PS4 due next week

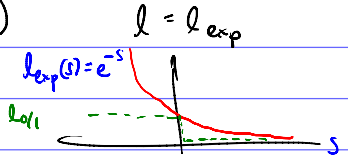
Last time: Boosting as an optimization alg.
Linear learning paradigm

Today: More Linear learning paradigm
- Convex risks
- SGD

Linear learning paradigm

1. Define feature map $\phi : \mathcal{X} \to \mathbb{R}^d$
2. Learn $h_w \in HLTF_{\mathbb{R}^d}$ from $S = ((\phi(x_1), y_1), \ldots, (\phi(x_n), y_n)) \in (\mathbb{R}^d \times \{-1, 1\})^n$ by (approx.) minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \ell(y_i \langle \phi(x_i), w \rangle)$$



3. Return $\hat{h}(x) = \text{sign}(\langle \phi(x), w \rangle)$

If $\ell = \ell_{0/1}$, step 2 could be NP-hard

This is why we don't use $\ell_{0/1}$

Notation: For $f : \mathcal{X} \to \mathbb{R}$, define

$$R(f) = E[\ell_{0/1}(y f(x))] \quad \text{where } (x, y) \sim P$$

↑ (classification) risk = error rate of sign of f

$$R^* = \inf_{f : \mathcal{X} \to \mathbb{R}} R(f) \quad \text{– optimal risk}$$

$$R_{exp}(f) = E[\ell_{exp}(y f(x))]$$

↑ exp. loss risk

$$R^*_{exp} = \inf_{f : \mathcal{X} \to \mathbb{R}} R_{exp}(f) \quad \text{– optimal exp loss risk}$$

Q: If $R_{exp}(f) - R^*_{exp}$ is small, then what do we know about $R(f) - R^*$?

R^* vs R^*_{exp}

$$R^* = \inf_f E[\ell_{0/1}(y f(x))]$$

$$E[\ell_{0/1}(y f(x))] = E[E[\ell_{0/1}(y f(x)) | x]]$$

Define $p(x) = P_r(y=1|x)$

$$E[\ell_{0/1}(y f(x)) | x] = p(x) \cdot 1\{f(x) \le 0\} + (1-p(x)) 1\{f(x) > 0\}$$

To minimize this ↷

$$f^*(x) = \begin{cases} \text{positive} & \text{if } p(x) > 1-p(x) \\ \text{negative} & \text{if } p(x) < 1-p(x) \end{cases}$$

e.g. $f^*(x) = \text{sign}\left(p(x) - \frac{1}{2}\right)$

$$R^* = R(f^*) = E[\min\{p(x), 1-p(x)\}]$$

$$R^*_{exp} = \inf_f E[e^{-y f(x)}]$$

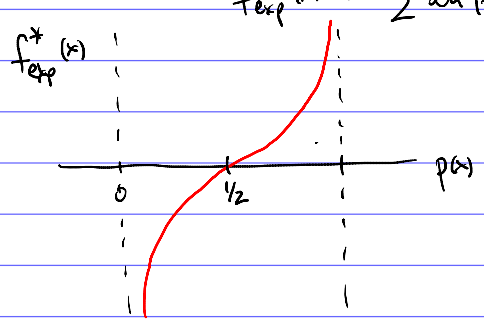
$$E[e^{-y f(x)} | x] = p(x) e^{-f(x)} + (1-p(x)) e^{f(x)}$$

$$\frac{\partial\{p(x) e^{-f(x)} + (1-p(x)) e^{f(x)}\}}{\partial f(x)} = -p(x) e^{-f(x)} + (1-p(x)) e^{f(x)} = 0$$

solve for $f(x)$

given

$$f^*_{exp}(x) = \frac{1}{2} \ln \frac{p(x)}{1-p(x)}$$

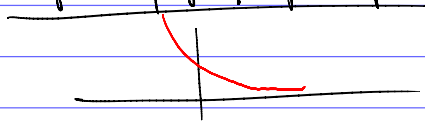


$$R^*_{exp} = R_{exp}(f^*_{exp}) = \cdots = 2E[\sqrt{p(x)(1-p(x))}]$$

Thm $R(f) - R^* \le \sqrt{2(R_{exp}(f) - R^*_{exp})}$

Pf Lots of algebra ☹

Important property of ℓ_{exp}: convex



$f(x) = \langle \phi(x), w \rangle$

Claim: $\ell_{exp}(y \langle \phi(x), w \rangle)$ is a convex fn of w.

Pf: ① $\ell_{exp}(s)$ is convex s

② $y \langle \phi(x), w \rangle$ is affine in w

③ composition of convex fn w/ affine fn is convex

$\therefore$ $\ell_{exp}(y \langle \phi(x), w \rangle)$ is convex fn of w. ☒

Other properties of convex fns:

① scaling by non-negative value $\longrightarrow$ fn stays convex

② sums / integrals

$\therefore$ $\frac{1}{n} \sum_{i=1}^{n} \ell_{exp}(y_i \langle \phi(x_i), w \rangle)$ } both convex in w.

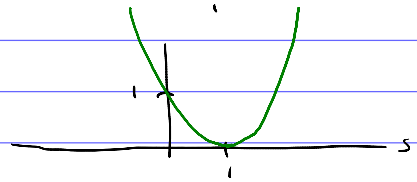
& $E[\ell_{exp}(y \langle \phi(x), w \rangle)]$

Many other convex 'surrogates' for ℓ_{01} exist

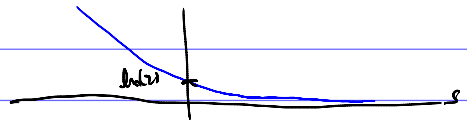
$\ell_{hinge}(s) = \max\{0, 1-s\}$



$\ell_{sq}(s) = (1-s)^2$



$\ell_{logistic}(s) = \ln(1+e^{-s})$



for all of these (and more), $\exists\ \psi^{-1}$

$$R(f) - R^* \leq \psi^{-1}(R_\ell(f) - R_\ell^*)$$

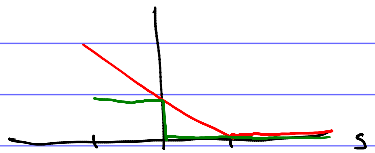
s.t. $R_\ell(f) \to R_\ell^* \implies R(f) \to R^*$

[Zhang, 2004, Bartlett, Jordan, McAuliffe, 2006]

Particularly nice : ℓ_{hinge}

- R^*_{hinge} is achieved by $f^*_{hinge}(x) = \text{sign}(p(x) - \frac{1}{2})$ ← also achieves smallest $R(f)$
- $R(f) - R^* \leq R_{hinge}(f) - R^*_{hinge} = R_{hinge}(f) - 2R^*$
- Can relate R^* & R^*_{hinge} easily:

$$R^*_{hinge} = 2R^*$$



- Slight improvement:

$$\inf_{t \in [-1,1]} R(f-t) - R^* \leq \frac{1}{2}\left(R_{hinge}(f) - R^*_{hinge}\right)$$

Q: Suppose we consider fns of the form

$$f(x) = \langle \phi(x), w \rangle$$

- How to minimize R_{hinge} wrt w? Convex opt.
- How to choose ϕ to ensure $\exists\, w^* \in \mathbb{R}^d$ s.t. $f^*(x) \approx \langle \phi(x), w \rangle$? Approx theory.

SGD

$J: \mathbb{R}^d \to \mathbb{R}$ convex

$$J(w) \geq J(w_0) + \langle \nabla J(w_0), w - w_0 \rangle \qquad \forall\, w, w_0 \in \mathbb{R}^d$$

$$J(w) = E[L(w, z)] \qquad \text{where } z \sim P$$

$L(\cdot, z)$ is convex

SGD Start $w_1 \in \mathbb{R}^d$ (eg. $w_1 = 0$)

For $t = 1, 2, \ldots, T$:

Get $z_t \sim P$ (indpt of everything seen so far)

Compute $\nabla L(w_t; z_t)$ ← $\nabla L(\cdot; z_t)$ evaluated at w_t

Update: $w_{t+1} = w_t - \eta \nabla L(w_t; z_t)$

(Return w_{T+1})

Conditional on $w_1, \ldots, w_t$

$$E[\nabla L(w_t, z_t) \mid w_1, \ldots, w_t] = \nabla J(w_t)$$
$$= \nabla E[L(w_t, z_t) \mid w_1, \ldots, w_t]$$
$$= \nabla J(w_t)$$

Thm For any $u \in \mathbb{R}^d$,

$$E\left[\frac{1}{T}\sum_{t=1}^{T} J(w_t) - J(u)\right] \leq \frac{\|w_1 - u\|^2}{2\eta T} + \frac{\eta G^2}{2}$$

where $G \geq \|\nabla L(w, z)\| \quad \forall\, w, z$

eg. $L(w, z) = \max\{0, 1 - y\langle \phi(x), w\rangle\}$, $\quad z = (x, y)$

then $\|\nabla L(w, z)\| = \|\phi(x)\|$

Pf: For any $u \in \mathbb{R}^d$,

$$\frac{1}{2}\|w_{t+1} - u\|^2 - \frac{1}{2}\|w_t - u\|^2$$

$$= \frac{1}{2}\|w_t - u - \eta \nabla L(w_t, z_t)\|^2 - \frac{1}{2}\|w_t - u\|^2$$

$$= \langle w_t - u, -\eta \nabla L(w_t, z_t)\rangle + \frac{\eta^2}{2}\|\nabla L(w_t, z_t)\|^2$$

$$\leq -\eta \langle w_t - u, \nabla L(w_t, z_t)\rangle + \frac{\eta^2}{2} G^2$$

Avg over $t = 1, \ldots, T$:

$$\frac{1}{T}\sum_{t=1}^{T} \langle w_t - u, \nabla L(w_t, z_t)\rangle \leq \frac{1}{T} \cdot \frac{\|w_1 - u\|^2}{2\eta} - \cancel{\frac{1}{T}\frac{\|w_{T+1} - u\|^2}{2\eta}}^{\leq 0} + \frac{\eta G^2}{2}$$

$$E[\quad \text{"} \quad]$$
$$= \frac{1}{T}\sum_{t=1}^{T} E[\langle w_t - u, \nabla J(w_t)\rangle]$$

$$\geq \frac{1}{T}\sum_{t=1}^{T} E[J(w_t) - J(u)] \qquad \text{by convexity}$$

$$= E\left[\frac{1}{T}\sum_{t=1}^{T} J(w_t) - J(u)\right]$$

$\boxtimes$