دانشکده‌ی مهندسی کامپیوتر

نظریه یادگیری ماشین

نیم‌سال اول ۱۴۰۳-۱۴۰۴

مدرس: دکتر امیر نجفی

تمرین دوم

مسئله‌ی ۱. وابستگی پیچیدگی راداماخر به توزیع (۱۰ نمره)

فرض کنید توزیع P را به گونه‌ای انتخاب کرده‌ایم که تک‌نقطه‌ای باشد یعنی $z = \delta(z_0)$ و نمونه‌ها از این توزیع می‌آیند. فرض کنید خانواده‌ی $\mathcal{F}$ را به گونه‌ای انتخاب کرده‌ایم که برای هر $x \in \mathbb{R}$ و $f \in \mathcal{F}$، $-1 \leqslant f(x) \leqslant 1$. نشان دهید که:

$$\mathcal{R}_n(\mathcal{F}) \leqslant \frac{1}{\sqrt{n}}$$

حل.

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{z_1,\dots z_n}\mathbb{E}_{\sigma_1,\dots,\sigma_n}\frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(z_i)$$

$$= \mathbb{E}_{\sigma_1,\dots,\sigma_n}\frac{1}{n}\sup_{f\in\mathcal{F}} f(z_0)\sum_{i=1}^{n}\sigma_i$$

$$\leqslant \mathbb{E}_{\sigma_1,\dots,\sigma_n}\frac{1}{n}\sup_{f\in\mathcal{F}}|f(z_0)||\sum_{i=1}^{n}\sigma_i| \leqslant \mathbb{E}_{\sigma_1,\dots,\sigma_n}\frac{1}{n}\sqrt{\sum_{i=1}^{n}\sigma_i^2} = \frac{1}{\sqrt{n}}$$

$\triangleright$

مسئله‌ی ۲. خواص بعد VC (۱۵ نمره)

بعد VC یک خانواده از مجموعه های $\mathcal{S}$ را برابر بعد VC توابع به شکل 1_S تعریف می‌کنیم که $S \in \mathcal{S}$. فرض کنید $\mathcal{S}, \mathcal{T}$ دو خانواده از مجموعه ها با بعد VC محدود باشند نشان دهید خانواده های زیر نیز بعد VC محدود دارند:

الف)

$$\mathcal{S}^c := \{S^c | S \in \mathcal{S}\}$$

ب)

$$\mathcal{S} \sqcap \mathcal{T} = \{S \cap T | S \in \mathcal{S}, T \in \mathcal{T}\}$$

ج)

$$\mathcal{S} \sqcup \mathcal{T} = \{S \cup T | S \in \mathcal{S}, T \in \mathcal{T}\}$$

حل.

الف) اگر $\mathcal{S}$ بتواند یک مجموعه‌ی n عضوی را shatter کند $\mathcal{S}^c$ نیز می‌تواند این مجموعه را shatter کند و بالعکس (کافیست لیبل های برعکس را در نظر بگیریم و این ترکیب را بسازیم و آن را مکمل بگیریم) که نتیجه می‌دهد که بعد VC هر دو $\mathcal{S}$ و $\mathcal{S}^c$ یکسان است.

ب) فرض کنید بعد VC $\mathcal{S}, \mathcal{T}$ به ترتیب برابر ν, μ باشد در این صورت در یک مجموعه‌ی n عضوی هر کدام به ترتیب $(n+۱)^{\mu}$ و $(n+۱)^{\nu}$ ترکیب را تولید می‌کنند و لذا در کل تعداد ترکیب هایی که با اشتراک این ها می‌توان تولید کرد برابر است با:

$$(n+۱)^{\mu+\nu}$$

و لذا اگر n را به گونه‌ای انتخاب کنیم که:

$$(n+۱)^{\mu+\nu} < ۲^n$$

آنگاه یک مجموعه‌ی n عضوی را نمی‌توان با اشتراک اعضای $\mathcal{S}$ و $\mathcal{T}$ shatter کرد.

ج) مشابه قبلی :)

$\triangleright$

مسئله‌ی ۳. خواص پیچیدگی راداماخر (۱۵ نمره)

خواص زیر از پیچیدگی راداماخر را ثابت کنید:

الف)

$$\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(\text{conv}(\mathcal{F}))$$

ب)

$$\mathcal{R}_n(\mathcal{F}+\mathcal{G}) \leqslant \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$$

و مثالی ارائه دهید که نتوان کران را بهبود داد.

ج) برای تابع ثابت و کراندار g:

$$\mathcal{R}_n(\mathcal{F}+g) \leqslant \mathcal{R}_n(\mathcal{F}) + \frac{||g||_{\infty}}{\sqrt{n}}$$

حل.

الف) کافیست توجه کنیم که:

$$\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(z_i)\sigma_i = sup_{f\in \mathrm{conv}(\mathcal{F})}\sum_{i=1}^{n} f(z_i)\sigma_i$$

برای اثبات فرض کنیم مقدار بالا برای $\sum_{j=1}^{m}\alpha_j f_j$ بیشینه گردد در این صورت این ضرب داخلی برای همه‌ی f_j ها باید یکسان باشد (در غیر این صورت اگر ضریب آن را صفر کنیم عبارت بیشتر می‌گردد) و کافیست آن f را برای بیشینه کردن روی $\mathcal{F}$ در نظر بگیریم.

ب) از خاصیت زیر برای sup نتیجه می‌گردد که در کلاس اثبات شد:

$$\sup_{x\in\mathcal{X}} f(x)+g(x) \leqslant \sup_{x\in\mathcal{X}} f(x) + \sup_{x\in\mathcal{X}} g(x)$$

ج)

$$\mathcal{R}_n(\mathcal{F}+g) = \mathbb{E}_{z_1,\ldots z_n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}(f(z_i)+g(z_i))\sigma_i$$

$$\leqslant \mathbb{E}_{z_1,\ldots z_n}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(z_i)\sigma_i + ||g||_\infty \sum_{i=1}^{n}\sigma_i$$

از طرفی بنابر نابرابر ینسن داریم:

$$\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\sum_{i=1}^{n}\sigma_i \leqslant \sqrt{\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\sum_{i=1}^{n}\sigma_i^2} = \sqrt{n}$$

و حکم نتیجه می‌گردد.

▷

مسئله‌ی ۴. محاسبه بعد VC (۱۰ نمره)

بعد VC مجموعه‌های زیر را حساب کنید:

الف) مجموعه دایره‌های در صفحه.

ب) مجموعه مربع‌های در صفحه.

حل.

الف) اول دایره ها می‌توانند هر سه نقطه‌ی غیر هم خط را shatter کنند. زیرا می‌توان شعاع دایره را به اندازه کافی زیاد کرد تا مشابه یک خط باشد و می‌دانیم که با خط ها می‌توان سه نقطه‌ی غیر هم خط را shatter کرد. برای اثبات این که هیچ ۴ نقطه‌ای را نمی‌توان با این دایره ها shatter کرد کافیست توجه کنیم که اگر پوش محدب این نقاط شامل سه نقطه باشد کافیست نقطه‌ی داخل پوش محدب را منفی و بقیه را مثبت کنیم در غیر این صورت circle minimum-enclosing را در نظر می‌گیرم اگر دو نقطه‌ی روبه‌رو قطری روی آن بود و دو نقطه‌ی دیگر در دو سمت مختلف قطر حضور داشتند کافیست دو نقطه‌ی روی قطری را مثبت و دو تای دیگر را منفی کنیم. اگر هر دو یک سمت قطر بودند آن نقطه‌ی که همراه با دو نقطه روی قطر دایره‌ی محیطی کوچکتری را تشکیل می‌دهد را مثبت و سایر را منفی می‌کنیم و اگر سه نقطه روی این دایره بودند کافیست دو نقطه مقابل در پوش محدب که هر دو روی دایره هستند را مثبت و سایر را منفی کنیم.

ب) مشابه قبلی

$\triangleright$

مسئله‌ی ۵. SVM تنک (۱۵ نمره)

یکی دیگر از روش‌های فرموله کردن الگوریتم SVM، بر پایه تنکی (sparsity) بردارهای کمکی است. فرض کنید به جای بیشینه کردن مقدار حاشیه می‌خواهیم تنکی بردارها را بیشینه کنیم. این کار را با کمینه کردن $||\alpha||_p$ که با آن بردار w به دست می‌آید انجام می‌دهیم ($p \geqslant ۱$). ابتدا حالتی را درنظر بگیرید که $p = ۲$. به مساله بهینه سازی زیر می‌رسیم:

$$\min_{\alpha,b} \frac{۱}{۲}\sum_{i=۱}^{m}\alpha_i^۲ + C\sum_{i=۱}^{m}\xi_i$$

$$\text{subject to} \quad y_i\left(\sum_{j=۱}^{m}\alpha_j y_j \mathbf{x}_i\cdot\mathbf{x}_j + b\right) \geqslant ۱ - \xi_i, \quad i\in[m]$$

$$\xi_i, \alpha_i \geqslant ۰, \quad i\in[m].$$

الف) نشان دهید که با در نظر گرفتن قید نامنفی بودن برای α این مسئله را می‌توانیم به یک نمونه از مسئله اولیه SVM تبدیل کنیم.

ب) مساله بهینه سازی دوگان مساله بالا را بنویسید.

ج) اگر قرار دهیم $p = ۱$، به یک α تنک‌تر دست می‌یابیم. مساله دوگان را در این حالت به دست آورید.

حل. الف) فرض کنید

$$\mathbf{x}'_i = (y_۱(\mathbf{x}_۱\cdot\mathbf{x}_i), \ldots, y_m(\mathbf{x}_m\cdot\mathbf{x}_i)).$$

سپس مسئله بهینه‌سازی به صورت زیر خواهد بود:

$$\min_{\alpha,b,\xi}\frac{۱}{۲}||\alpha||^۲ + C\left(\sum_{i=۱}^{m}\xi_i\right)$$

با قیود زیر:

$$y_i(\alpha \cdot \mathbf{x}_i' + b) \geqslant ۱ - \xi_i, \quad \xi_i \geqslant ۰, \quad \alpha_i \geqslant ۰, \quad i \in [m].$$

ب) متغیر های لاگرانژ زیر را تعریف میکنیم و مساله لاگرانژین را می نویسیم:

$$p_i, q_i, r_i \geqslant ۰$$

$$L(\alpha, b, \xi, p, q, r) = \frac{۱}{۲}\|\alpha\|^۲ + C\left(\sum_{i=۱}^{m} \xi_i\right) - \sum_{i=۱}^{m} p_i \left\{ y_i \left[\left(\sum_{j=۱}^{m} \alpha_j y_j x_j \right) \cdot x_i + b \right] - ۱ + \xi_i \right\}$$

$$- \sum_{i=۱}^{m} q_i \xi_i - \sum_{i=۱}^{m} r_i \alpha_i.$$

پس از برابر قرار دادن گرادیان عبارت بالا نسبت به متغیر های primal با صفر داریم:

$$L(\alpha, b, \xi, p, q, r) = \frac{۱}{۲}\|\alpha\|^۲ + C\left(\sum_{i=۱}^{m} \xi_i\right) - \sum_{i=۱}^{m} \alpha_i(\alpha_i - r_i) - \sum_{i=۱}^{m} p_i y_i b - \sum_{i=۱}^{m} p_i - \sum_{i=۱}^{m} (p_i + q_i)\xi_i - \sum_{i=۱}^{m} r_i \alpha_i$$

$$= -\frac{۱}{۲}\left(\sum_{i=۱}^{m} p_i y_i x_i\right)^۲ + \sum_{i=۱}^{m} p_i.$$

پس در نهایت با قرار دادن عبارات کنار هم داریم:

$$\max_{p,r} \quad \sum_{i=۱}^{m} p_i - \frac{۱}{۲}\left\|\sum_{i=۱}^{m} p_i y_i x_i + r\right\|^۲$$

$$subject\ to\ ۰ \leqslant p_i \leqslant C \quad \forall i, \quad \sum_{i=۱}^{m} p_i y_i = ۰, \quad r \in \mathbb{R}.$$

ج) راه حل این قسمت بسیار مشابه قسمت ب است با این تفاوت که عبارت

$$\frac{۱}{۲}\|\alpha\|^۲$$

با عبارت

$$\|\alpha\|_۱ = \sum_{i=۱}^{m} |a_i|$$

جایگزین خواهد شد و از تکرار راه حل صرف نظر می کنیم. $\triangleright$

مسئله‌ی ۶. قضیه رادون (۱۰ نمره)

برای مجموعه $S = \{\mathbf{x}_۱, \ldots, \mathbf{x}_m | \forall i \in [m] : \mathbf{x}_i \in \mathbb{R}^n\}$، پوش محدب S مجموعه‌ای است به صورت:

$$\{\mathbf{z} \in \mathbb{R}^n \mid \exists \lambda_۱, \dots, \lambda_m \in [۰, ۱], \sum_{i=۱}^{m} \lambda_i = ۱, \mathbf{z} = \sum_{i=۱}^{m} \lambda_i \mathbf{x}_i\}.$$

اگر $m \geqslant n + ۲$، ثابت کنید S دارای دو زیرمجموعه مجزا $S_۱$ و $S_۲$ است که پوش محدب آن‌ها تداخل دارد.

حل. لم: اگر $P = \{p_۱, \dots, p_n\} \subset \mathbb{R}^d$ مجموعه ای متناهی از نقاط درفضای dـ بعدی باشد، اگر $n > d$، آنگاه برای بعضی ضرایب $\mu_۱, \dots, \mu_n$داریم:

$$۰ = \sum_{i=۱}^{n} \mu_i p_i \quad \text{with} \quad \mu_۱, \dots, \mu_n \text{not all zero} ,$$

و اگر $n > d + ۱$، می توانیم عبارت بالا را با شرط اضافی

$$\sum_{i=۱}^{n} \mu_i = ۰.$$

داشته باشیم. در حالت دوم، بعضی از $\mu_۱, \dots, \mu_n$ مثبت و بعضی منفی هستند.

اثبات: اگر $n = d + ۲$ و $P = \{p_۱, \dots, p_n\} \subset \mathbb{R}^d$، طبق لم بالا ، برای ضرایب $\mu_۱, \dots, \mu_n$ (که بعضی مثبت و بعضی نامنفی هستند)داریم:

$$۰ = \sum_{i=۱}^{n} \mu_i p_i \quad \text{with} \quad \sum_{i=۱}^{n} \mu_i = ۰.$$

قرار دهید $J = \{j \in \{۱, \dots, n\} : \mu_j > ۰\}$ و $I = \{۱, \dots, n\} \setminus J$. نه J و نه I تهی نیستند. قرار دهید

$$A = \sum_{j \in J} \mu_j \quad \text{که} \ A > ۰.$$

پس

$$\sum_{i \in I} \mu_i = -A.$$

قرار دهید $P_۱ = \{p_j : j \in J\}$ ، $P_۲ = \{p_i : i \in I\}$. پوش محدب $\mathrm{Conv}(P_۱)$ ، $\mathrm{Conv}(P_۲)$ شامل نقطه

$$p = \sum_{j \in J} \frac{\mu_j}{A} x_j = \sum_{i \in I} \frac{-\mu_i}{A} x_i,$$

است چرا که حاصل جمع اول یک ترکیب محدب از نقاط $P_۱$ و حاصل جمع دوم ترکیب محدب از نقاط $P_۲$است.
پس،

$$\mathrm{Conv}(P_۱) \cap \mathrm{Conv}(P_۲) \neq \emptyset.$$

$\triangleright$

مسئله‌ی ۷. مثال نقض (۱۰ نمره)

اغلب این‌گونه است که بُعد VC یک کلاس فرضیه برابر (یا از بالا محدود شونده) با تعداد پارامترهایی است که برای تعریف هر فرضیه در کلاس نیاز است. برای مثال، اگر $\mathcal{H}$ کلاس مستطیل‌های هم‌تراز با محور در $\mathbb{R}^d$ باشد، آنگاه $VCdim\mathcal{H} = ۲d$ است که برابر با تعداد پارامترهای مورد نیاز برای تعریف مستطیلی در $\mathbb{R}^d$ می‌باشد. در اینجا مثالی وجود دارد که نشان می‌دهد این همیشه درست نیست. ما خواهیم دید که یک کلاس فرضیه می‌تواند بسیار پیچیده باشد و حتی یادگیری‌پذیر نباشد، اگرچه تعداد کمی پارامتر دارد.

دامنه $\mathcal{X} = \mathbb{R}$ و کلاس فرضیه

$$\mathcal{H} = \{x \mapsto \lceil \sin(\theta x) \rceil : \theta \in \mathbb{R}\}$$

را در نظر بگیرید. (اینجا، $۰ = \lceil -۱ \rceil$ را می‌گیریم). اثبات کنید که $VCdim\mathcal{H} = \infty$.

نکته: راه‌های متعددی برای اثبات این نتیجه وجود دارد. یک راه با اعمال لم زیر است: اگر $۰.x_۱x_۲x_۳\ldots$ بسط دودویی $x \in (۰, ۱)$ باشد، آنگاه برای هر عدد طبیعی m، $\lceil \sin(۲^m \pi x) \rceil = (۱ - x_m)$ به شرطی که $\exists k \geqslant m$ که در آن $x_k = ۱$.

حل. ابتدا لم مطرح شده را اثبات میکنیم:

$$\begin{aligned}\sin(۲^m\pi x) &= \sin\left(۲\pi(۰.x_۱x_۲\ldots)\right) = \sin\left(۲\pi(x_۱x_۲\ldots x_{m-۱}x_m x_{m+۱}\ldots)\right)\\ &= \sin\left(۲\pi(x_۱x_۲\ldots x_{m-۱}x_m x_{m+۱}\ldots) - ۲\pi(x_۱x_۲\ldots x_{m-۱}/۰)\right) = \sin\left(۲\pi(۰.x_m x_{m+۱}\ldots)\right).\end{aligned}$$

حال، اگر $x_m = ۰$، آنگاه $۲\pi(۰.x_m x_{m+۱}\ldots) \in (۰, \pi)$ ($\exists k \geqslant m$ s.t. $x_k = ۱$)، پس عبارت بالا مثبت است و، $\lceil \sin(۲^m\pi x) \rceil = ۱$. همچنین، اگر $x_m = ۱$، آنگاه $۲\pi(۰.x_m x_{m+۱}\ldots) \in [\pi, ۲\pi)$، پس عبارت بالا نامثبت است و، $\lceil \sin(۲^m\pi x) \rceil = ۰$. در نتیجه داریم:

$$\lceil \sin(۲^m\pi x) \rceil = ۱ - x_m.$$

برای اینکه ثابت کنیم $\mathrm{VC}(\mathcal{H}) = \infty$، باید n نقطه انتخاب کنیم که توسط shatter$\mathcal{H}$ می شوند، برای هر n. پس، n نقطه $x_۱, \ldots, x_n \in [۰, ۱]$می سازیم، به گونه ای که مجموعه بیت mـام در نمایش باینری، وقتی که m از ۱ تا ۲^n تغییر میکند، روی تمام حالات labeling $x_۱, \ldots, x_n$ تغیر میکند:

$$x_۱ = ۰.۰۰۰۰۰\ldots۱۱$$

$$x_۲ = ۰.۰۰۰۰۰\ldots۱۱$$

$$\vdots$$

$$x_{n-۱} = ۰.۰۰۱۱\ldots۱۱$$

$$x_n = ۰.۰۱۰۰\ldots۰۱$$

برای مثال برای آنکه لیبل ۱ برای همه نمونه هارا بسازیم، $h(x) = \lceil \sin(۲^۱ x) \rceil$را انتخاب می کنیم، که اولین بیت یا ستون از بسط باینری را میسازد. اگر بخواهیم لیبل ۱ برای $x_۱, \ldots, x_{n-۱}$، و لیبل ۰ برای x_n داشته باشیم، $h(x) = \lceil \sin(۲^۲ x) \rceil$را انتخاب میکنیم، که بیت دوم از بسط باینری را میسازد و به همین ترتیب میتوان ادامه داد.

پس نتیجه میگیریم که $x_1, \ldots, x_n$ میتواند هر labeling برای برخی مقادیر $h \in \mathcal{H}$را تولید کند، پس .شدهshatter این کار را میتوان برای هر nانجام داد، پس $\mathrm{VCdim}(\mathcal{H}) = \infty$. ▷

مسئله‌ی ۸. رأی اکثریت (۱۵ نمره)

نشان دهید که اگر کلاس فرضیه $\mathcal{H}$ دارای VC-dimension برابر d باشد، آنگاه کلاس $\mathrm{MAJ}_k(\mathcal{H})$ دارای VC-dimension برابر $O(kd \log kd)$ است. منظور از $\mathrm{MAJ}_k(\mathcal{H})$ کلاس توابعی است که با گرفتن رأی اکثریت بر روی k تابع از $\mathcal{H}$ به دست می‌آید.

حل. فرض کنید D ، VC-dimension $\mathrm{MAJ}_k(\mathcal{H})$باشد، پس طبق تعریف، وجود دارد مجموعه S متشکل از D نقطه که با $\mathrm{MAJ}_k(\mathcal{H})$ shatter شده. طبق لم ساور میدانیم که حداکثر D^d روش دسته بندی نقاط S با استفاده از توابع $\mathcal{H}$وجود دارد.

حال چون هر تابع h در $\mathrm{MAJ}_k(\mathcal{H})$ با برخی k تابع $h_1, h_2, \ldots, h_k$ در $\mathcal{H}$مشخص شده، این یعنی دسته بندی S توسط h با دسته بندی S توسط $h_1, \ldots, h_k$مشخص شده. چون ماکسیمم $(D^d)^k = D^{dk}$ حالت انتخاب k دسته بندی از S مطابق با $\mathcal{H}$وجود دارد (احتمالا با تکرار)، این یعنی حداکثر D^{kd} حالت دسته بندی نقاط S با استفاده از توابع $\mathrm{MAJ}_k(\mathcal{H})$ وجود دارد.

از طرفی، چون S توسط $\mathrm{MAJ}_k(\mathcal{H})$shatter شده، میدانیم همه 2^D دسته بندی ممکن هستند. در نتیجه داریم

$$2^D \leqslant D^{kd},$$

و $D \leqslant 2kd \log(kd)$ (برای $kd \geqslant 4$). ▷