دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین بهار ۱۴۰۰

پاسخ تمرین سری هفتم

مدرس: دکتر محمّدحسین رهبان

پاسخ سوال ۱ *AdaBoost Algorithm*

آ) می‌دانیم که در الگوریتم *Adaboost* از فرض یادگیری ضعیف استفاده می‌کنیم. به این معنا که در هرمرحله دسته‌بند $h \in H$ وجود دارد که خطای آن مطابق توزیع آن مرحله، کمتر از $\frac{1}{2}$ است. حال تلاش می‌کنیم خطای h_t را در مرحله‌ی $t+1$، یعنی مطابق توزیع D_{t+1} پیدا کنیم.

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

حال چون خطای آن $\frac{1}{2}$ شد، در نتیجه این دسته‌بند نمی‌تواند در مرحله‌ی $t+1$ انتخاب شود.

ب) با ضرب داخلی این دو بردار داریم:

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

که Z_t همان ضریب نرمال‌سازی است. دقت کنید در هر مرحله، α_t را طوری انتخاب می‌کردیم که نسبت به Z_t کمینه باشد. پس منطقا برابری $\frac{dZ_t}{d\alpha_t} = 0$ برقرار است. درنتیجه ضرب داخلی این دو بردار صفر است.

پاسخ سوال ۲ *Density Estimation with Non–Parametric Methods*

آ) وقتی از روش پارامتری استفاده کنیم، به این معناست که یک مدل پارامتری برای داده‌ها در نظر می‌گیرم. حال یکی از ساده‌ترین رویکردهای پارامتری که می‌توانیم درنظر بگیریم این است که داده‌ها را روی چند مدل پارامتری فیت کنیم و بیشینه درست‌نمایی را برای هر مدل بدست آوریم. سپس مدلی را انتخاب کنیم که بیشینه درست‌نمایی آن از باقی مدل‌ها بیشتر است. همانطور که مشخص است این روش معایب زیادی

دارد؛ ممکن است توزیع به اشتباه انتخاب شود و وابستگی بسیار زیادی به داده‌های ورودی دارد. همچنین ممکن است هزینه‌ی محاسبات آن بسیار زیاد باشد. حال اگر بخواهیم درمورد مزایای این روش صحبت کنیم، فیت کردن توزیع به داده این امکان را برای ما به وجود می‌آورد که بتوانیم آماره‌های مختلفی را محاسبه کنیم. مشخصا روش‌های متفاوتی می‌توان پیشنهاد داد و روش گفته شده تنها پیشنهاد مورد قبول نیست.

ب)

$$\begin{aligned}
\mathbb{E}\left[\widehat{p}_n(x)\right] &= H \cdot P\left(X_i \in B_\ell\right) \\
&= H \int_{\frac{\ell-1}{H}}^{\frac{\ell}{H}} p(u) du \\
&= H\left(F\left(\frac{\ell}{H}\right) - F\left(\frac{\ell-1}{H}\right)\right) \\
&= \frac{F\left(\frac{\ell}{H}\right) - F\left(\frac{\ell-1}{H}\right)}{1/H} \\
&= \frac{F\left(\frac{\ell}{H}\right) - F\left(\frac{\ell-1}{H}\right)}{\frac{\ell}{H} - \frac{\ell-1}{H}} \\
&= p\left(x^*\right), \quad x^* \in \left[\frac{\ell-1}{H}, \frac{\ell}{H}\right]
\end{aligned}$$

دقت کنید در خط آخر از قضیه‌ی مقدار میانی استفاده کردیم! حال مطابق این قضیه نقطه‌ی x^{**} بین x و x^* وجود دارد به گونه‌ای که:

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

حال برای محاسبه‌ی میزان اریبی داریم:

$$\begin{aligned}
\operatorname{bias}\left(\widehat{p}_n(x)\right) &= \mathbb{E}\left[\widehat{p}_n(x)\right] - p(x) \\
&= p\left(x^*\right) - p(x) \\
&= p'\left(x^{**}\right) \cdot \left(x^* - x\right) \\
&\leq \left|p'\left(x^{**}\right)\right| \cdot \left|x^* - x\right| \\
&\leq \frac{\beta}{H}
\end{aligned}$$

پ) درواقع پیدا کردن حد بالا برای میزان اریبی یک نتیجه بسیار مهم است. زیرا این اطمینان حاصل می‌شود که میزان اریبی از یک حدی بالاتر نخواهد شد و همچنین با افزایش H این حد کمتر نیز می‌شود. همچنین این روش مشخصا ناپارامتری است زیرا هیچ مدلی روی داده‌ها فرض نمی‌شود و پارامترهای مساله با افزایش داده تغییر می‌کنند.

ت) این روش علیرغم محدودیت‌هایی که به همراه دارد ولی به صورت مجانبی، ناریب است. می‌دانیم که برای مساله‌ی پارامتری این تضمین وجود نخواهد داشت. همچنین می‌توان حد بالایی برای واریانس این تخمینگر نیز محاسبه کرد که درنتیجه یک حد بالا برای خطای تخمینگر داریم و بنابراین مساله واگرا نخواهد شد.

پاسخ سوال ۳ *Ensemble Learning*

آ) هرگاه بیشتر از نصفی از دسته‌بندها اشتباه تشخیص دهند، آنگاه H دچار خطا می‌شود.

ب) با استفاده از نامساوی هافدینگ برای توزیع دوجمله‌ای داریم:

$$P(H(\boldsymbol{x}) \neq f(\boldsymbol{x})) = \sum_{k=0}^{\lfloor T/2 \rfloor} \begin{pmatrix} T \\ k \end{pmatrix} (1-\epsilon)^k \epsilon^{T-k} \leq \exp\left(-\frac{1}{2}T(2\epsilon-1)^2\right)$$

دقت کنید برای رخ‌دادن خطا، حداکثر تعداد دسته‌بندهایی که می‌توانند درست پیش‌بینی کرده باشند $\lfloor T/2 \rfloor$ است. پس از تعداد 0 تا این عدد جمع زده‌ایم. همچنین با توجه به فرم تابع به دست آمده، مشخص است با افزایش T این تابع به سمت صفر میل می‌کند.

پاینده باشید