بسمه تعالی

پاسخ سری هفتم تمرینات درس یادگیری ماشین

امیرحسین رمضانی بناب (۹۹۲۱۰۲۹۴)

۱ سوال ۱

۱.۱ الف

طبق فرض سوال در هر *iteration* از الگوریتم، *weak learner* به شکل زیر تعیین می‌شود.

$$h_t(x_i) = \min_h \epsilon_t = \min_h \sum_{i=1}^{m} \mathcal{D}_t(i) \mathbb{1}_{(h(x_i) \neq y_i)}$$

حال فرض کنیم در مرحله‌ی tام الگوریتم، تابع h_t انتخاب شده باشد. در آن صورت داریم:

$$\epsilon_t = \sum_{i=1}^{m} \mathcal{D}_t(i) \mathbb{1}_{(h_t(x_i) \neq y_i)}$$
$$\alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

پس برای مرحله‌ی بعدی، توزیع وزن‌ها به شکل زیر محاسبه می‌شود.

$$\mathcal{D}_{t+1}(i) = \frac{\mathcal{D}_t(i)}{Z_t} \exp(-\alpha_t y_i h_t(x_i))$$

و به طرز مشابه داریم

$$h_{t+1}(x_i) = \min_h \sum_{i=1}^{m} \mathcal{D}_{t+1}(i) \mathbb{1}_{(h(x_i) \neq y_i)}$$

حال فرض کنیم $h_{t+1} = h_t$. یعنی

$$\begin{aligned} h_t = \underset{h}{argmin}\ \epsilon_{t+1} &= \underset{h}{argmin} \sum_{i=1}^{m} \mathcal{D}_{t+1}(i) \mathbb{1}_{(h(x_i) \neq y_i)} \\ &= \underset{h}{argmin} \sum_{i=1}^{m} \frac{\mathcal{D}_t(i)}{Z_t} \exp(-\alpha_t y_i h(x_i)) \mathbb{1}_{(h(x_i) \neq y_i)} \\ &= \underset{h}{argmin} \underbrace{\sum_{i=1}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h(x_i)) \mathbb{1}_{(h(x_i) \neq y_i)}}_{E(h)} \end{aligned}$$

فرض کنیم عدد jای وجود دارد که $h_t(x_j) \neq y_j$. پس

$$h_t = \underset{h}{argmin} \sum_{\substack{i=1 \\ i\neq j}}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h(x_i)) \mathbb{1}_{(h(x_i)\neq y_i)} + \mathcal{D}_t(j) \exp(-\alpha_t y_j h(x_j))$$

تابع h'_t را به شکل زیر می‌سازیم. تنها تفاوت h'_t با h_t در نقطه‌ی x_j است.

$$h'_t(x) = \begin{cases} y_j & x = x_j \\ h_t(x) & o.w \end{cases}$$

اگر E را براساس این تابع جدید به دست آوریم خواهیم داشت

$$\begin{aligned}
E(h') &= \sum_{i=1}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h'(x_i)) \mathbb{1}_{(h'(x_i)\neq y_i)} \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h'(x_i)) \mathbb{1}_{(h'(x_i)\neq y_i)} + \mathcal{D}_t(j) \exp(-\alpha_t y_j h'(x_j)) \mathbb{1}_{(h'(x_j)\neq y_j)} \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h'(x_i)) \mathbb{1}_{(h'(x_i)\neq y_i)} + 0 \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h(x_i)) \mathbb{1}_{(h(x_i)\neq y_i)} \\
&< E(h)
\end{aligned}$$

که با اینکه $h_t = \underset{h}{argmin} E(h)$ در تناقض است. پس دو *iteration* متوالی نمی‌توانند توابع *weak learner* یکسان داشته باشند.

۲.۱ ب

ضرب داخلی $(y_1 h_t(x_1), y_2 h_t(x_2), ..., y_m h_t(x_m))$ و $(\mathcal{D}_{t+1}(1), \mathcal{D}_{t+1}(2), ..., \mathcal{D}_{t+1}(m))$ برابر است با:

$$\begin{aligned}
d &= \sum_{i=1}^{m} \mathcal{D}_{t+1}(i) y_i h_t(x_i) \\
&= \sum_{i=1}^{m} \frac{\mathcal{D}_t(i)}{Z_t} \exp(-\alpha_t y_i h_t(x_i)) y_i h_t(x_i) \\
&= \frac{1}{Z_t} \sum_{i=1}^{m} \mathcal{D}_t(i) \exp(-\alpha_t y_i h_t(x_i)) y_i h_t(x_i) \\
&= \frac{1}{Z_t} \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{m} \mathcal{D}_t(i) \exp(-\alpha_t) - \frac{1}{Z_t} \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{m} \mathcal{D}_t(i) \exp(\alpha_t) \\
&= \frac{1}{Z_t} \Big[\exp(-\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{m} \mathcal{D}_t(i) - \exp(\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{m} \mathcal{D}_t(i) \Big] \qquad (۱)
\end{aligned}$$

از طرف دیگر داریم

$$\epsilon_t = \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} \mathcal{D}_t(i)$$

و

$$\alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

پس

$$\frac{\exp(\alpha_t)}{\exp(-\alpha_t)} = \exp(2\alpha_t) = \frac{1-\epsilon_t}{\epsilon_t} = \frac{\sum_{\substack{i=1 \\ h_t(x_i) = y_i}}^{m} \mathcal{D}_t(i)}{\sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} \mathcal{D}_t(i)}$$

با طرفین وسطین داریم:

$$\exp(\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} \mathcal{D}_t(i) = \exp(-\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i) = y_i}}^{m} \mathcal{D}_t(i)$$

پس

$$\exp(-\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i) = y_i}}^{m} \mathcal{D}_t(i) - \exp(\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} \mathcal{D}_t(i) = 0$$

با جایگذاری در رابطه‌ی ۱ به دست می‌آید:

$$d = \frac{1}{Z_t} \left[\exp(-\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i) = y_i}}^{m} \mathcal{D}_t(i) - \exp(\alpha_t) \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} \mathcal{D}_t(i) \right] = 0$$

۲ سوال ۲

۱.۲ آ

همانطور که در کلاس بحث شد، یکی از روش‌های تخمین توزیع داده ها، استفاده از روش‌های Parametric است. در این روش‌ها می‌توانیم شکل کلی توزیع داده‌ها را فرض کرده و پارامترهای توزیع را بیابیم. به عنوان مثال فرض می‌کنیم داده‌های ما به شکل i.i.d از یک توزیع گوسین آمده باشند و پارامترهای توزیع را توسط *Log Liklihood* می‌یابیم. فرض کنیم $\mathcal{D} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ دیتاست ما باشد. همچنین چون توزیع گوسین است پس:

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

پس Liklihood به شکل زیر خواهد بود

$$p(\mathcal{D}; \mu, \sigma) = \prod_{i=1}^{n} p(\mathbf{x}_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(\mathbf{x}_i-\mu)^2}{2\sigma^2}\right)$$

پس

$$\mathcal{L}(\mu, \sigma) = \log p(\mathcal{D}; \mu, \sigma) = \sum_{i=1}^{n} \Big(-\log(\sigma) - \log\Big(\sqrt{2\pi}\Big) - \frac{(\mathbf{x}_i-\mu)^2}{2\sigma^2}\Big)$$

برای محاسبه‌ی $\hat{\mu}_{ML}$ داریم:

$$\hat{\mu}_{ML} = \underset{\mu}{argmax} \mathcal{L}(\mu, \sigma)$$

پس:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{i=1}^{n} \frac{2(\mathbf{x}_i-\mu)}{2\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} (\mathbf{x}_i - \mu) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

همچنین برای محاسبه‌ی $\hat{\sigma}^2_{ML}$ داریم:

$$\hat{\sigma}^2_{ML} = \underset{\sigma}{argmax} \mathcal{L}(\mu, \sigma)$$

پس

$$\frac{\partial \mathcal{L}}{\partial \sigma} = \sum_{i=1}^{n} \Big(-\frac{1}{\sigma} + \frac{(\mathbf{x}_i-\mu)^2}{\sigma^3}\Big) = 0 \Rightarrow \sum_{i=1}^{n} \Big(-\sigma^2 + (\mathbf{x}_i-\mu)^2\Big) = 0 \Rightarrow \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i-\mu)^2$$

و در نتیجه

$$p(x; \mu, \sigma) \sim \mathcal{N}(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML})$$

مزیت چنین روشهایی این است که تعداد پارامترهای استفاده شده در این مدل‌ها اندک است که باعث می‌شود تعداد داده‌های زیادی برای آموزش نیاز نداشته باشیم.
ایراد عمده‌ی این روش‌ها محدودیت به توزیع انتخاب شده است و نمی‌توانیم نوع توزیع را تغییر دهیم. پس اگر نوع توزیع داده‌ها به درستی تعیین نشود، عملکرد خوبی از این مدل‌ها شاهد نخواهیم بود.

۲.۲ ب

طبق تعریف، مقدار بایاس برابر است با:

$$bias(\hat{p}(x)) = \mathbb{E}[\hat{p}(x)] - p(x)$$

اولا داریم:

$$\mathbb{E}[\hat{p}(x)] = \mathbb{E}[\ \frac{H}{n}\sum_{i=1}^{n}\mathbb{1}_{(\mathbf{x}_i\in I_l)}\ ] = \frac{H}{n}\,\mathbb{E}[\ \sum_{i=1}^{n}\mathbb{1}_{(\mathbf{x}_i\in I_l)}\ ] \quad (۱)$$

$$= \frac{H}{n}\ \sum_{i=1}^{n}\mathbb{E}[\mathbb{1}_{(\mathbf{x}_i\in I_l)}] \quad (۲)$$

$$= \frac{H}{n}\ \sum_{i=1}^{n}P(\mathbf{x}_i \in I_l) \quad (۳)$$

$$= \frac{H}{n}\ \sum_{i=1}^{n}\int_{\frac{l-1}{H}}^{\frac{l}{H}} p(x)dx \quad (۴)$$

$$= \frac{H}{n}\ \sum_{i=1}^{n}[F(\frac{l}{H}) - F(\frac{l-1}{H})] \quad (۵)$$

$$= H[F(\frac{l}{H}) - F(\frac{l-1}{H})] \quad (۶)$$

$$= \frac{F(\frac{l}{H}) - F(\frac{l-1}{H})}{\frac{l}{H} - \frac{l-1}{H}} \quad (۷)$$

برای رسیدن به ۱ و ۲ از خواص امیدریاضی استفاده شد. همچنین برای رسیدن از ۲ به ۳ نیز از خاصیت امیدریاضی روی Indicator استفاده شد. از آنجایی که بازه‌ی I_l طبق تعریف شامل نقاط $\frac{l-1}{H}$ تا $\frac{l}{H}$ است، توانستیم از ۳ به ۴ برسیم.برای رسیدن به ۵ از تابع توزیع تجمعی کمک گرفتیم و با توجه به اینکه مقدار داخل براکت برای همه‌ی i ها یکسان بود $\sum$ را حذف نمودیم و عبارت را در n ضرب کردیم. به این ترتیب به ۶ رسیدیم. در نهایت برای رسیدن از ۶ به ۷ مقدار H را به مخرج کسر بردیم و فرمی استخراج کردیم که شبیه مشتق باشد تا بتوانیم از قضیه‌ی مقدار میانگین استفاده کنیم.
حال طبق قضیه‌ی مقدار میانگین یک x' در بازه‌ی I_l داریم که:

$$F'(x') = p(x') = \frac{F(\frac{l}{H}) - F(\frac{l-1}{H})}{\frac{l}{H} - \frac{l-1}{H}}$$

پس با استفاده از ۷ نتیجه می‌شود:

$$\mathbb{E}[\hat{p}(x)] = p(x')$$

حال اگر مجددا به تعریف بازگردیم:

$$\begin{aligned} bias(\hat{p}(x)) = \mathbb{E}[\hat{p}(x)] - p(x) &= p(x') - p(x) \\ &= \frac{p(x') - p(x)}{x' - x}(x' - x) \\ &\le |\frac{p(x') - p(x)}{x' - x}|.|(x' - x)| \\ &\le \beta.\frac{1}{H} \\ &= \frac{\beta}{H} \end{aligned}$$

۳.۲ پ

از این رابطه نتیجه می‌شود هر چه تعداد Bin ها بیشتر باشد، Bias محدودتر می‌شود. در واقع با میل کردن H به سمت ∞ مقدار Bias به صفر میل می‌کند. در واقع با افزایش تعداد Bin ها درون هر Bin نویز بیشتری وجود خواهد داشت و واریانس به شدت افزایش می‌یابد. اما Bias کاهش خواهد یافت.

وقتی H به اندازه‌ی کافی بزرگ باشد، درون هر Bin حداکثر یک داده خواهیم داشت و در نتیجه تعداد Bin ها از اردر تعداد داده‌ها خواهد بود. به همین دلیل این روش یک روش Non-Parametric است.

۴.۲ ت

روشی که در بخش الف معرفی شد به دلیل اینکه از یک توزیع که شکل آن از قبل مشخص است استفاده می‌کنیم، دارای عملکرد بسیار خوبی روی دادگان از آن توزیع می‌باشد. به عبارتی اگر انتخاب شکل توزیع داده‌ها به خوبی انجام شده باشد، انتظار عملکرد بسیار خوبی خواهیم داشت. از طرف دیگر همانطور که اشاره شد اگر تعداد داده‌های کمی برای آموزش داشته باشیم، این روش مفیدتر است و به دلیل تعداد پارامترهای کمتر می‌تواند داده‌ها را بهتر مدل کند. در حالی که روش‌های ناپارامتری ممکن است روی تعداد داده‌های محدود Overfit شده و تعمیم‌پذیری خوبی نداشته باشد.

از طرف دیگر، اگر تعداد بسیاز زیادی داده داشته باشیم، از روش قسمت ب عملکرد بهتری انتظار خواهیم داشت. زیرا وقتی داده‌های زیادی در اختیار داشته باشیم، شکل کلی هیستوگرام در صورت انتخاب مناسب H به شکل توزیع نزدیک‌تر می‌شود و می‌تواند به خوبی آن را مدل کند. همچنین اگر انتخاب شکل کلی توزیع ممکن نباشد و یا به سختی بتوان شکل داده‌ها را به یک توزیع نسبت داد(که معمولا در مسائل دنیای واقع به همین شکل است) روش‌های ناپارامتری انتخاب بهتری هستند.

۳ سوال ۳

۱.۳ آ

در صورتی که حداقل نصف دسته‌بند ها برچسب داده را به درستی پیش بینی کنند، در این صورت الگوریتم به درستی پاسخ خواهد داد و در غیر اینصورت پاسخ اشتباه خواهد داد. به شکل دقیق تر فرض کنیم برچسب داده $+1$ باشد. همچنین فرض کنیم M دسته‌بند برچسب $+1$ و $T-M$ دسته‌بند برچسب -1 پیش‌بینی کنند. در اینصورت خواهیم داشت:

$$H(x) = sign((M)(+1) + (T-M)(-1)) = sign(2M - T)$$

زمانی برچسب صحیح پیش‌بینی می‌شود که:

$$sign(2M-T) = +1 \Rightarrow 2M - T \geq 0 \Rightarrow M \geq \frac{T}{2}$$

که این به معنی درست بودن پیش‌بینی حداقل نصف دسته‌بندها است. به طرز مشابه اگر برچسب داده‌ی x برابر -1 نیز باشد، به همین نتیجه می‌رسیم.

۲.۳ ب

طبق نامساوی هافدینگ اگر یک متغیر تصادفی از توزیع برنولی با پارامتر ϵ داشته باشیم، احتمال آنکه در n آزمایش مستقل حداکثر $k = (\epsilon + p)n \ , p > 0$ بار موفق شویم برابر است با

$$P(H(n) \leq (\epsilon + p)n) \leq \exp\bigl(-2p^2 n\bigr)$$

از آنجایی که هر دسته‌بند به شکل مستقل، دارای احتمال خطای ϵ است، پس احتمال اینکه در T دسته بند، حداقل نصف آن‌ها دچار خطا شوند برابر است با:

$$P(H(T) \geq \frac{1}{2}T) = P(H(T) \geq (\epsilon + (\frac{1}{2} - \epsilon))T) \leq \exp\left(-2(\frac{1}{2} - \epsilon)^2 T\right)$$

پس

$$\lim_{T\to\infty} P(H(T) \geq \frac{1}{2}T) \leq \lim_{T\to\infty} \exp\left(-2(\frac{1}{2} - \epsilon)^2 T\right) = 0$$

پس با افزایش T احتمال خطای کل دسته‌بندی به صفر میل می‌کند.

۴ سوال ۴

نوت‌بوک خواسته شده در قالب فایل *ML2021S_HW7* به پیوست تقدیم شده است.