دانشکده علوم ریاضی
دانشگاه صنعتی شریف

نظریه یادگیری محاسباتی

امید اعتصامی، محمدهادی فروغمنداعرابی
بهار ۱۳۹۳

منظم‌سازی

جلسه‌های ؟؟؟ تا ؟؟؟؟

نگارنده: فرزاد جعفررحمانی

در بخش قبلی دو خانواده Convex-Lipschitz-Bounded و Convex-Smooth-Bounded را معرفی کردیم. در این بخش می‌خواهیم بگوییم که همه مسائل یادگیری در این دو خانواده قابل یادگیری هستند. بعضی از مسائل خاصیت همگرایی یکنواخت را دارند که با ERM می‌توان آنها را یاد گرفت. اما همیشه این طور نیست. بنابراین در این بخش الگوریتم دیگری را معرفی می‌کنیم. الگوریتم جدید RLM نام دارد. در RLM که مخفف Regularized Loss Minimization است مجموع خطای روی نمونه[1] و تابع منظم‌سازی[2] را کمینه می‌کنیم که تابع منظم‌سازی معیاری از پیچیدگی یک فرضیه است. به صورت رسمی یک تابع منظم‌ساز به صورت $R : \mathbb{R}^d \to \mathbb{R}$ است و RLM معادل است با:

$$\arg\min_{w}(L_s(w) + R(w)) \tag{۱}$$

تابع‌های R مختلفی وجود دارد، اما در این بخش تمرکز بر روی $R(w) = \lambda ||w||^2$ خواهد بود. که $\lambda > 0$. در نتیجه خروجی

[1] empirical risk
[2] regularization function

RLM به صورت زیر می‌توان نوشت.

$$\arg\min_{w}(L_s(w) + \lambda||w||^2) \tag{2}$$

این نوع R، تیخونوف[3] نام دارد.

در ادامه می‌خواهیم RLM را روی رگرسیون خطی انجام دهید. در نتیجه خروجی به صورت زیر خواهد بود.

$$\arg\min_{w\in\mathbb{R}^d}(\lambda||w||^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\langle w, x_i\rangle - y_i)^2) \tag{3}$$

انجام دادن رگرسیون خطی با استفاده از معادله قبل رگرسیون $ridge$ نامیده می‌شود که برای حل آن از آن گرادیان گرفته و برابر صفر قرار می‌دهیم. در نتیجه خواهیم داشت.

$$\begin{aligned}
(2\lambda mI + A)w &= b \Rightarrow w = (2\lambda mI + A)^{-1}b && (4)\\
A &= (\sum_{i=1}^{m} x_i x_i^T) && (5)\\
b &= \sum_{i=1}^{m} y_i x_i && (6)\\
& && (7)
\end{aligned}$$

قضیه ۱. *توزیع D روی $\chi \times [-1,1]$ را در نظر بگیرید که $\chi = \{x \in \mathbb{R}^d : ||x|| \leq 1\}$ و $H = \{w \in \mathbb{R}^d : ||w|| \leq B\}$. برای هر $\epsilon \in (0,1)$، m را به این صورت در نظر بگیرید که $m > \frac{150B^2}{\epsilon^2}$. آنگاه با جرای رگرسیون $ridge$ با $\lambda = \frac{\epsilon}{3B^2}$ خواهیم داشت:*

$$E[L_D(A(S))] \leq min_{w\in H} L_D(w) + \epsilon \tag{8}$$

۱ منظم بودن

به طور شهودی می‌گوییم یک الگوریتم یادگیری منظم است اگر تغییرات کوچک در ورودی، خروجی الگوریتم را زیاد تغییر ندهد. حال می‌خواهیم تعریف منظم بودن را بیان کنیم. فرض کنید A یک الگوریتم یادگیری است. $S = (z_1, ..., z_m)$ را در نظر بگیرید که m تا نمونه هستند. فرض کنید یک نمونه اضافی z' داده شده است. $s^{(i)}$ را به صورت $s^{(i)} = (z_1, ..., z_{i-1}, z', z_{i+1}, ..., z_m)$ در نظر می‌گیریم. حال الگوریتم یادگیری، الگوریتم خوبی است اگر $l(A(S^{(i)}), z_i) - l(A(S), z_i) \geq 0$.

قضیه ۲. *فرض کنید $S = (z_1, ..., z_m)$ یک دنباله $i.i.d$ از مثال‌ها باشد و z' یک مثال دیگر. $U(m)$ را توزیع یکنواخت روی $[m]$ در نظر بگیرید. آنگاه برای هر الگوریتم یادگیری داریم:*

$$E_{S\sim D^m}[L_D(A(S)) - L_S(A(S))] = E_{(S,z')\sim D^{m+1}, i\sim U(m)}[l(A(S^{(i)}, z_i)) - L_S(A(S), z_i)] \tag{9}$$

اثبات. به دلیل اینکه S و z' هر دو $i.i.d$ هستند، در نتیجه برای هر i داریم:

$$E_S[L_D(A(S))] = E_{S,z'}[l(A(S), z')] = E_{S,z'}[l(A(S^{(i)}), z_i)] \tag{10}$$

$$E_S[L_S(A(S))] = E_{S,i}[l(A(S), z_i)] \tag{11}$$

که ترکیب دو معادله اثبات را کامل خواهد کرد. □

[3] Tikhonov

تعریف ۳. *تابع $\epsilon : \mathbb{N} \to \mathbb{R}$ را در نظر بگیرید که به صورت یکنوا نزولی باشد. می‌گوییم الگوریتم A با نرخ $\epsilon(m)$، $On-Average-Replace-One-Stable$ است اگر برای هر توزیع D داشته باشیم:*

$$E_{(S,z')\sim D^{m+1}, i\sim U(m)}[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \le \epsilon(m) \quad (12)$$

تعریف ۴. *تابع f را λــ قویا محدب می‌گویند اگر برای هر w، u و $\alpha \in (0,1)$ داشته باشم:*

$$f(\alpha w + (1-\alpha)u) \le \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)||w-u||^2 \quad (13)$$

لم ۵. ۱. *تابع $f(w) = \lambda ||w||^2$، λــ قویا محدب است.*

۲. *اگر f و g، λــ قویا محدب باشد آنگاه $f+g$، λــ قویا محدب خواهد بود.*

۳. *اگر f، λــ قویا محدب باشد و u نقطه کمینه f باشد آنگاه برای هر w داریم:*

$$f(w) - f(u) \ge \frac{\lambda}{2}||w-u||^2 \quad (14)$$

اثبات. قسمت ۱ و ۲ به طور مستقیم از تعریف بدست می‌آید. برای قسمت سوم با توجه تعریف می‌توان نتیجه گرفت:

$$\frac{f(u+\alpha(w-u)) - f(u)}{\alpha} \le f(w) - f(u) - \frac{\lambda}{2}\alpha(1-\alpha)||w-u||^2 \quad (15)$$

وقتی $\alpha \to 0$ سمت چپ، مشتق تابع $g(\alpha) = f(u+\alpha(w-u))$ خواهد بود و سمت راست برابر $f(w)-f(u)-\frac{\lambda}{2}||w-u||^2$ خواهد بود. از طرفی به دلیل اینکه u کمینه تابع f است در نتیجه $\alpha = 0$ نیز کمینه تابع g خواهد بود. بنابراین سمت چپ، وقتی $\alpha \to 0$ برابر صفر خواهد بود و حکم ثابت می‌شود. □

حال می‌خواهیم نشان دهیم که RLM منظم است. تابع f_s به صورت $f_s(w) = L_S(w) + \lambda ||w||^2$ تعریف می‌کنیم. با استفاده از لم قبل f_s، 2λــ قویا محدب خواهد بود. حال با توجه به قسمت دوم لم قبل داریم:

$$f_s(v) - f_s(A(S)) \ge \lambda ||v - A(S)||^2 \quad (16)$$

از طرفی برای هر v، u و i داریم:

$$\begin{aligned} f_s(v) - f_s(u) &= L_S(v) + \lambda||v||^2 - (L_S(u) + \lambda||u||^2) && (17)\\ &= L_{S^{(i)}}(v) + \lambda||v||^2 - (L_{S^{(i)}}(u) + \lambda||u||^2) + \frac{l(v,z_i) - l(u,z_i)}{m} + \frac{l(v,z') - l(u,z')}{m} && (18)\end{aligned}$$

قرار دهید $v = A(S^{(i)})$ و $u = A(S)$ و همچنین می‌دانیم که v کمینه $L_{S^{(i)}}(w) + \lambda||w||^2$ است. در نتیجه داریم:

$$f_s(A(S^{(i)})) - f_s(A(S)) \le \frac{l(A(S^{(i)}), z_i) - l(A(S), z_i)}{m} + \frac{l(A(S), z') - l(A(S^{(i)}), z')}{m} \quad (19)$$

ترکیب معادله اخیر با معادله $f_s(v) - f_s(u) = L_S(v) + \lambda||v||^2 - (L_S(u) + \lambda||u||^2)$ نتیجه می‌دهد:

$$\lambda||A(S^{(i)}) - A(S)||^2 \le \frac{l(A(S^{(i)}), z_i) - l(A(S), z_i)}{m} + \frac{l(A(S), z') - l(A(S^{(i)}), z')}{m} \quad (20)$$

اگر تابع $l(., z_i)$، ρ-Lipschitz باشد، داریم:

$$l(A(S^{(i)}), z_i) - l(A(S), z_i) \le \rho||A(S^{(i)}) - A(S)|| \quad (21)$$

و همچنین به طور مشابه می‌توان نوشت:

$$l(A(S^{(i)}), z') - l(A(S), z') \leq \rho||A(S^{(i)}) - A(S)|| \tag{۲۲}$$

جمع معادله اخیر با معادله بدست آمده در مرحله آخر بخش منظم‌سازی نتیجه خواهد داد:

$$\rho||A(S^{(i)}) - A(S)||^{۲} \leq \frac{۲\rho||A(S^{(i)}) - A(S)||}{m} \tag{۲۳}$$

در نتیجه می‌توان نوشت:

$$||A(S^{(i)}) - A(S)|| \leq \frac{۲\rho}{\lambda m} \tag{۲۴}$$

جمع معادله اخیر با معادله اولی، نتیجه می‌دهد:

$$l(A(S^{(i)}), z_i) - l(A(S), z_i) \leq \frac{۲\rho^{۲}}{\lambda m} \tag{۲۵}$$

که این معادله برای هر S، i و z' درست است. حال می‌توان نتیجه مستقیم زیر را گرفت:
نتیجه: فرض کنید تابع خطای l محدب و Lipschitz-ρ باشد. آنگاه RLM با تابع $\lambda||w||^{۲}$ یک $on-average-replace-one-stable$ با نرخ $\frac{۲\rho^{۲}}{\lambda m}$ خواهد بود.
حال با توجه به قضیه ذکر شده در این بخش داریم:

$$E_{S\sim D^m}[L_D(A(S)) - L_S(A(S))] \leq \frac{۲\rho^{۲}}{\lambda m} \tag{۲۶}$$

اگر تابع خطا βــ هموار و نامنفی باشد، در نتیجه خواهیم داشت:

$$|| \bigtriangledown f(w)||^{۲} \leq ۲\beta f(w) \tag{۲۷}$$

همچنین فرض کنید که $\lambda \geq \frac{۲\beta}{m}$، حال با توجه به فرض هموار بودن داریم:

$$l(A(S^{(i)}), z_i) - l(A(S), z_i) \leq \langle \bigtriangledown l(A(S), z_i), A(S^{(i)}) - A(S)\rangle + \frac{\beta}{۲}||A(S^{(i)}) - A(S)||^{۲} \tag{۲۸}$$

با استفاده از نامساوی کوشی شوارتز و معادله $|| \bigtriangledown f(w)||^{۲} \leq ۲\beta f(w)$ داریم :

$$\begin{aligned} l(A(S^{(i)}), z_i) - l(A(S), z_i) &\leq || \bigtriangledown l(A(S), z_i)||\ ||A(S^{(i)}) - A(S)|| + \frac{\beta}{۲}||A(S^{(i)}) - A(S)||^{۲} && (۲۹)\\ &\leq \sqrt{۲\beta\ l(A(S), z_i)}\ ||A(S^{(i)}) - A(S)|| + \frac{\beta}{۲}||A(S^{(i)}) - A(S)||^{۲} && (۳۰)\end{aligned}$$

به طور مشابه می‌توان برای z' همین عبارت را نوشت که به صورت زیر خواهد شد:

$$\begin{aligned} l(A(S^{(i)}), z') - l(A(S), z') &\leq || \bigtriangledown l(A(S), z')||\ ||A(S^{(i)}) - A(S)|| + \frac{\beta}{۲}||A(S^{(i)}) - A(S)||^{۲} && (۳۱)\\ &\leq \sqrt{۲\beta\ l(A(S), z')}\ ||A(S^{(i)}) - A(S)|| + \frac{\beta}{۲}||A(S^{(i)}) - A(S)||^{۲} && (۳۲)\end{aligned}$$

جمع این معادله با معادله بدست آمده در آخر بخش منظم‌سازی نتیجه خواهد داد:

$$||A(S^{(i)}) - A(S)|| \leq \frac{\sqrt{۲\beta}}{\lambda m - \beta}(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}) \tag{۳۳}$$

ترکیب این معادله‌ها با فرض $\beta \leq \frac{\lambda m}{۲}$ نتیجه خواهد داد:

$$||A(S^{(i)}) - A(S)|| \leq \frac{\sqrt{۸\beta}}{\lambda m}(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}) \tag{۳۴}$$

با توجه به معادله بدست آمده و همچنین فرض $\beta \leq \frac{\lambda m}{2}$ می‌توان نوشت:

$$\begin{aligned} l(A(S^{(i)}), z_i) - l(A(S), z_i) &\leq \sqrt{2\beta l(A(S), z_i)}||A(S^{(i)}) - A(S)|| + \frac{\beta}{2}||A(S^{(i)}) - A(S)||^2 && (35) \\ &\leq \left(\frac{4\beta}{\lambda m} + \frac{8\beta^2}{(\lambda m)^2}\right)\left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}\right)^2 && (36) \\ &\leq \frac{8\beta}{\lambda m}\left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S^{(i)}), z')}\right)^2 && (37) \\ &\leq \frac{24\beta}{\lambda m}\left(l(A(S), z_i) + l(A(S^{(i)}), z')\right) && (38) \end{aligned}$$

با گرفتن امیدریاضی از دو طرف معادله بالا و همچنین با توجه به اینکه $E[l(A(S), z_i)] = E[l(A(S^{(i)}), z')] = E[L_s A(S)]$ می‌توان نتیجه زیر را گرفت:

نتیجه: فرض کنید تابع خطا βـ هموار و نامنفی باشد. آنگاه اجرای RLM با $\lambda||w||^2$ که $\lambda \geq \frac{2\beta}{m}$ نتیجه خواهد داد:

$$E[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \frac{48\beta}{\lambda m} E[L_s A(S)] \quad (39)$$

با توجه به نتیجه نیر می‌توان نوشت:

$$E[l(A(S^{(i)}), z_i) - l(A(S), z_i)] \leq \frac{48\beta C}{\lambda m} \quad (40)$$

که می‌دانیم برای z داریم $l(0, z) \leq C$

۲ کنترل حد بین منظم‌سازی و نزدیک شدن به مثال‌ها

می‌توان امید ریاضی خطا را به صورت زیر نوشت:

$$E_s[L_D A(S)] = E_s[L_s A(S)] + E_s[L_D A(S) - L_s A(S)] \quad (41)$$

جمله اول نشان‌دهنده این است که چقدر به نمونه داده شده نزدیک هستیم. همان‌طور که در قضیه دوم نشان داده شد، جمله دوم، معادل با منظم بودن الگوریتم A است. برای کمینه کردن خطا الگوریتم نیازمندیم که جمع هردو ترم کمینه شود. حال می‌خواهیم برای خطای $emprical$ که با استفاده از RLM حاصل شده است، کران بالای بدست آوریم. بردار دلخواه w^* را در نظر بگیرید. در نتیجه داریم:

$$L_s A(S) \leq L_s A(S) + \lambda||A(S)||^2 \leq L_s(w^*) + \lambda||w^*||^2 \quad (42)$$

با گرفتن امید ریاضی از دو طرف نامساوی بالا می‌توان نوشت:

$$E_s[L_s A(S)] \leq L_D(w^*) + \lambda||w^*||^2 \quad (43)$$

جمع معادله با معادله داده شده در ابتدای بخش نتیجه خواهد داد:

$$E_s[L_D A(S)] \leq L_D(w^*) + \lambda||w^*||^2 + E_s[L_D(A(S)) - L_s(A(S))] \quad (44)$$

ترکیب این معادله با نتیجه بدست آمده در بخش قبل می‌توان نتیجه زیر را نوشت:

نتیجه: فرض کنید که تابع خطا محدب و ρ-Lipschitz است. آنگاه با اجرای RLM با $\lambda||w||^2$ خواهیم داشت:

$$\forall w^* \;\; E_s[L_D A(S)] \leq L_D(w^*) + \lambda||w^*||^2 + \frac{2\rho^2}{\lambda m} \quad (45)$$

این کران بالا نامساوی $oracle$ نامیده می‌شود. اگر w^* یک فرضیه با خطای کم باشد، این کران به ما می‌گوید که چه تعداد مثال نیاز است که $A(S)$ به خوبی w^* باشد.

با استفاده از این نتیجه می‌توان نتیجه زیر را گرفت:

نتیجه: فرض کنید (H, Z, l) یک مساله یادگیری $Convex - Lipschitz - Bounded$ با پارامترهای B و ρ باشد. برای هر نمونه mــ تایی در نظر بگیرید $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$. آنگاه با اجرای RLM با $\lambda||w||^2$ خواهیم داشت:

$$E_s[L_D(A(S))] \leq min_{w\in H} L_D(w) + \rho\beta\sqrt{\frac{8}{m}} \tag{46}$$

به طور خاص برای هر $\epsilon > 0$ اگر $m \geq \frac{8\rho^2 B^2}{\epsilon^2}$ آنگاه برای هر توزیع D داریم:

$$E_s[L_D(A(S))] \leq min_{w\in H} L_D(w) + \epsilon \tag{47}$$

نتیجه: اگر تابع خطا محدب، βــ هموار و نامنفی باشد با اجرای RLM با $\lambda||w||^2$ برای $\lambda \geq \frac{2\beta}{m}$ خواهیم داشت:

$$E_s[L_D(A(S))] \leq (1 + \frac{48\beta}{\lambda m})E_s[L_s(A(S))] \leq (1 + \frac{48\beta}{\lambda m})(L_D(w^*) + \lambda||w^*||^2) \tag{48}$$

نتیجه: فرض کنید (H, Z, l) یک مساله یادگیری $Convex - Smooth - Bounded$ با پارامترهای B و β باشد. همچنین فرض کنید که $l(0, z) \leq 1$. برای هر $z \in Z$ و هر $\epsilon \in (0, 1)$، در نظر بگیرید $m \geq \frac{150\beta B^2}{\epsilon^2}$ و $\lambda = \frac{\epsilon}{3B^2}$. در نتیجه برای هر توزیع D خواهیم داشت:

$$E_s[L_D(A(S))] \leq min_{w\in H} L_D(w) + \epsilon \tag{49}$$