دانشکده علوم ریاضی
دانشگاه صنعتی شریف

نظریه یادگیری محاسباتی

امید اعتصامی، محمدهادی فروغمنداعرابی
بهار ۱۳۹۳

یادگیری محدب

جلسه‌های ؟؟؟ تا ؟؟؟

نگارنده: فرزاد جعفررحمانی

در این بخش می‌خواهیم مساله یادگیری محدب را معرفی کنیم. یادگیری محدب یک خانواده مهم از مسائل یادگیری است، به این دلیل که بیشتر آنچه می‌توانیم به صورت کارا یادبگیریم در این خانواده قرار می‌گیرد. به صورت کلی یک مساله یادگیری محدب مساله‌ای است که کلاس فرضیه آن مجموعه محدب است و تابع خطای آن نیز تابع محدب برای هر مثال است. برا همین در ابتدای این فصل با بعضی از تعریف‌ها مانند محدب، هموار، و $Lipschitz$ آشنا خواهیم شد.

۱ محدب، $Lipschitz$، هموار

تعریف ۱. *مجموعه C در یک فضای برداری محدب است، اگر برای هر دو بردار u و v در C و برای هر $\alpha \in [0,1]$ داشته باشیم $\alpha u + (1-\alpha)v \in C$.*

تعریف ۲. *مجموعه محدب C را در نظر بگیرید. تابع $f : C \to \mathbb{R}$ محدب است، اگر برای هر $u, v \in C$ و $\alpha \in [0,1]$ داشته باشیم:*

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v) \tag{1}$$

یک ویژگی مهم توابع محدب این است که هر کمینه موضعی تابع یک کمینه تابع نیز است. به صورت رسمی در نظر بگیرید $B(u,r)=\{v:||v-u||\le r\}$ یک مجموعه از نقاط به شعاع r حول u باشد. می‌گوییم $f(u)$ یک کمینه موضعی تابع f است، اگر وجود داشته باشد $r>0$ که $v\in B(u,r)$. این موضوع نشان می‌دهد که برای هر v، یک $\alpha>0$ وجود دارد که $u+\alpha(v-u)\in B(u,r)$. بنابراین داریم:

$$f(u)\le f(u+\alpha(v-u)) \tag{2}$$

اگر تابع f محدب باشد، همچنین داریم:

$$f(u+\alpha(v-u))=f(\alpha v+(1-\alpha)u)\le(1-\alpha)f(u)+\alpha f(v) \tag{3}$$

در نتیجه با ترکیب این معادله، می‌توان نوشت $f(u)\le f(v)$ که به دلیل اینکه v دلخواه بود، پس $f(u)$ کمینه تابع f خواهد بود.

گرادیان تابع f در نقطه w که با $\bigtriangledown f(w)$ نمایش می‌دهیم، برابر است با:

$$\bigtriangledown f(w)=\left(\frac{\partial f(w_1)}{\partial w_1},...,\frac{\partial f(w_d)}{\partial w_d}\right) \tag{4}$$

در نتیجه برای هر تابع محدب، مشتق پذیر f داریم:

$$\forall u, f(u)\ge f(w)+\langle\bigtriangledown f(w),u-w\rangle \tag{5}$$

لم ۳. *تابع $f:\mathbb{R}\to\mathbb{R}$ را در نظر بگیرید. شرایط زیر معادل خواهند بود:*

۱. *f محدب است.*

۲. *f' به صورت یکنوا غیرنزولی است.*

۳. *f'' نامنفی است.*

مثال: تابع $f(x)=x^2$ محدب است. و همچنین $f'(x)=2x$ و $f''(x)=2>0$.

لم ۴. *فرض کنید تابع $f:\mathbb{R}^d\to\mathbb{R}$ را بتوان به صورت $f(w)=g(\langle w,x\rangle+y)$ نوشت که $x\in\mathbb{R}^d$، $y\in\mathbb{R}$ و $\alpha\in[0,1]$. آنگاه اگر g محدب باشد، f نیز محدب خواهد بود.*

اثبات. $w_1,w_2\in\mathbb{R}^d$ و $\alpha\in[0,1]$ را در نظر بگیرید.

$$\begin{aligned}
f(\alpha w_1+(1-\alpha)w_2) &= g(\langle\alpha w_1+(1-\alpha)w_2,x\rangle+y) && (6)\\
&= g(\alpha\langle w_1,x\rangle+(1-\alpha\langle w_2,x\rangle)+y) && (7)\\
&= g(\alpha(\langle w_1,x\rangle+y)+(1-\alpha)(\langle w_2,x\rangle+y)) && (8)\\
&\le \alpha g(\langle w_1,x\rangle+y)+(1-\alpha)g(\langle w_2,x\rangle+y) && (9)
\end{aligned}$$

□

مثال: تابع $f(w)=(\langle w,x\rangle-y)^2$ را در نظر بگیرید. f ترکیب تابع $g(a)=a^2$ و یک تابع خطی است. در نتیجه f محدب است.

مثال: تابع $f(w)=\log(1+\exp(-y\langle w,x\rangle))$ که $y=\{1,-1\}$ را در نظر بگیرید. f ترکیب تابع $g(a)=\log(1+\exp(a))$ و یک تابع خطی است. در نتیجه f محدب است.

لم ۵. *برای $r, ..., 1 = i$ ، $f_i : \mathbb{R}^d \to \mathbb{R}$ را محدب در نظر بگیرید. توابع زیر نیز محدب هستند:*

۱. $g(x) = max_{i\in[r]} f_i(x)$

۲. $g(x) = \sum_{i=1}^{r} w_i f_i(x)$ که $w_i \geq 0$

اثبات. ابتدا به اثبات شماره ۱ می‌پردازیم.

$$g(\alpha u + (1-\alpha)v) = \max_i f_i(\alpha u + (1-\alpha)v) \quad (10)$$
$$\leq \max_i [\alpha f_i(u) + (1-\alpha) f_i(v)] \quad (11)$$
$$\leq \alpha \max_i f_i(u) + (1-\alpha) \max_i f_i(v) \quad (12)$$
$$= \alpha g(u) + (1-\alpha) g(v) \quad (13)$$

برای نشان شماره ۲ نیز می‌توان نوشت:

$$g(\alpha u + (1-\alpha)v) = \sum w_i f_i(\alpha u + (1-\alpha)v) \quad (14)$$
$$\leq \alpha \sum w_i f_i(u) + (1-\alpha) \sum w_i f_i(v) \quad (15)$$
$$= \alpha g(u) + (1-\alpha) g(v) \quad (16)$$

□

مثال: تابع $g(x) = |x|$ که برابر است با $\max\{x, -x\}$ و از طرفی $f_1(x) = x$ و $f_2(x) = -x$ محدب هستند، محدب خواهد بود.

۲ تابع Lipschitz

تعریف ۶. *در نظر بگیرید $C \subseteq \mathbb{R}^d$. تابع $f : \mathbb{R}^d \to \mathbb{R}$ ، روی C ، ρ-Lipschitz است، اگر برای هر $w_1, w_2 \in C$ داشته باشیم:*

$$||f(w_1 - f(w_2))|| \leq \rho ||w_1 - w_2|| \quad (17)$$

مثال: $f(x) = \log(1 + \exp(x))$ یک تابع 1-Lipschitz روی $\mathbb{R}$ است.

$$|f'(x)| = \left|\frac{\exp(x)}{1 + \exp(x)}\right| = \left|\frac{1}{\exp(-x) + 1}\right| \leq 1 \quad (18)$$

در مثال بالا از قضیه مقدار میانی استفاده شده که اگر f مشتق پذیر باشد، در نتیجه $f(w_1) = f(w_2) \leq f'(u)(w_1 - w_2)$
مثال: $f(x) = x^2$ تابع ρ-Lipschitz نیست. $x_1 = 0$ و $x_2 = 1 + \rho$ را در نظر بگیرید. می‌توان نوشت:

$$f(x_2) - f(x_1) = (1+\rho)^2 > \rho(1+\rho) = \rho|x_2 - x_1| \quad (19)$$

اما این تابع روی مجموعه $C = \{x : |x| \leq \frac{\rho}{2}\}$، ρ-Lipschitz خواهد بود.
تابع خطی $f : \mathbb{R}^d \to \mathbb{R}$ که برابراست با $f(w) = \langle v, w \rangle + b$ که $v \in \mathbb{R}^d$ یک تابع $||v||$-Lipschitz است.

$$|f(w_1) - f(w_2)| = |\langle v, w_1 - w_2 \rangle| \leq ||v|| \; ||w_1 - w_2|| \quad (20)$$

لم ۷. *تابع $f(x) = g_1(g_2(x))$ را در نظر بگیرید. اگر g_1 و g_2 به ترتیب ρ_1-Lipschitz و ρ_2-Lipschitz باشند، آنگاه f ، $\rho_1\rho_2$-Lipschitz خواهد بود.*

اثبات.

$$|f(w_1) - f(w_2)| = |g_1(g_2(w_1)) - g_1(g_2(w_2))| \leq \rho_1||g_2(w_2) - g_2(w_1)|| \leq \rho_1\rho_2||w_2 - w_1||$$

□

۳ توابع هموار

تعریف ۸. *تابع مشتق‌پذیر f را βـ هموار است، اگر گرادیان آن β-Lipschitz باشد. یعنی برای هر v و w داشته باشیم:*

$$|| \nabla f(v) - \nabla f(w)|| \leq \beta||v - w|| \quad (۲۱)$$

اگر f هموار باشدآنگاه برای هر v و w داریم:

$$f(v) \leq f(w) + \langle \nabla f(w), v - w\rangle + \frac{\beta}{2}||v - w||^2 \quad (۲۲)$$

اگر v را برابر با $w - \frac{1}{\beta} \nabla f(w)$ قرار دهیم، عبارت زیر بدست می‌آید:

$$\frac{1}{2\beta}|| \nabla f(w)||^2 \leq f(w) - f(v) \quad (۲۳)$$

همچنین اگر برای هر v، $f(v) \geq 0$ آنگاه $|| \nabla f(w)||^2 \leq 2\beta f(w)$

مثال: تابع $f(x) = \log(1 + \exp(x))$، $\frac{1}{4}$ـ هموار است.

لم ۹. *تابع $f(w) = y(\langle w, x\rangle + b)$ که $g : \mathbb{R} \to \mathbb{R}$ و g و، βـ هموار است. را در نظر بگیرید. آنگاه f ، $\beta||x||^2$ـ هموار خواهد بود.*

با استفاده از قاعده زنجیره‌ای می‌توان نوشت:

$$\nabla f(w) = g'(\langle w, x\rangle + b)x \quad (۲۴)$$

حال با استفاده از هموار بودن g و نامساوی کوشی شوارتز می‌توانیم نتیجه را به صورت زیر بدست آوریم.

$$\begin{aligned}
f(v) &= g(\langle w, x\rangle + b) && (۲۵)\\
&\leq g(\langle w, x\rangle + b) + g'(\langle w, x\rangle + b)\langle v - w, x\rangle + \frac{\beta}{2}(\langle v - w, x\rangle)^2 && (۲۶)\\
&\leq g(\langle w, x\rangle + b) + g'(\langle w, x\rangle + b)\langle v - w, x\rangle + \frac{\beta}{2}(||v - w||\ ||x||)^2 && (۲۷)\\
&= f(w) + \langle \nabla f(w), v - w\rangle + \frac{\beta|x||^2}{2}||v - w||^2 && (۲۸)
\end{aligned}$$

مثال: تابع $f(w) = (\langle w, x\rangle - y)^2$ یک تابع $2||x||^2$ـ هموار است.

مثال: تابع $f(w) = \log(1 + \exp(-y\langle w, x\rangle))$ که $y \in \{+1, -1\}$، یک تابع $\frac{||x||^2}{4}$ـ هموار است.

۴ مسائل یادگیری محدب

تعریف ۱۰. *مساله یادگیری (H, Z, l) را محدب می‌گوییم، اگر کلاس فرضیه H محدب باشد و برای هر $z \in Z$ تابع خطای $l(., z)$ تابع محدب باشد.*

مثال: (رگرسیون خطی با خطای مربع) در این مساله کلاس H برابر $\{x \to \langle w, x\rangle : w \in \mathbb{R}^d\}$ و خطا به صورت $l(h,(x,y)) = (h(x)-y)^2$ است. این مساله یک مساله یادگیر محدب است. زیرا مجموعه H و تابع خطا هر دو محدب هستند.

لم ۱۱. *اگر تابع خطای l یک تابع محدب باشد و H نیز محدب باشد. آنگاه ERM_H یک مساله بهینه کردن محدب است.*

اثبات.

$$ERM_H(S) = \underset{w \in H}{\arg\min}\, L_S(w) \tag{۲۹}$$

به دلیل اینکه برای نمونه $S = z_1, ..., z_m$ و هر w، داریم $L_S(w) = \frac{1}{m}\sum_{i=1}^{m} l(w, z_i)$، در نتیجه $L_S(w)$ محدب است. بنابراین حکم ثابت شد. □

۵ یادگیری مسائل «یادگیری محدب»

موارد زیادی است که انجام دادن ERM برای مسائل یادگیری محدب کارا خواهد بود. اما آیا محدب بودن شرط کافی برای یادگیری این دست مسائل است؟ جواب این سوال منفی است. در مثال بعد نشان داده می‌شود که همه مسائل یادگیری محدب قابل یادگیری نیستند.

مثال: می‌خواهیم نشان دهیم مساله یادگیری رگرسیون اگر $d = 1$ باشد قابل یادگیری نیست. فرض کنید که این گونه نباشد. یعنی الگوریتم A وجود دارد که برای این مساله یک یادگیری PAC است. در نتیجه تابع $m(.,.)$ وجود دارد به طوری که برای هر توزیع D و برای هر ϵ و δ، اگر A با ورودی $m \geq m(\epsilon, \delta)$ اجرا شود با احتمال حداقل $1 - \delta$، عبارت $L_D(A(S)) - \min_w L_D(w) \leq \epsilon$ برقرار خواهد بود.

حال در نظر بگیرید $\epsilon = \frac{1}{100}$، $\delta = \frac{1}{2}$ و $\mu = \frac{\log(100/99)}{2m}$. دو توزیع تعریف می‌کنیم که الگوریتم A در برابر آنها شکست خواهد خورد. D_1 را روی دو مثال $z_1 = (1, 0)$ و $z_2 = (\mu, -1)$ در نظر بگیرید که چگالی احتمال برای مثال اول برابر μ و برای مثال دوم برابر $1 - \mu$ خواهد بود. همچنین توزیع D_2 رو به این صورت در نظر بگیرید که روی مثال ۱ باشد و جاهای دیگر برابر صفر باشد. به وضوح اگر $A(S) < \frac{-1}{2\mu}$ الگوریتم در برابر توزیع اول شکست خواهد خورد و در غیر اینصورت در برابر توزیع دوم شکست خواهد خورد.

همان‌طور که در این مثال دیدم شرط محدب بودن برای یادگیری همواره کافی نیست. در نتیجه نیازمند شرایط دیگری به مساله هستیم. بنابراین برای رسیدن به این هدف نیازمند تعریف کردن انواع مسائل دیگر با شرایط اضافی‌تر خواهیم بود که در ادامه پرداخته شده است.

تعریف ۱۲. *مساله یادگیری (H, Z, l) را با پارامترهای ρ و B Convex-Lipschitz-Bounded می‌گوییم، اگر شرایط زیر برقرار باشد:*

۱. *کلاس فرضیه H محدب باشد و برای هر $w \in H$ داشته باشیم $||w|| \leq B$.*

۲. *برای هر $z \in Z$، تابع خطای $l(., z)$، تابع محدب و ρ-Lipschitz باشد.*

مثال: مجموعه‌های $\{X \in \mathbb{R}^d : ||x|| \leq \rho\}$، $y = \mathbb{R}$ و $H = \{w \in \mathbb{R}^d : ||w|| \leq B\}$ را در نظر بگیرید و تابع خطا را به صورت $l(w,(x,y)) = |\langle w, x\rangle - y|$ بگیرید. این مساله یک مساله یادگیری Convex-Lipschitz-Bounded است.

تعریف ۱۳. *مساله یادگیری (H, Z, l) را با پارامترهای β و B، Convex-Lipschitz-Bounded می‌گوییم، اگر شرایط زیر برقرار باشد:*

۱. *کلاس فرضیه H محدب باشد و برای هر w داشته باشیم* $||w|| \leq B$.

۲. *برای هر $z \in Z$، تابع خطای $l(.,z)$، تابع محدب، نامنفی و βــ هموار باشد.*

مثال: مجموعه $X = \left\{x \in \mathbb{R}^d : ||x|| \leq \frac{\beta}{۲}\right\}$، $y = \mathbb{R}$ و $H = \{w \in \mathbb{R}^d : ||w|| \leq B\}$ را در نظر بگیرید. همچنین تابع خطای $l(۲,(x,y)) = (\langle w,x \rangle - y)^۲$. این مساله یک مساله یادگیری Convex-Lipschitz-Bounded است.