (\lambda\nu)\frac{\|\tilde{w}\|^2}{2}$$

i.e. is equivalent to just rescale λ ...

Theory basics

decision theory setup

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

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

ultimate goal is find $w^* = \operatorname{argmin}_{w \in W} L_p(w)$

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

suppose $\underbrace{(x^{(i)}, y^{(i)})_{i=1}^n}_{\triangleq D_n} \overset{iid}{\sim} P$ training data

learning algorithm $\hat{w}_n = A(D_n)$ ← algorithm

→ we could look at $\hat{L}_n(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h_w(x^{(i)}))$

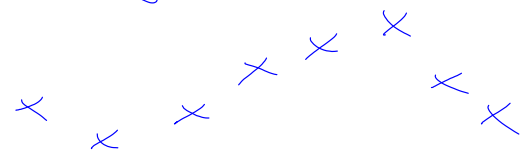
from statistics/prob

$\hat{L}_n(w) \overset{a.s.}{\to} L_p(w) \quad \forall w$ (LLN)

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

*note: minimizing training error gives no guarantee in general

e.g. linear regression



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

⇒ overfitting

in learning theory, want to study properties of learning algorithm

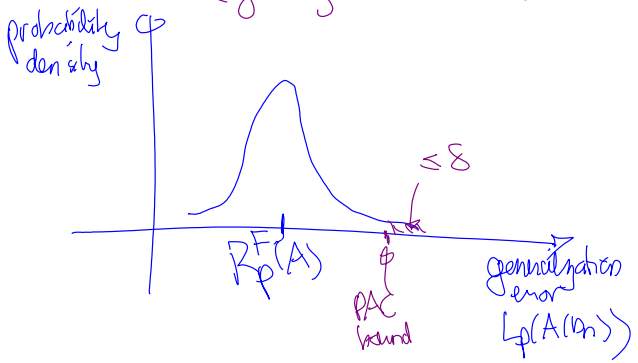
in particular, what can we say about $L_p(\underbrace{A(D_n)}_{\hat{w}_n})$?

different approaches: "frequentist risk" $R^F_p(A) \triangleq \mathbb{E}_{D_n \sim p^n}[L_p(A(D_n))]$ (D_n is random)

PAC framework
"probably approximately correct"

$$P\{L_p(A(D_n)) > \text{some bound}\} \leq \delta$$

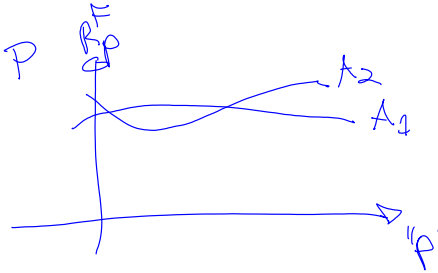
i.e. $L_p(A(D_n)) \leq$ some bound with prob $\geq 1-\delta$
← generalization error bound



$$\mathbb{E}X = \underbrace{\mathbb{E}X\mathbb{I}\{X \leq B\}}_{B\cdot(1-\delta)} + \underbrace{\mathbb{E}X\mathbb{I}\{X > B\}}_{\leq B_X \cdot \delta}$$

↑ tail bound limit

Issue with R_p^F → depends on P



weighted frequentist risk: $\mathbb{E}_{\theta\sim p(\theta)}[R_{p_\theta}^F(A)]$

"Bayesian posterior risk" $R^{post}(w|D_n) = \mathbb{E}_{\theta\sim p(\theta|D_n)}[L_{p_\theta}(w)]$

prior $p(\theta)$ over distributions
$p(D_n|\theta)$
⇒ posterior $p(\theta|D_n)$

Bayesian estimate $\hat{w}_n = \underset{w}{\text{argmin}}\, R^{post}(w|D_n)$

A Bayesian is optimal for weighted frequentist risk using $p(\theta)$

No free lunch

frequentist risk analysis of learning algorithm A

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

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 $\sup_{P \in \mathcal{P}} \left[R^F_P(A) - L_P(h^*_P) \right] < \varepsilon$

"uniform regret"

$h^*_P = \operatorname{argmin}_{\text{all } h: X \to Y} L_P(h)$

terminology:

- A is consistent for dist. P

  if $\lim_{n \to \infty} \underbrace{R^F_P(A; n)}_{\mathbb{E}_{D_n}[L_P(A(D_n))]} - L_P(h^*_P) = 0$

- A is uniformly consistent for family $\mathcal{P}$

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

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

I) if X is finite; then the "voting procedure" (assign the most frequent label to a input x)

is uniformly and universally consistent

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

with (universal) sample complexity $n(\Theta, \varepsilon, A_{voting}) \leq \frac{|X|}{\varepsilon^2}$ (free lunch? !!)

II) If X is infinite

no free lunch theorem:

(for binary classification with & the 0-1 loss)

for any n and any learning algorithm A

then $\sup_P [R_P^F(A;n) - L_P(h_P^*)] \geq \frac{1}{2}$

i.e. $\exists$ always a distribution P s.t. your algorithm A is not doing better than random prediction (?)

NFT II)

[thm. 7.2 in Devroye et al. 1996]

let ε_n be any non-increasing sequence converging to zero (could be an arbitrarily slow) and any alg. A

then $\exists P$ s.t. $R_P^F(A;n) - L_P(h_P^*) \geq \varepsilon_n \ \forall n$

~~(X)(X)~~ consequence: we need assumptions on P to say anything

Occam's generalization error bound

- binary classification and 0-1 loss
- consider W to be a countable set

Let's define a prior probability over W: $\pi(w)$ i.e. $\sum_{w \in W} \pi(w) = 1 \quad \pi(w) \geq 0$

$|w|_\pi$ = "description length" of w

$\triangleq \log_2 \frac{1}{\pi(w)}$

Occam's bound

For any P, with probability $\geq 1-\delta$ over training set $D_n \sim P^{\otimes n}$

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

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

(*) bound is only useful for distribution P s.t. $|w_P^*|_\pi$ is small

$w_P^* = \operatorname{argmin}_{w \in \mathcal{W}} L_P(w)$

$|w|_\pi = \log_2 \frac{1}{\pi(w)}$ if $\pi(w) \propto \exp(\gamma \|w\|^2)$ then $|w|_\pi = \|w\|^2 + \text{cst.}$

proof: use 3 things

1) Chernoff bound (concentration inequality) $\quad P\{D_n : \hat{L}_n(w) \leq L(w) - \varepsilon\} \leq \exp(-2n\varepsilon^2) \quad \forall \varepsilon \geq 0$
2) union bound $\quad P\{\exists x \text{ s.t. } prop(x) \text{ is true}\} \leq \sum_x P\{prop(x) \text{ is true}\}$
3) "Kraft's inequality" $\quad \sum_w 2^{-|w|_\pi} \leq 1$

we say w is bad if bound fails

we say w is bad if bound fails

$$\text{bad}(w) = \mathbb{1}\{ L(w) > \hat{L}_n(w) + \underbrace{\frac{1}{\sqrt{2n}} \Omega_{\pi}(w;\delta)}_{\varepsilon_n} \}$$

using Chernoff, $\hat{L}_n(w) \leq L(w) - \varepsilon_n$

$$P\{\text{bad}(w)\} \leq \exp(-2n\varepsilon_n^2) = \exp\left(-2n \frac{1}{2n}\left((\ln 2)|w|_\pi + \ln\frac{1}{\delta}\right)\right)$$

$$= \delta 2^{-|w|_\pi}$$

using union bound

$$P\{\exists w : \text{bad}(w)\} \leq \sum_w P\{\text{bad}(w)\} \leq \sum_w \delta 2^{-|w|_\pi} \leq \delta$$

//

example of complexity: $|w|_\pi = \|w\|_2^2$ ⇝ get ℓ_2-regularization?

Surrogate loss:

NP-hard to minimize $\hat{L}_n(w)$; replace with $\hat{\mathcal{L}}_n(w)$ which is "surrogate" e.g. hinge loss