1

مسئله‌ی نامتناهی بُعدی داده شده را به صورت مسئله با بُعد محدود می‌توان بازنویسی کرد

تابع محدب $f(x)$ را می‌توان با piecewise linear convex function به صورت زیر تقریب زد :

$$f(x) = \max_{j=1\dots m} \{ a_j x + b_j \}$$

m تعداد piece هاست

$$a_1 \leqslant \cdots \leqslant a_m$$

$$\xrightarrow{\text{جاگذاری}} \min_{a_j, b_j} \sum_{i=1}^{n} \left(\max_{j=1\dots m} \{ a_j x + b_j \} - y_i \right)^2$$

$$\text{s.t.} \quad a_1 \leqslant \cdots \leqslant a_m$$

مسئله بهینه سازی با بُعد محدود به دست آمده را می‌توان توسط QP و یا gradient descent حل کرد

زیرا تابع هزینه quadratic است و قید ها نیز خطی هستند

بنابراین به دلیل محدب بودن توابع و محدود بودن n مسئله قابل یادگیری است

2

می‌دانیم که ریسک هر weak learner به تنهایی باید کمتر از $\frac{1}{2}$ باشد. ریسک h_t را در مرحله $t+1$ محاسبه می‌کنیم :

$$R_{D^{t+1}}(h_t) = \sum_{i=1}^{n} \underbrace{D^{t+1}(i)}_{\frac{D^t(i)\, e^{-\alpha_t y_i h_t(x_i)}}{Z_t}} \mathbb{1}[y_i \neq h_t(x_i)] = \sum_{y_i h_t(x_i) < 0} \frac{D^t(i)\, e^{\alpha_t}}{Z_t}$$

$$= \frac{e^{\alpha_t}}{Z_t} \underbrace{\sum_{i=1}^{n} D^t(i)\, \mathbb{1}[y_i \neq h_t(x_i)]}_{\varepsilon_t} \overset{\mathrm{I}}{=} \frac{\sqrt{\frac{1}{\varepsilon_t} - 1}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}}\, \varepsilon_t = \frac{\sqrt{\varepsilon_t - \varepsilon_t^2}}{2\sqrt{\varepsilon_t - \varepsilon_t^2}} = \frac{1}{2}$$

بنابراین h_t برای مرحله $t+1$ یادگیرنده پایه‌ی مناسبی نیست و h_{t+1} باید متفاوت از h_t باشد

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

$$= e^{\alpha_t} \underbrace{\sum_{i \in incorr} D^t(i)}_{\varepsilon_t} + e^{-\alpha_t} \underbrace{\sum_{i \in corr} D^t(i)}_{1-\varepsilon_t} \qquad \mathrm{I}$$

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

4

✻ فرض می‌کنیم x یک بردار با d المان غیرصفر است و w یک بردار با k المان غیر صفر است

الف) بدون فرض

$$\|x\|_\infty \|w\|_1 \approx \|w\|_1 \approx k$$

$$\|x\|_2 \|w\|_2 \approx \sqrt{d}\,\|w\|_2 \approx \sqrt{dk}$$

$\longrightarrow$ if $k<d$ (sparse w) $\longrightarrow$ L1 is better

if $k>d$ (dense w) $\longrightarrow$ L2 is better

در حالت کلی حرفی برای کران نمی‌توان زد ولی L1 به طور کلی جواب را sparse تر می‌کند

ب) sparse w

$$\|x\|_\infty \|w\|_1 \approx k$$

$$\|x\|_2 \|w\|_2 \approx \sqrt{dk}$$

$\xrightarrow{k<d}$ L1 is better

ج) dense w $\left(\|w\|_1 \approx \sqrt{d}\,\|w\|_2\right)$

$$\|x\|_\infty \|w\|_1 \approx \sqrt{d}\,\|w\|_2$$

$$\|x\|_2 \|w\|_2 \approx \sqrt{d}\,\|w\|_2$$

$\longrightarrow$ comparable

در این حالت نیز حرفی برای کران نمی‌توان زد و تقریبا برابر هستند

5.a

$$R_\ell(h) = \mathbb{E}_{(x,y)\sim \mathcal{D}}\left[e^{-y h(x)}\right] = \mathbb{E}_x \mathbb{E}_{y|x}\left[e^{-y h(x)}\right]$$

$$\overset{y\in\{-1,1\}}{=\!=\!=} \mathbb{E}_x\left[\eta(x) e^{-h(x)} + (1-\eta(x))\, e^{h(x)}\right] \geqslant \mathbb{E}_x\left[2\sqrt{\eta(x)\,(1-\eta(x))}\right]$$

حال اگر نشان دهیم کران پایین $R_\ell(h)$ به ازای چه h ای achieve می‌شود، آن h همان h_{exp} خواهد بود

$$\text{if } h(x) = \frac{1}{2}\lg\frac{\eta(x)}{1-\eta(x)} \longrightarrow R_\ell(h) = \mathbb{E}_x\Bigg[\underbrace{\eta(x)\,\frac{1}{e^{\lg\sqrt{\frac{\eta(x)}{1-\eta(x)}}}}}_{\sqrt{(1-\eta(x))\,\eta(x)}} + \underbrace{(1-\eta(x))\, e^{\lg\sqrt{\frac{\eta(x)}{1-\eta(x)}}}}_{\sqrt{\eta(x)(1-\eta(x))}}\Bigg]$$

$$= \mathbb{E}_x\left[2\sqrt{\eta(x)\cdot(1-\eta(x))}\right]$$

بنابراین، $h_{exp}(x) = \frac{1}{2}\lg\frac{\eta(x)}{1-\eta(x)}$ پاسخ بهینه نیز است

$\boxed{5.b}$

$$R(h) = \mathbb{E}_{(x,y)\sim D}\Big[\mathbb{1}\big[\operatorname{sign}(h(x)) \neq y\big]\Big] = \mathbb{E}_{x}\,\mathbb{E}_{y|x}\Big[\mathbb{1}\big[\operatorname{sign}(h(x)) \neq y\big]\Big]$$

$$\underset{y\in\{-1,1\}}{=}\ \mathbb{E}_{x}\Big[\eta(x)\,\mathbb{1}\big[h(x)<0\big] + (1-\eta(x))\,\mathbb{1}\big[h(x)\geq 0\big]\Big]$$

$$\geq \mathbb{E}_{x}\Big[\min\{\eta(x),\, 1-\eta(x)\}\Big]$$

چون $\mathbb{1}$ ها اشتراکی ندارند و در هر حالت فقط یکی از آنها انتخاب می‌شود
پس اگر min را برداریم همواره کوچکتر مساوی می‌شود

اگر h_{exp} بخش قبل را در نظر بگیریم

$$\longrightarrow R(h_{exp}) \overset{\mathrm{I}}{=} \mathbb{E}_{x}\Big[\eta(x)\,\mathbb{1}_{\eta(x)<\frac{1}{2}} + (1-\eta(x))\,\mathbb{1}_{\eta(x)\geq\frac{1}{2}}\Big]$$

$$\mathrm{I}:\quad h_{exp}(x) = \frac{1}{2}\lg\frac{\eta(x)}{1-\eta(x)} < 0 \longrightarrow \eta(x) < \frac{1}{2}$$

$$\qquad\qquad \sim \quad \sim \quad \sim \quad \geq 0 \longrightarrow \eta(x) \geq \frac{1}{2}$$

عبارت داخل $\mathbb{E}_x$ برابر $\min\{\eta(x), 1-\eta(x)\}$ است زیرا
اگر $\eta(x) < \frac{1}{2}$ باشد خودش انتخاب می‌شود و اگر
$\eta(x) \geq \frac{1}{2}$ باشد مکمل آن انتخاب می‌شود

بنابراین کمترین مقدار $R(h)$ که آنرا R^* می‌نامیم
همان $R(h_{exp})$ است

6

می دانیم که ریسک هر weak learner به تنهایی باید کمتر از $\frac{1}{2}$ باشد. ریسک h_t را در مرحله $t+1$ محاسبه می کنیم:

$$R_{D^{t+1}}(h_t) = \sum_{i=1}^{n} \underbrace{D^{t+1}(i)}_{\frac{D^t(i)\, e^{-\alpha_t y_i h_t(x_i)}}{Z_t}} \mathbb{1}\left[y_i \neq h_t(x_i)\right] = \sum_{y_i h_t(x_i) < 0} \frac{D^t(i)\, e^{\alpha_t}}{Z_t}$$

$$= \frac{e^{\alpha_t}}{Z_t} \underbrace{\sum_{i=1}^{n} D^t(i)\, \mathbb{1}\left[y_i \neq h_t(x_i)\right]}_{\varepsilon_t} \overset{\mathrm{I}}{=} \frac{\sqrt{\frac{1}{\varepsilon_t} - 1}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}}\, \varepsilon_t = \frac{\sqrt{\varepsilon_t - \varepsilon_t^2}}{2\sqrt{\varepsilon_t - \varepsilon_t^2}} = \frac{1}{2}$$

بنابراین h_t برای مرحله $t+1$ یادگیرنده پایه‌ی مناسبی نیست و h_{t+1} باید متفاوت از h_t باشد

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

$$= e^{\alpha_t} \underbrace{\sum_{i \in incorr} D^t(i)}_{\varepsilon_t} + e^{-\alpha_t} \underbrace{\sum_{i \in corr} D^t(i)}_{1-\varepsilon_t}$$

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

I

در برخی سوال ها با علی بابابیگ همفکری داشتم