نظریه یادگیری ماشین

نیم‌سال اول ۱۴۰۳-۱۴۰۴

مدرس: دکتر امیر نجفی

دانشکده‌ی مهندسی کامپیوتر

تمرین چهارم

مسئله‌ی ۱. رگرسیون محدب (۱۰ نمره)

فرض کنید می‌خواهیم مسئله‌ی رگرسیون به شکل زیر را حل کنیم:

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

که در اینجا $\mathcal{F}$ مجموعه همه توابع محدب است. روشی برای حل این مسئله‌ی بهینه‌سازی نامتنهای بعدی ارائه دهید.
(نکته جالب: برخلاف حالت classification این مسئله قابل یادگیری است)

حل. ابتدا توجه می‌کنیم که برای هر تابع محدب f داریم:

$$f(x) \geqslant f(x_i) + (x - x_i)^T z_i$$

که در اینجا z_i زیرگرادیان تابع محدب در نقطه‌ی z_i می‌باشد. و حال مسئله‌ی متناهی البعد زیر را حل می‌کنیم:

$$\min_{\{\tilde{y_i}, z_i\}_{i=1}^{n}} (y_i - \tilde{y_i})^2$$

$$\text{s.t.} \quad \tilde{y_j} \geqslant \tilde{y_i} + \tilde{z_i}^T (x_j - x_i)$$

و از جواب نهایی برای تعریف تابع نهایی به شکل زیر استفاده می‌کنیم:

$$\hat{f}(x) = \max_{i=1,2,\ldots,n} \{\hat{y_i} + \hat{z_i}^T (x - x_i)\}$$

که با توجه به این که max تعدادی تابع آفین(محدب) است خود محدب خواهد بود. ▷

مسئله‌ی ۲. به‌روزرسانی وزن‌ها (۱۵ نمره)

فرض کنید که فرض main weak learner در الگوریتم AdaBoost برقرار باشد. فرض کنید h_t یادگیرنده پایه‌ای باشد که در مرحله t انتخاب شده است. نشان دهید که یادگیرنده پایه h_{t+1} که در مرحله $t+1$ انتخاب شده است باید متفاوت از h_t باشد.

حل. با توجه به فرض یادگیری ضعیف، یک فرضیه $h \in \mathcal{H}$ وجود دارد که خطای D_{t+1} آن کمتر از نصف است. خطای تجربی h_t را برای توزیع D_{t+1} بررسی کنید. از آنجا که $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ و $\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$،

$$\hat{R}_{\mathcal{D}_{t+1}}(h_t) = \sum_{i=1}^{m} \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \cdot 1_{[h_t(x_i)y_i<0]}$$

$$= \sum_{h_t(x_i)y_i<0}^{m} \frac{D_t(i)\exp\left(\alpha t\right)}{Z_t}$$

$$= \frac{e^{\alpha t}}{Z_t} \sum_{h_t(x_i)y_i<0}^{m} D_t(i)$$

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

این نشان می دهد که h_t نمی تواند در دور t + ۱ انتخاب شود. ▷

مسئله‌ی ۳. لاسو گروهی با گروه های همپوشان (۱۵ نمره)

نشان دهید اگر از لاسو گروهی برای گروه های همپوشان استفاده کنیم در این که پرامتر های غیر صفر ما در اجتماعی از این گروه ها قرار بگیرند کمکی نمی‌کند. به طور دقیق تر استفاده از این هموارساز باعث می‌شود که پارامتر ها در مکمل اجتماعی از این گروه ها قرار بگیرند. (نیازی به اثبات این مورد نیست، برای حل این مشکل Overlapping group lasso) طراحی شده است که می‌توانید در مورد آن بیشتر بخوانید.

حل. فرض کنید که $\theta \in \mathbb{R}^3$ فضای پارامتر های ما باشد. و دو گروه مدنظر ما برابر $\{1,2\}$ و $\{1,3\}$ باشند. فرض کنیم که $\theta_1 = \theta_2 = 1$ در این صورت:

$$f(\theta_3) = \sqrt{1+\theta_3^2} - 1$$

این تابع در نقطه‌ی صفر مشتق پذیر است و مشتقی برابر صفر دارد، این موضوع باعث می‌گردد که θ_3 به تنک بودن تشویق نشود. با عوض کردن جای θ_2, θ_3 می‌توان حکم مشابه را برای θ_2 نتیجه گرفت و لذا اگر متغیر اول فعال باشد معمولا دو متغیر دیگر نیز فعال خواهند بود. ▷

مسئله‌ی ۴. در جستجوی هموارساز خوب (۱۰ نمره)

در دو کرانی که برای پیچیدگی راداماخر مدل های خطی به دست آوردیم دو عبارت ظاهر می‌شد:

الف) برای نرم ۱ عبارتی که ظاهر می‌شد برابر $||w||_1||x||_\infty$ بود (در واقع کران هایی روی این ها ولی برای سادگی فعلا این عبارت ها را در نظر می‌گیریم.)

ب) برای نرم ۲ عبارت $||w||_2||x||_2$ ظاهر می‌شد.

فرض کنید w, x متغیر های تصادفی نزدیک مقادیر $\{-۱, ۱\}$ باشند. در حالت های زیر این دو کران را بررسی کنید و نتیجه بگیرید هموارسازی با نرم ۱ و ۲ چه زمان هایی مناسب تر هستند.

الف) بدون هیچ فرض اضافه‌ای.

ب) w تنک با حداکثر k مولفه ناصفر باشد.

ج) w چگال باشد مثلا با فرض $||w||_۱ \approx \sqrt{d}||w||_۲$

حل.

الف) بدون هیچ فرض اضافه‌ای عبارت شامل نرم ۲ تقریبا برابر $\sqrt{d}\sqrt{d}$ خواهد بود عبارت شامل نرم ۱ برابر $d۱$ و لذا تفاوتی بین دو عبارت نخواهد بود.

ب) در این حالت عبارت شامل نرم ۱ برابر $k۱$ خواهد بود و عبارت شامل نرم ۲ برابر $\sqrt{k}\sqrt{d}$ و لذا استفاده از نرم ۱ در این گونه مسائل مفید خواهد بود.

ج) در این حالت خواهیم داشت:

$$||w||_۲||x||_۲ \leqslant \frac{۱}{\sqrt{d}}||w||_۱\sqrt{d}||x||_\infty = ||w||_۱||x||_\infty$$

و لذا در این حالت بهتر است از نرم ۲ به عنوان هموارساز استفاده کنیم.

$\triangleright$

مسئله‌ی ۵. خطای نمایی **Bayes** (۱۵ نمره)

فرض کنید مجموعه ورودی $\mathcal{X}$ و فضای برچسب $Y = \{-۱, +۱\}$ باشد. در الگوریتم AdaBoost از تابع خطای نمایی زیر استفاده می‌شود:

$$\ell(h(x), y) = \exp(-yh(x)).$$

برای یک توزیع $\mathcal{D}$ روی $\mathcal{X} \times Y$، خطای نمایی به صورت زیر تعریف می‌شود:

$$R_l(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)].$$

خطای نمایی بیز برای یک توزیع $\mathcal{D}$ به صورت خطای کمینه روی توابع قابل اندازه‌گیری تعریف می‌شود:

$$R_l^* = \inf_{h:\mathcal{X}\to\mathbb{R}\text{measurable}} R_l(h).$$

فرضیه‌ای $h_{\exp}$ با $R_l(h_{\exp}) = R_l^*$ به‌عنوان راه‌حل بهینه بیز نامیده می‌شود. $\eta(x)$ را به صورت $\eta(x) = P[y = +۱|x]$ تعریف کنید.

(a) عبارت رابطه راه‌حل بهینه بیز $h_{\exp}$ را برای خطای نمایی بر حسب $\eta(x)$ بدهید.

(b) خطای generalization و خطای بیز برای طبقه‌بندی باینری را به صورت زیر تعریف کنید:

$$R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}[sign(h(x)) \neq y]], \quad R^* = \inf_{h:\mathcal{X}\to\mathcal{Y} \text{ measurable}} R(h).$$

که $\text{sign}(t) = ۱_{t\geqslant ۰} - ۱_{t<۰}$.

نشان دهید $R(h_{\exp}) = R^*$.

حل. الف)

بر اساس تعریف، $R_\ell(h)$ به صورت زیر بیان می‌شود:

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

$$= \mathbb{E}_x\mathbb{E}_{y|x}\left[\exp(-yh(x))\right]$$

$$= \mathbb{E}_x\left[\eta(x)\exp(-h(x)) + (۱-\eta(x))\exp(h(x))\right]$$

$$\geqslant \mathbb{E}_x\left[۲\sqrt{\eta(x)(۱-\eta(x))}\right],$$

که برابری برقرار است اگر و تنها اگر برای هر $x \in X$، داشته باشیم:

$$h(x) = \frac{۱}{۲}\log\left(\frac{\eta(x)}{۱-\eta(x)}\right).$$

بنابراین،

$$R_\ell^* = \mathbb{E}_x\left[۲\sqrt{\eta(x)(۱-\eta(x))}\right]$$

خطای بیز برای خطای نمایی است و

$$h_{\exp} : x \mapsto \frac{۱}{۲}\log\left(\frac{\eta(x)}{۱-\eta(x)}\right)$$

جواب بهینه بیز است.

ب)

بر اساس تعریف، $R(h)$ به صورت زیر بیان می‌شود:

$$R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[۱_{\text{sign}(h(x))\neq y}\right],$$

$$R^* = \inf_{h:X\to\mathbb{R}} R(h),$$

که در آن $\text{sign}(t) = 1_{t\geqslant 0} - 1_{t<0}$. اثبات کنید که $R(h_{\exp}) = R^*$.

بر اساس تعریف، $R(h)$ به صورت زیر بیان می‌شود:

$$\begin{aligned} R(h) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{\mathrm{sign}(h(x))\neq y}\right] \\ &= \mathbb{E}_x\mathbb{E}_{y|x}\left[\mathbb{1}_{\mathrm{sign}(h(x))\neq y}\right] \\ &= \mathbb{E}_x\left[\eta(x)\mathbb{1}_{h(x)<0} + (1-\eta(x))\mathbb{1}_{h(x)\geqslant 0}\right] \\ &\geqslant \mathbb{E}_x\left[\min\{\eta(x), 1-\eta(x)\}\right], \end{aligned}$$

که برابری برقرار است اگر و تنها اگر برای هر $x \in X$، $\mathrm{sign}(h(x)) = \mathrm{sign}(\eta(x) - 1/2)$ باشد. از آنجا که برای هر $x \in X$، $\mathrm{sign}(h_{\exp}(x)) = \mathrm{sign}(\eta(x) - 1/2)$، ثابت شد که $R(h_{\exp}) = R^*$.

$\triangleright$

مسئله‌ی ۶. خطای طبقه‌بند ضعیف (۱۰ نمره)

نشان دهید خطای h_t با توجه به توزیع $D^{(t+1)}$ دقیقاً برابر با ۱/۲ است. به عبارت دیگر، نشان دهید که برای هر $t \in [T]$ داریم:

$$\sum_{i=1}^{m} D_i^{(t+1)}\mathbb{I}[y_i \neq h_t(x_i)] = \frac{1}{2}.$$

حل. طبق تعریف ϵ_t :

$$\sum_{i=1}^{m} D_i^{(t)} \exp\left(-w_t y_i h_t(x_i)\right) \cdot \mathbb{1}_{[h_t(x_i)\neq y_i]} = \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = \sqrt{\epsilon_t(1-\epsilon_t)}. \quad (1)$$

به طور مشابه،

$$\sum_{j=1}^{m} D_j^{(t)} \cdot \exp\left(-w_t y_j h_t(x_j)\right) = \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} + (1-\epsilon_t) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} = 2\sqrt{\epsilon_t(1-\epsilon_t)}. \quad (2)$$

برابری مورد نظر با مشاهده این موضوع به دست می‌آید که خطای h_t نسبت به توزیع D_{t+1} با تقسیم معادله (۱) بر معادله (۲) قابل محاسبه است. $\triangleright$