Lecture 14

* Time slots for presentation review have been posted on the project homepage.

RECAP

Boosting

Weak-learning: An algorithm A is a weak learner with **edge / advantage** γ for class $\mathcal{C}$ if: for any dist. D & any target $c \in \mathcal{C}$, given access to $EX(c, D)$, w.p. $(1-\delta)$ A produces a hypothesis with $\text{error}(h, c, D) \leq \frac{1}{2} - \gamma$.

If A runs in time $\text{poly}(d, \frac{1}{\delta})$ & $\gamma > \frac{1}{\text{poly}(d)}$

then $\mathcal{C}$ is efficiently weakly-PAC learnable

Thm If $\mathcal{C}$ is weakly-PAC learnable (efficiently), then $\mathcal{C}$ is PAC-learnable (efficiently.

TODAY

Proof AdaBoost (Freund & Schapire)

(Godel prize '03, very influential in practice)

Training set: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

$x_i \in \mathcal{X}$

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

By realizability,

$\exists\ c \in \mathcal{C}$ s.t. $y_i = c(x_i)\ \forall i$.

We assume $\exists$ weak learning algorithm (WL) for $\mathcal{C}$.

AdaBoost (Freund & Schapire)

1. $D_1(i) = \frac{1}{n} \quad \forall i \in [n]$
2. For $t = 1, \ldots, T$
3. Use WL with dist D_t to get h_t
4. Let $\varepsilon_t = \mathbb{P}_{x \sim D_t}[h_t(x) \neq y]$ // $\varepsilon_t \leq \frac{1}{2} - \gamma$ (w.p. $1-\delta$)
5. Choose $\alpha_t = \frac{1}{2} \log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
6. $$D_{t+1}(i) = \frac{D_t(i)}{Z_{t+1}} \cdot \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$
$$= \frac{D_t(i)}{Z_{t+1}} \cdot e^{(-\alpha_t h_t(x_i) y_i)}$$

(where Z_{t+1} is normalizing constant)

7. Output $\text{sign}\left(H(x)\right)$ where $H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

Note:

1) We can assume - $\varepsilon_t \leq \frac{1}{2} - \gamma$ (union bound)

2) Can emulate $E_t(c, D_t)$, because D_t has finite support, just reweigh.

Aside: AdaBoost fits into the "Multiplicative weight updates" framework. General framework with many algorithmic applications.

Lemma For $T = \frac{1}{2\gamma^2} \log(2n)$, training error is 0.

Proof

Training error $= \mathbb{P}_{D_1}\left[\text{sign}(H(x_i)) \neq y_i\right]$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(\text{sign}(H(x_i)) \neq y_i\right)$$

Note that

$$\mathbb{1}(\text{sign}(H(x)) \neq y) \leq e^{-yH(x)}$$

Define $H_t(x) = \sum_{s=t}^{T} \alpha_s h_s(x)$

$$H_t(x) = \alpha_t h_t(x) + H_{t+1}(x)$$

$$H_1(x) = H(x)$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(\text{sign}\left(H(x_i)\right) \neq y_i\right) \leq \sum_{i=1}^{n} D_1(i)\, e^{-y_i H_1(x_i)}$$

$$= \sum_{i=1}^{n} D_1(i)\, e^{-\alpha_1 y_i h_1(x_i) - y_i H_2(x_i)}$$

$$= z_2 \sum_{i=1}^{n} D_2(i)\, e^{-y_i H_2(x_i)}$$

$$= z_2 \sum_{i=1}^{n} D_2(i)\, e^{-\alpha_2 y_i h_2(x_i) - y_i H_3(x_i)}$$

$$= z_2 z_3 \sum_{i=1}^{n} D_3(i)\, e^{-y_i H_3(x_i)}$$

$$\vdots$$

$$= \prod_{t=2}^{T+1} z_t$$

What is z_{t+1}?

$$z_{t+1} = \sum_{i=1}^{n} D_t(i)\, e^{-\alpha_t y_i h_t(x_i)}$$

$$= \sum_{i : y_i = h_t(x_i)} D_t(i)\, e^{-\alpha_t} + \sum_{i : y_i \neq h_t(x_i)} D_t(i)\, e^{\alpha_t}$$

$$= (1 - \epsilon_t)\, e^{-\alpha_t} + \epsilon_t \cdot e^{\alpha_t}$$

Recall $\alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

$$\therefore \quad Z_{t+1} = (1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

$$= 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Define $\gamma_t = \frac{1}{2} - \epsilon_t$ Note that $\gamma_t \geq \gamma$

$$Z_{t+1} = 2\sqrt{\left(\tfrac{1}{2}-\gamma_t\right)\left(\tfrac{1}{2}+\gamma_t\right)}$$

$$= \sqrt{1-4\gamma_t^2}$$

$$\leq \left(e^{-4\gamma_t^2}\right)^{1/2}$$

$$= e^{-2\gamma_t^2}$$

$$\leq e^{-2\gamma^2}$$

$$\therefore \prod_{t=1}^{T+1} z_t \leq e^{-2T\gamma^2} < \frac{1}{2n}$$

if $T \geq \frac{1}{2\gamma^2}\log(2n)$.

What about test error?

Suppose that WL (weak learning algo.) always outputs a hypothesis from some class $\mathcal{H}$ whose VC-dim is d.

$$LC(\mathcal{H}, T) = \left\{ \text{sign}\left(\sum_{i=1}^{T} \alpha_i h_i(x) \right) : h_i \in \mathcal{H}, \alpha_i \in \mathbb{R} \right\}$$

Exercise: $VC\text{-}dim(LC(\mathcal{H}, T)) \leq c \cdot T \cdot d \cdot \log T$ for some constant c.

Hint: As in HW1, compute growth function & use Sauer's lemma.

∴ Due to VC Theorem,

if $n \geq \frac{c}{\varepsilon}\left(\log\left(\frac{1}{\delta}\right) + T\, d \log T \log\left(\frac{1}{\varepsilon}\right)\right)$

then the hypothesis produced by AdaBoost has error $\leq \varepsilon$, w.p. $1-\delta$.

Putting in the bound of T,

if $n \geq \frac{c}{\varepsilon}\left(\log\left(\frac{1}{\delta}\right) + \frac{1}{2\gamma^2}\log(2n)\, d \log\left(\frac{1}{2\gamma^2}\log(2n)\right) \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$

we will get error $\leq \varepsilon$ w.p. $1-\delta$.

This is satisfied for

$$n \geq \frac{c}{\varepsilon}\left(\frac{d}{\gamma^2}\,\mathrm{polylog}\left(d, \frac{1}{\gamma^2}, \frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right).$$

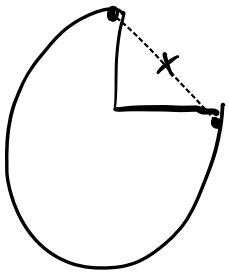
[$n \geq a \log n$ is satisfied for $n \geq a \log a$]

Convex Optimization

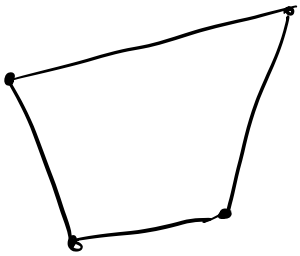
Basic properties,

Def (convex set) A set $C \subseteq \mathbb{R}^d$ is convex if $x, y \in C \Rightarrow tx + (1-t)y \in C$ for all $0 \leq t \leq 1$.

Non convex



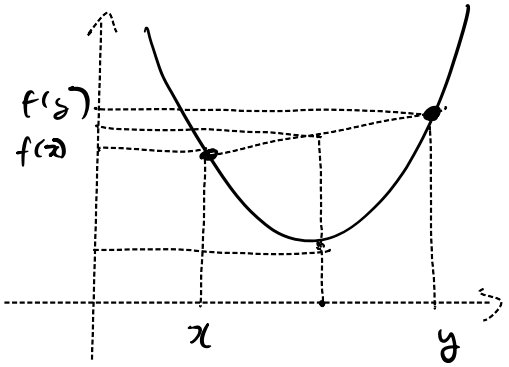
Convex

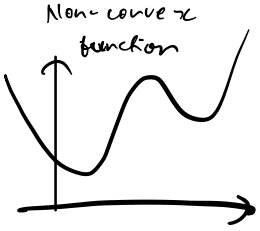


Def (convex functions) A function $f: \mathbb{R}^d \to \mathbb{R}$ is convex if domain $(f) \subseteq \mathbb{R}^d$ is convex &

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$$

$\forall\ t \in [0,1]$ & $x, y \in$ domain(f).





Lemma If f is differentiable, then f is convex if & only if domain (f) is convex & $f(y) - f(x) \geq \langle y-x, \nabla f(x) \rangle$

Proof Exercise, use definition of convexity.

Also,

$$f(y) - f(x) \leq \langle y-x, \nabla f(y) \rangle.$$

Corollary If f is convex & differentiable then $\nabla f(x) = 0$ implies x is a global minima of f (local minimum implies global minimum).

Def (strong convexity) A function f is λ-strongly convex if its domain is convex & $\forall\, x, y$

$$f(y) \geq f(x) + \langle y-x, \nabla f(x) \rangle + \frac{\lambda}{2} \|y-x\|^2.$$

E.g. $f(x) = \|x\|^2$

Lemma (Some properties of convex functions)

① If f is twice differentiable, then f is convex if & only if $\text{domain}(f)$ is convex & $\nabla^2 f(x) \succeq 0 \quad \forall \quad x \in \text{domain}(f)$.

(Recall $A \succeq 0 \iff x^T A x \geq 0 \quad \forall \; x$).

② If $f_i(x)$ is convex $\forall\, i \in [n]$ then $g(x) = \sum_{i=1}^{n} w_i f_i(x)$ is convex, where $w_i \geq 0$.

③ If $f_i(x)$ is convex $\forall\, i \in [n]$ then $g(x) = \max_{i \in [n]} f_i(x)$ is convex.

Convex Learning Problems

An optimization problem

$$\min_{x \in A} f(x)$$

is called a convex optimization problem if ① $f(x)$ is convex ② A is convex.

Convex optimization problems can be solved efficiently.

Recall the ERM problem of finding the ERM wrt some hypothesis $\mathcal{H}$ on some training set $S = (z_1, \dots, z_n)$ where $z_i = (x_i, y_i)$.

$$ERM_{\mathcal{H}}(S) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i)$$

Let $\mathcal{H} \subseteq \mathbb{R}^d$ parameterized by $w \in \mathcal{H}$

$$ERM_{\mathcal{H}}(S) = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$$

Lemma If ℓ is a convex loss (in terms of w), & $\mathcal{H}$ is convex, then $ERM_{\mathcal{H}}(S)$ is a convex optimization problem.

Proof Average of convex functions is convex..

Example (linear regression with squared loss)

Let $\mathcal{H} = \{x \mapsto \langle w, x \rangle, w \in \mathbb{R}^d\}$

$$\ell(h, (x, y)) = (h(x) - y)^2$$

$$\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$$

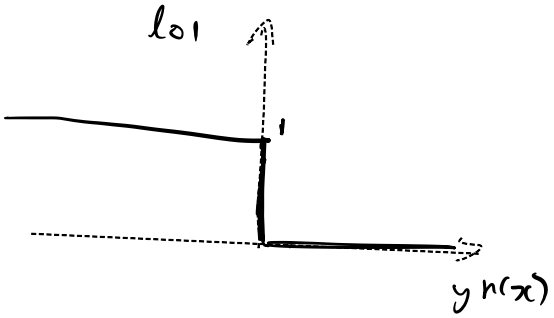
$$\mathcal{H} = \mathbb{R}^d \quad \text{(convex)}$$

Exercise $\ell(w, (x, y))$ is convex in terms of w.

$\therefore$ $ERM_{\mathcal{H}}(S)$ is a convex problem.

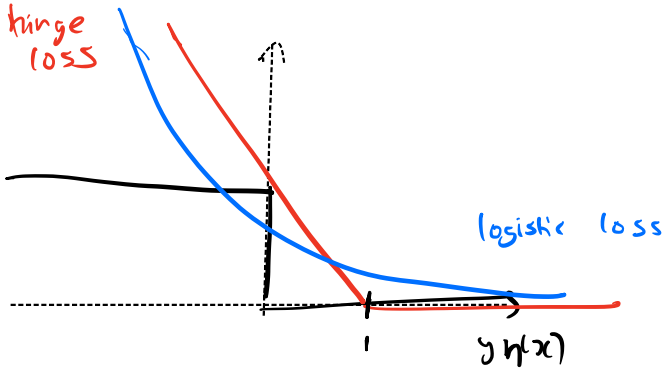
What about the 0/1 loss?

$$\ell_{0,1}(h, (x, y)) = \mathbb{1}(h(x) \neq y) = \mathbb{1}(y\,h(x) \leq 0)$$



Non-convex!

A common technique to handle a non-convex loss is to instead consider a convex surrogate.



hinge loss: $\ell(h, (x, y)) = \max(1 - y\,h(x), 0)$

logistic loss: $\ell(h, (x, y)) = \log\left(1 + e^{-y\,h(x)}\right)$