دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین زمستان ۱۳۹۹

پاسخ تمرین سری اول

مدرس: دکتر محمّدحسین رهبان

پاسخ سوال ۱ اگر فرض کنیم که 2^n تیم در این لیگ وجود دارد در آن صورت n دور مسابقات در این لیگ داریم. در دور اول 2^{n-1} بازی انجام می‌شود در دور دوم 2^{n-2} بازی تا دور nام که 1 بازی خواهد داشت. پس تعداد کل بازی‌ها برابر $2^{n-1} + 2^{n-2} + \cdots + 2 + 1 = 2^n - 1$ است. حال یک بازی به خصوص مانند g را در نظر بگیرید. I_g متغیری است که مشخص می‌کند فرد امتیاز این بازی را می‌گیرد یا خیر. اگر این بازی در دور $r = r(g)$ام انجام بگیرد، میزان امتیازی که این فرد کسب خواهد کرد برابر 2^{r-1} است. پس به طور متوسط امتیاز کسب شده از این بازی برابر است با:

$$Excepted\,points\,g = E[2^{r-1}.I_g] = 2^{r-1}E[I_g] = 2^{r-1}P_g$$

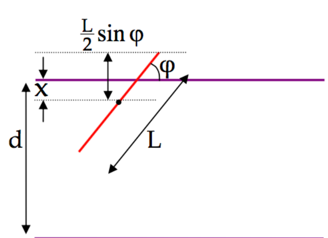
که P_g احتمال تشخیص برنده این بازی است. تشخیص درست برنده این بازی به منزله تشخیص درست نتیجه تمام بازی‌های قبلی آن تیم برنده در دور rام است. پس $P_g = 2^{-r}$ و

$$Excepted\,points\,g = E[2^{r-1}.I_g] = 2^{r-1}E[I_g] = 2^{r-1}P_g = 2^{r-1}2^{-r} = \frac{1}{2}$$

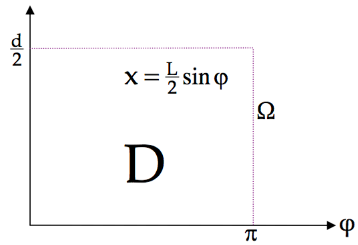
که مستقل از بازی است. امید ریاضی امتیاز کسب شده از هر بازی برابر $\frac{1}{2}$ می‌باشد. با توجه به این که $2^n - 1$ بازی در این لیگ داریم، امید ریاضی امتیاز فرد برابر با $\frac{2^n-1}{2}$ خواهد بود. در این لیگ 64 تیم داریم یعنی $n = 6$ پس :

$$Expected\,points = \frac{2^6 - 1}{2} = \frac{63}{2} = 31.5$$

پاسخ سوال ۲ فرض کنید که فاصله سوزن از نزدیک‌ترین خط x باشد و ϕ زاویه با این خط باشد. در صورت تقاطع خواهیم داشت $x < \frac{L}{2}\sin(\phi)$ که مقادیر ممکنه برای x عبارت است از $0 < x < \frac{d}{2}$. برای $x > \frac{d}{2}$ خط دیگر نزدیک‌تر خواهد بود مساله عینا تکرار می‌شود.



برای ϕ نیز مقادیر ممکنه برابر هستند با $0 < \phi < \pi$. در نتیجه احتمال را روی فضای زیر باید حساب کنیم:



$$P = \int\int_D \frac{1}{\pi\frac{d}{2}} dx d\phi = \frac{2}{\pi d}\int_0^{\pi} \frac{L}{2}\sin(\phi) d\phi = \frac{2L}{\pi d}$$

پاسخ سوال ۳ آ)

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} xf(x)dx = \int_0^{\alpha} xf(x)dx + \int_{\alpha}^{\infty} xf(x)dx \geq \int_{\alpha}^{\infty} xf(x)dx \geq \int_{\alpha}^{\infty} \alpha f(x)dx$$

$$\Rightarrow E[X] \geq \alpha \int_{\alpha}^{\infty} f(x)dx = \alpha P(X \geq \alpha) \Rightarrow P(X \geq \alpha) \leq \frac{E[X]}{\alpha}$$

ب) اگر Z یک متغیر تصادفی دلخواه باشد می‌توان متغیر تصادفی نامنفی X را به صورت زیر تعریف کرد:

$$X = (Z - \mu)^2$$

در آن صورت باتوجه به بخش آ، داریم:

$$P(X \geq \alpha) = P((Z-\mu)^2 \geq \alpha) \leq \frac{E[(Z-\mu)^2]}{\alpha} = \frac{\sigma^2}{\alpha} \Rightarrow P(|Z-\mu| \geq \sqrt{\alpha}) \leq \frac{\sigma^2}{\alpha}$$

با قرار دادن $\epsilon = \sqrt{\alpha}$ به رابطه مورد نظر می‌رسیم.

$$P(|Z-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

پ) ۱) فرض کنید که n نقطه تصادفی ایجاد کرده‌ایم. در این‌صورت تخمین‌گر ما برای عدد π متغیر تصادفی $M_n = \frac{4}{n}\sum_{i=1}^{n} x_i$ است که در آن x_i متغیر تصادفی است که نشان می‌دهد نقطه iام درون دایره قرار گرفته است یا خیر. x_i یک متغیر تصادفی برنولی با احتمال $\frac{\pi}{4}$ است. امید ریاضی و واریانس یک متغیر تصادفی برنولی با احتمال p به ترتیب برابر p و $p(1-p)$ است. پس:

$$E[M_n] = E[\frac{4}{n}\sum_{i=1}^{n} x_i] = \frac{4}{n}\sum_{i=1}^{n} E[x_i] = \frac{4}{n}\sum_{i=1}^{n}\frac{\pi}{4} = \frac{4}{n}n\frac{\pi}{4} = \pi$$

$$Var[M_n] = Var[\frac{4}{n}\sum_{i=1}^{n} x_i] = (\frac{4}{n})^2\sum_{i=1}^{n} Var[x_i] = (\frac{4}{n})^2\sum_{i=1}^{n}\frac{\pi}{4}(1-\frac{\pi}{4}) = \frac{\pi(4-\pi)}{n}$$

حال طبق قضیه چبیشف

$$P(|M_n - E[M_n]| \geq 0.01) \leq \frac{Var(M_n)}{0.01^2} \Rightarrow P(|M_n - \pi| \geq 0.01) \leq \frac{\pi(4-\pi)}{0.01^2 n} \leq 0.05 \Rightarrow n \geq 539354$$

از آن‌جایی که مسئله، مسئله تخمین مقدار π است نمی‌توانیم از مقدار آن به صورت مستقیم استفاده کنیم. ولی با توجه به تابع $f(x) = x(4-x) \leq 4$ است می‌توانیم بنویسیم :

$$P(|M_n - \pi| \geq 0.01) \leq \frac{4}{0.01^2 n} \leq 0.05 \Rightarrow n \geq 800000$$

(در اینجا به هر دو مقدار نمره کامل تعلق گرفته است)

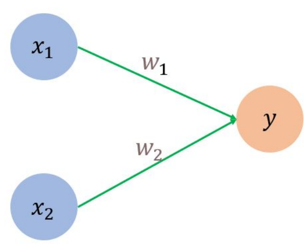
۲)

$$P(|M_n - \pi| \geq 0.01) = P(|\frac{M_n}{4} - \frac{\pi}{4}| \geq 0.0025) \leq 2e^{-n\frac{2.0.0025^2}{(1-0)^2}} \leq 0.05 \Rightarrow n \geq -\ln(\frac{0.05}{2}.\frac{1}{2.0.0025^2})$$

$$\Rightarrow n \geq 295111$$

پاسخ سوال ۴ آ) مدل‌های پرسپترونی مورد نظر مطابق شکل دارای دو ورودی دودویی هستند. در این خروجی را می‌توان به شکل زیر نوشت:

$$y^{(i)} = sgn(w_1 x_1^{(i)} + w_2 x_2^{(i)} + b)$$



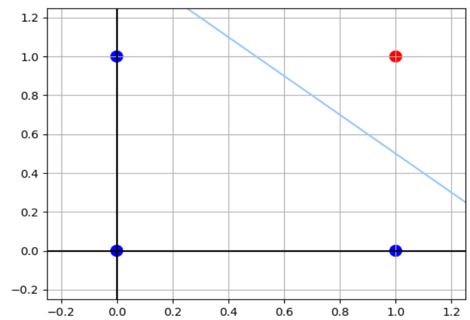
در دو نمودار زیر، خروجی‌های توابع AND و OR به ازای ورودی‌های ممکن رسم شده‌است. نقاط با رنگ قرمز، برچسب $+1$ و نقاط آبی برچسب‌های 0 را نشان می‌دهند.خط تفکیک کننده را باید به نحوی کشید تمام نقاط قرمز در یک سمت آن و تمام نقاط آبی در سمت دیگر آن باشد. یک نمونه این خط در شکل‌ها رسم شده است.

$$AND: w_1 = 1, w_2 = 1, b = 1.5 \Rightarrow y = sgn(x_1^{(i)} + x_2^{(i)} + 1.5)$$

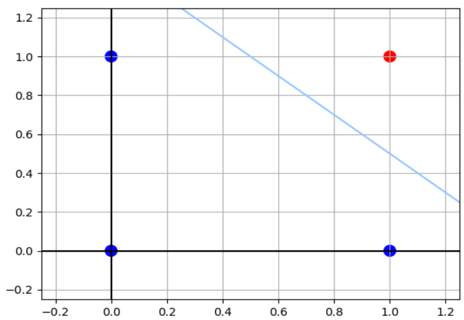
$$OR: w_1 = 1, w_2 = 1, b = 0.5 \Rightarrow y = sgn(x_1^{(i)} + x_2^{(i)} + 0.5)$$

ب) برای تابع XOR همانطور که از شکل مشخص است، نمی‌توان آن را با یک خط به نحوی تفکیک کرد که خطای مدل‌سازی صفر باشد. در ادامه به صورت ریاضی ناممکن بودن این مسئله را نشان می‌دهیم.

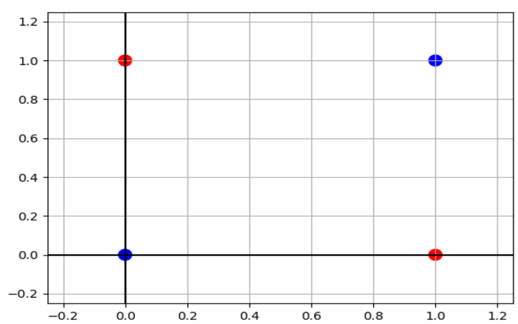
$$y = f(x_1, x_2) = sgn(w_1 x_1 + w_2 x_2 + b)$$



(آ) خروجی تابع AND



(ب) خروجی تابع OR



شکل ۲: خروجی تابع XOR

$$0 = f(0,0) = sgn(0+0+b) = sgn(b) \Rightarrow b < 0$$

$$1 = f(1,0) = sgn(w_1+0+b) = sgn(w_1+b) \Rightarrow w_1 + b > 0$$

$$1 = f(0,1) = sgn(0+w_2+b) = sgn(w_2+b) \Rightarrow w_2 + b > 0$$

$$0 = f(1,1) = sgn(w_1+w_2+b) \Rightarrow w_1 + w_2 + b < 0$$

اگر معادلات دوم و سوم را با هم جمع کنیم:

$$w_1 + b + w_2 + b > 0 \Rightarrow w_1 + w_2 + b > -b > 0$$

که این نتیجه با معادله چهارم در تناقض است. پس هیچ تابع پرسپترونی وجود ندارد که بتواند تابع منطقی XOR را با خطای صفر مدل‌سازی کند.

پ)

$$\langle w^*, w_k\rangle = \langle w^*, (w_{k-1} + y_k x_k)\rangle = \langle w^*, w_{k-1}\rangle + y_k\langle w^*, x_k\rangle \geq \langle w^*, w_{k-1}\rangle + \rho \geq \langle w^*, w_{k-2}\rangle + 2\rho \geq \cdots$$

$$\geq \langle w^*, w_1\rangle + k\rho = k\rho \Rightarrow \langle w^*, w_k\rangle \geq k\rho$$

به روزرسانی وزن زمانی انجام می‌گیرد که یک داده با برچسب داشته باشیم. پس:

$$||w_k||^2 = ||w_{k-1} + y_k x_k||^2 = ||w_{k-1}||^2 + 2y_k\langle w_{k-1}, x_k\rangle + ||y_k||^2||x_k||^2 = ||w_{k-1}||^2 + 2y_k\langle w_{k-1}, x_k\rangle + ||x_k||^2$$

$$\leq ||w_{k-1}||^2 + ||x_k||^2 \leq ||w_{k-1}||^2 + r^2 \leq ||w_{k-2}||^2 + 2r^2 \leq \cdots \leq ||w_0||^2 + kr^2 = kr^2 \Rightarrow ||w_k||^2 \leq kr^2$$

حال زاویه میان دوبردار w^* و w_k را محاسبه می‌کنیم.

$$\cos(w^*, w_k) = \frac{\langle w^*, w_k\rangle}{||w^*||.||w_k||} \geq \frac{k\rho}{||w^*||.||w_k||} \geq \frac{k\rho}{||w^*||.r\sqrt{k}} = \frac{\sqrt{k}\rho}{||w^*||.r}$$

$$\cos(w^*, w_k) \leq 1 \Rightarrow \frac{\sqrt{k}\rho}{||w^*||.r} \leq 1 \Rightarrow k \leq ||w^*||(\frac{r}{\rho})^2$$

با توجه به ویژگی‌های تابع sgn، همواره می‌توان وزن‌های مدل را در یک ضریب ثابت مثبت ضرب کرد، بدون آنکه مدل پرسپترون تغییری کند. به همین خاطر همواره می‌توان بردار وزن را به ۱ نرمالیزه کرد. پس :

$$k \leq (\frac{r}{\rho})^2$$

پاسخ سوال ۵ آ) A_1 طبق گفته سوال، هموراه فرضیه‌ای را انتخاب می‌کند که روی مجموعه دادگان خطای کمتری داشته باشد. به ازای هر تعدادی نمونه که از کل فضا نمونه برداری شده باشد، احتمالی وجود دارد که نمونه‌های وارد شده به $\mathcal{D}$ به درستی نمایانگر توزیع روی $\mathcal{X}$ نباشند. به عنوان مثال فرض کنید که $p = 0.9$ است. در این صورت احتمال کمی وجود دارد که اکثریت نمونه‌های داخل مجموعه دادگان -1 باشد. با افزایش تعداد نمونه‌ها این احتمال کم می‌شود اما همواره ناصفر است و در حدهای به سمت بی‌نهایت مقدار آن به صفر میل می‌کند (یعنی بینهایت نمونه از $\mathcal{X}$ گرفته باشیم). در این شرایط A_1 فرضیه h_2 را انتخاب می‌کند که عملکردی بدتر از انتخاب تصادفی برای نقاط خارج دارد. به همین خاطر این الگوریتم هیچ تضمینی برای عملکرد بهتر از $random$ روی فضای خارج نمی‌دهد.

ب) مطابق پاسخ قسمت قبل، هیچ تضمینی برای این‌که نمونه‌های داخل $\mathcal{D}$ به درستی نمایانگر توزیع روی $\mathcal{X}$ باشند وجود ندارد به همین خاطر ممکن است غالب نمونه‌های $\mathcal{X}$ برچسب -1 داشته باشند (یعنی $p < 0.5$) اما این مسئله در مجموعه دادگان ما صدق نکند. در این شرایط A_2 فرضیه‌ای با خطای کمتر از A_1 ایجاد می‌کند.

پ) چون فرض شده است که تمام نمونه‌های داخل مجموعه دادگان دارای برچسب $+1$ هستند A_1 همواره فرضیه h_1 را انتخاب می‌کند که برای نقاط خارج از $\mathcal{D}$ نیز خطای آماری 0.1 را به همراه خواهد داشت که از خطای آماری A_2 (0.9) کمتر است. به همین خاطر A_1 در این شرایط همواره فرضیه‌ی بهتری تولید می‌کند.

ت) به ازای $p < 0.5$ احتمال کمی وجود دارد که تمام دادگان $\mathcal{D}$ دارای برچسب $+1$ باشند در این شرایط A_1 همواره فرضیه h_1 را انتخاب می‌کند اما A_2 فرضیه‌ای را انتخاب می‌کند که روی دادگان واقعی خطای کمتری دارد.

پاینده باشید