نظریه یادگیری ماشین

نیم‌سال دوم ۱۴۰۰-۱۴۰۱

مدرس: دکتر بیگی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری پنجم

موعد تحویل: ۹ تیر

مسئله‌ی ۱. (۲۰ نمره)

در مسائل دسته‌بندی به دست آوردن کران تعمیم پذیری برمبنای بعد VC را مرور کردیم. در مسئله رگرسیون و فضاهای فرضیه آن از Pdim به عنوان معیار پیچیدگی استفاده می‌شود. میخواهیم کران تعمیم‌پذیری برمبنای این معیار به دست بیاوریم.
فرض کنید $\mathcal{H}$ فضای فرضیه توابع با خروجی حقیقی باشد، M کران بالا برای تابع هزینه مسئله است و $\mathcal{G} = \{x \mapsto L(h(x), f(x)) : h \in \mathcal{H}\}$ مجموعه متناظر با هزینه توابع این فضای فرضیه باشد. می دانیم $Pdim = d$.
نشان دهید به ازای هر $\delta > ۰$ با احتمال حداقل $۱ - \delta$ مربوط به انتخاب m نمونه آموزش، کران تعمیم‌پذیری زیر $\forall h \in \mathcal{H}$ برقرار است:

$$R(h) \leqslant \hat{R}(h) + M\sqrt{\frac{2d \log \frac{em}{d}}{m}} + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{1}$$

مسئله‌ی ۲. (۳۰ نمره)

میخواهیم PAC-Learnable بودن را برای الگوریتم رگرسیون خطی یک بعدی بررسی کنیم. می‌دانیم $\mathcal{H} = \{w \in \mathbb{R}\}$ و $l(w, (x, y)) = (wx - y)^2$.

الف) نشان دهید اگرچه شرط محدب بودن برقرار است اما این مسئله PAC-Learnable نیست.
(راهنمایی: می‌توانید از برهان خلف و تعریف PAC-Learnable استفاده کنید.)

ب) نشان دهید در حالتی که شرط کران دار بودن نیز اضافه شود، یعنی $\mathcal{H} = \{w \in \mathbb{R} : ||w|| \leqslant B\}$ ، مسئله PAC-Learnable نیست.

مسئله‌ی ۳. (۲۰ نمره)

یک مسئله یادگیری محدب با فضای فرضیه کراندار $\mathcal{W} \subseteq \mathbf{R}^d$ را به صورت $(\forall w \in \mathcal{W}, ||W|| \leqslant B)$ در نظر بگیرید.
می‌دانیم تابع هزینه $l(z, w)$ ، $\rho - lipschitz$ و محدب نسبت به w است. نسخه تنظیم سازی شده [۱] از آن را با تابع بهینه سازی زیر در نظر می‌گیریم:

$$\min_{w \in \mathcal{W}} \hat{L}(w) + \frac{\alpha}{2}||w||^2 \tag{2}$$

نشان دهید این تابع هزینه جدید نیز lipscitz و محدب است. (پارامترlipscitz را نیز مشخص کنید.)

مسئله‌ی ۴. (۳۰ + ۷ نمره)

با استفاده از رویکرد PAC-Bayes می‌توان کران‌های تعمیم‌پذیری برای الگوریتم‌های یادگیری به دست آورد. در درس با یکی از کران‌های شناخته شده به کمک این رویکرد آشنا شدیم. در این کران عبارت $KL(Q||P)$ ظاهر می‌شود که

---

[۱] Regularized

در آن P توزیع پیشین و Q توزیع پسین روی فضای فرضیه مورد نظر است. این عبارت در حالت کلی مقدار کوچکی ندارد و بنابراین کرانی که به دست می‌آید، مقدار نسبتا بزرگی ممکن است داشته باشد. اما این رویکرد قابلیت‌های فراتری نیز دارد و با برداشتن گام‌هایی کران‌های بهتر و تنگ‌تر [۲] تری نیز می‌توان به دست آورد. برای این منظور، یک جهت اصلی محدود کردن و انتخاب توزیع پیشین مناسب به صورتی است که مقدار عبارت $KL(Q||P)$ برای خروجی الگوریتم مشخص مورد نظر ما کمینه شود. در واقع به جای انتخاب یک توزیع پیشین عمومی، یک توزیع پیشین مختص توزیع دادگان مسئله و خروجی الگوریتم، انتخاب می‌کنیم.
در این سوال میخواهیم برمبنای این جهت به کران تعمیم‌پذیری جدید دیگری برای یک الگوریتم یادگیری دست پیدا کنیم.
در مسئله یادگیری، فضای ورودی و برچسب آن را به صورت $(x, y) = z \in \mathcal{Z}$ و تابع هزینه را به صورت $l(h, z) \in [۰, ۱]$ در نظر بگیرید.
الگوریتمی مجموعه داده‌های آموزش S، شامل m نمونه که از توزیع D^m روی فضای $\mathcal{Z}^m$ نمونه‌گیری شده است را به عنوان ورودی دریافت می‌کند و سپس توزیع پسین Q_Sرا که توزیعی روی فضای فرضیه $\mathcal{H}$ که تعداد متناهی عضو دارد، به عنوان خروجی تولید می‌کند.
طبق قضیه‌ای که در درس با آن آشنا شدید، با احتمال حداقل $۱ - \delta$ داریم:

$$R(Q_s) \leqslant \hat{R}(Q_s) + \sqrt{\frac{KL(Q||P) + \ln(\frac{m+1}{\delta})}{2m}} \tag{3}$$

الف) ابتدا توزیع پیشین را به گونه‌ای به دست آورید که امید ریاضی روی S عبارت $KL(Q_s||P)$ به ازای آن کمینه شود.
(توجه داشته باشید که این توزیع می‌تواند به توزیع دادگان(D) بستگی داشته باشد.)

ب) فرضیه h Q_s را با نمونه برداری $\mathcal{H}$ براساس توزیع Q_S در نظر می‌گیریم. نشان دهید با استفاده از این توزیع پیشین کران زیر به دست می‌آید:

$$R(Q_s) \leqslant \hat{R}(Q_s) + \sqrt{\frac{I(h; S) + \ln(\frac{m+1}{\delta})}{2m}} \tag{4}$$

که در آن $I(h; S)$ برابر اطلاعات متقابل [۳] بین دسته‌بند نهایی و دادگان آموزش است.

I نیز مشابه KL یک مفهوم نظریه اطلاعاتی است که می‌توان به صورت زیر آن را بیان کرد:
X و Y دو متغیر تصادفی با توزیع مشترک $P(X, Y)$ و توزیع‌های حاشیه‌ای P_X و P_Y هستند. داریم:

$$I(X; Y) = KL(P_{(X,Y)}||P_X \otimes P_Y) = \sum_{x \in X} \sum_{y \in Y} P_{(X,Y)}(x, y) \log \left(\frac{P_{(X,Y)}(x, y)}{P_X \, P_Y} \right) \tag{۵}$$

ج) (امتیازی) تفسیر و برداشت خود از این کران جدید و ارتباط آن با تعمیم‌پذیری و بیش‌برازش [۴] را با توجه به مفهوم اطلاعات متقابل بیان کنید.

موفق باشید :)

[۲] Tight
[۳] Mutual Information
[۴] Overfitting