دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

استاد درس: دکتر محمدحسین رهبان

بهار ۱۴۰۰

تمرین ششم

درس یادگیری ماشین

نام و نام خانوادگی: امیر پورمند

شماره دانشجویی: ۹۹۲۱۰۲۵۹

آدرس ایمیل **pourmand1376@gmail.com**

۱ سوال ۱

۱.۱ الف

با استفاده از فرض یادگیرنده ضعیف باید یادگیرنده ای وجود داشته باشد که خطای آن کمتر از $\frac{1}{2}$ باشد. حال برای این که این فرضیه را تست کنیم خطای فرضیه h_t هنگامی که از توزیع وزن D_{t+1} استفاده میکنیم باید مقداری کوچکتر از یک باشد که در اینجا میبینیم نیست.
البته توجه داریم که روابط زیر برقرار است:

$$\alpha_t = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})$$
$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$
$$\epsilon_t = \sum_{y_i h_t(x_i)<0} D_t(i)$$

پس داریم:

$$\begin{aligned}
\hat{R}_{D_{t+1}}(h_t) &= \sum_{i=1}^{m} \frac{D_t(i)\exp(-y_i h_t(x_i)\alpha_t)}{Z_t} I\{y_i h_t(x_i) < 0\} \\
&= \frac{e^{\alpha_t}}{Z_t} \sum_{y_i h_t(x_i)<0} D_t(i) \\
&= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \frac{1}{2\sqrt{\epsilon_t(1-\epsilon_t)}} \epsilon_t \\
&= \frac{1}{2}
\end{aligned}$$

که همان طور که گفتیم خطای نیم با فرض اولیه ما تناقض دارد در واقع مدلی که این پیش بینی را انجام میدهد به درد خاصی نمیخورد!

۲.۱ ب

با استفاده از تعریف همبستگی میتوانیم بنویسیم:

$$\sum_{i=1}^{m} y_i h_t(x_i) D_{t+1}(i) = \sum_{i=1}^{m} y_i h_t(x_i) D_t(i) \frac{\exp(-\alpha_t y_i h_t(x_i))}{Z_t} = \frac{1}{Z_t}\frac{dZ_t}{d\alpha_t}$$

رابطه آخر از انجا نوشته شده است که میدانیم $Z_t = \sum_{i=1}^{m} D_t(i) y_i h_t(x_i)$ برقرار است. با توجه به این که میدانیم a_t در واقع Z_t را مینیمم میکند پس مشتق آن صفر است و میتوان نتیجه گرفت که نسبت به یکدیگر ناهمبسته هستند.

۲ سوال ۲

۱.۲ الف

خب میخواهیم یک روش پارمتریک ارائه دهیم. ابتدا یک فرض باید داشته باشیم که مثلا داده از توزیع گوسی میاید. من این فرض را انتخاب میکنم و جلو میروم. البته توجه داریم که چون فرض ما ممکن است درست نباشد با مشکلاتی نیز رو به رو خواهیم شد که بعدا توضیح میدهم.

برای هر توزیع گوسی یک میانگین و یک واریانس کافی است. با این که میتوان برای تخمین میانگین و واریانس از روش هایی مانند method of moments و MLE استفاده کرد و نتیجه تقریبا مشابهی بدست آورد (مانند روش MLE) ، ما صرفا از میانگین نمونه و واریانس نمونه استفاده میکنیم که تخمین بدی هم نیست و بایاس آن اگر توزیع گوسی باشد صفر است!

پس داریم

$$\hat{\mu} = \bar{X} = \sum_{i=1}^{n} x_i, \hat{\sigma^2} = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

و برای تخمین pdf نیز میتوان صرفا این دو مقدار را پس از بدست آوردن در تابع جایگذاری کرد و روش تکمیل خواهد شد. پس pdf برابر خواهد بود با:

$$\hat{p}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-1}{2\hat{\sigma}^2}(x_i - \hat{\mu})^2)$$

مزایای این روش آن است که اولا محاسبه اش ساده هست و پیچیدگی محاسباتی بالایی ندارد و ثانیا به داده های خیلی زیادی نیاز ندارد یعنی مثل بعضی روش های نانپارامتریک نیست که نیاز به تعداد بسیار زیادی داده داشته باشد تا خوب عمل کند ولی یک مشکل اصلی دارد که از همان فرض اولیه ما ناشی میشود. یک بایاس غیرقابل اجتناب وجود دارد که کاری هم برایش نمیتوان کرد! و ممکن است اگر توزیع اولیه ما گوسی نباید این بایاس خیلی زیاد شود.

۲.۲ ب

خب باید ثابت کنیم بایاس کوچکتر مساوی مقدار گفته شده است. ابتدا بایاس را حساب کنیم:

$$\begin{aligned}
E[\bar{p}(x)] &= H.P(x_i \in I_l) \\
&= H \int_{\frac{l-1}{H}}^{\frac{l}{H}} p(u)du \\
&= H[F(\frac{l}{H}) - F(\frac{l-1}{H})] \\
&= \frac{F(\frac{l}{H}) - F(\frac{l-1}{H})}{\frac{1}{H}} \\
&= \frac{F(\frac{l}{H}) - F(\frac{l-1}{H})}{\frac{l}{H} - \frac{l-1}{H}} \\
&= p(x^*), x^* \in [\frac{l-1}{H}, \frac{l}{H}]
\end{aligned}$$

اخرین تساوی با قضیه مقدار میانگین بدست آمده است. به شکل مشابه میتوان برای p نیز با استفاده از همین قضیه نوشت:

$$\frac{p(x^*) - p(x)}{x^* - x} = p(x^{**})$$

پس بایاس برابر است با:

$$\begin{aligned}
bias &= E[\hat{p}(x)] - p(x) \\
&= p(x^*) - p(x) \\
&= p'(x^{**})(x^* - x) \\
&\leq |p'(x^{**})||x^* - x| \\
&\leq \frac{\beta}{H}
\end{aligned}$$

۳.۲ پ

این انالیز به ما میگوید که هر چه تعداد بیشتری bin داشته باشیم در واقع بایاس کمتری داریم و وقتی میتوانیم تعداد bin بیشتری داشته باشیم که داده های بیشتری داده باشیم که در هر بین جا شوند. از طرفی با بدست آوردن واریانس این مدل میتوان دید که اگر تعداد بین ها بیش از حد زیاد شود مدل overfit میشود و باید همواره یک trade-off را در این زمینه لحاظ کرد و مقدار مناسبی را برای تعداد ظرف ها در نظر گرفت. (ظرف = بین = bin)

علت غیرپارامتریک بودن روش نیز ان است که پارامتر یا مجموعه خاصی از پارامترها یادگرفته نمیشوند که بعد از آن دیگر با دیتاها کاری نداشته باشیم مانند کاری که با مدل گوسی در قسمت الف انجام دادیم و با بدست اوردن میانگین و واریانس دیگر کاری با خود دادگان نداشتیم. روش به گونه ای هست که همواره به کل دیتاست برای تخمین نیاز است بنابراین غیرپارامتریک است.

۴.۲ ت

مزایای روش پارامتریک اول نسبت به دومی این است که اولا هایپرپارامتر برای tune کردن ندارد که در اکثر موارد آموزش دادن راحت تر خواهد بود و نیازی به پیدا کردن مقدار بهینه hyperparameter نیست و ثانیا روش اول نسبت به روش دوم زودتر کانورج میکند یعنی که زودتر به نتیجه میرسد ولی ایراد های آن نسبت به روش دوم این است یک بایاس غیرقابل اجتناب دارد در حالی که روش دوم این را ندارد.

۳ سوال ۳

۱.۳ الف

چون در سوال جایی ذکر نشده من فرض میکنم کلاس ها -۱ و ۱ هستند وگرنه به مشکل برمیخوریم.

مشخصا در صورتی که بیشتر از نصف دسته بندهای دودویی ما به اشتباه دسته بندی را انجام دهند مشکل پیش می اید زیرا تابع ما دارد جمع مقادیر کلاس ها برای تک تک دسته بندها را در نظر میگیرد و نظر جمع به هر سمتی برود جواب همان است. پس به طور خلاصه اگر بیشتر از $\frac{T}{2}$ دسته بندها اشتباه باشند خطا داریم.

۲.۳ ب

حال برای باند خطا میتوانیم بنویسیم:

$$P(H(x) \neq f(x)) = \sum_{i=0}^{T/2} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq exp(\frac{-T}{2}[2\epsilon - 1]^2)$$

که همان طور که مشخص است وقتی که T به سمت بی نهایت میل کند باند به صفر میل خواهد کرد زیرا توان ما همواره منفی است و با افزایش T منفی تر میشود.