بسمه تعالی

پاسخ سری اول تمرینات درس یادگیری ماشین

امیرحسین رمضانی بناب (۹۹۲۱۰۲۹۴)

۱ سوال اول

بدون کاسته شدن از کلیت فرض میکنیم در صورتی که خروجی پرتاب سکه H باشد پیش بینی برد بازیکن اول، و در صورت خروجی T، پیش بینی برد بازیکن دوم آن بازی است.
برای پاسخ به این سوال ابتدا ثابت می‌کنیم احتمال پیش بینی صحیح **یک** بازی در دور i برابر $\frac{1}{2^i}$ است. بدین منظور از استقرا استفاده می‌کنیم.

- پایه‌ی استقرا : احتمال پیش بینی صحیح هر بازی در دور اول برابر $\frac{1}{2}$ است : برای اثبات این ادعا فضای پیشامد برای هر بازی را به شکل زیر در نظر می‌گیریم.

$$\Omega = \{(P_1, H), (P_1, T), (P_2, H), (P_2, T)\}$$

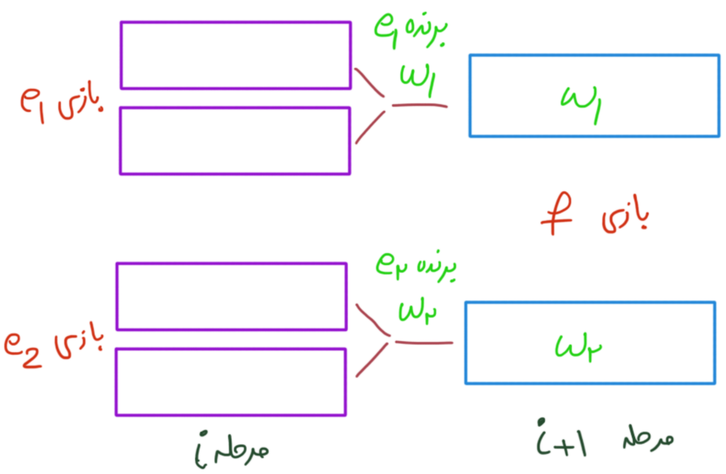
در هر زوج مرتب، مولفه‌ی اول نشان دهنده‌ی بازیکن برنده بازی و مولفه‌ی دوم نشان دهنده‌ی نتیجه‌ی پرتاب سکه است. حالت هایی که پیش بینی به درستی انجام می‌شود به شکل زیر است:

$$A = \{(P_1, H), (P_2, T)\}$$

پس احتمال پیش بینی صحیح هر بازی برابر $\frac{|A|}{|\Omega|}$ است که مساوی عدد $\frac{1}{2}$ است.

- فرض استقرا : احتمال پیش بینی صحیح یک بازی در دور i برابر $\frac{1}{2^i}$ است.

- حکم استقرا : احتمال پیش بینی صحیح یک بازی در دور $i+1$ برابر $\frac{1}{2^{i+1}}$ است :برای اثبات این حکم ابتدا یک بازی در دور $i+1$ در نظر می‌گیریم و آن را بازی f می‌نامیم. همچنین فرض می‌کنیم در جدول بازی ها تیم هایی که بازی f را انجام می‌دهند برندگان بازی e_1 و e_2 هستند که این بازیکنان را نیز w_1 و w_2 می‌نامیم که به ترتیب برنده‌ی بازی e_1 و e_2 هستند که در بازی f به مصاف هم می‌روند.



اگر w_1 برنده‌ی بازی f باشد فقط در دو حالت ممکن است پیش‌بینی‌کننده به درستی این پیش‌بینی را انجام داده باشد.

حالت اول : بازی e_1 و e_2 به درستی پیش‌بینی شده باشند و بازی f هم به درستی پیش‌بینی شود. احتمال این رویداد طبق فرض استقرا برابر عدد زیر است

$$\frac{1}{2^i} \times \frac{1}{2^i} \times \frac{1}{2}$$

حالت دوم : بازی e_1 به درستی پیش‌بینی شده باشد ولی پیش‌بینی بازی ۲ درست نباشد. همچنین پیش‌بینی بازی f درست باشد و بازیکن w_1 در آن بازی برنده شده باشد. احتمال چنین رویدادی طبق فرض استقرا برابر عدد زیر است.

$$\frac{1}{2^i} \times (1 - \frac{1}{2^i}) \times \frac{1}{2}$$

پس در کل در حالتی که w_1 برنده‌ی بازی f شود، به احتمال زیر پیش بینی درست است.

$$\frac{1}{2^i} \times \frac{1}{2^i} \times \frac{1}{2} + \frac{1}{2^i} \times (1 - \frac{1}{2^i}) \times \frac{1}{2} = \frac{1}{2^{i+1}}$$

همچنین در حالتی که w_2 برنده‌ی بازی شود با استدلالی مشابه به همین عدد می‌رسیم.

حال اگر احتمال پیش بینی صحیح را با $P(TruePredict)$ نمایش دهیم و احتمال برنده شدن بازیکن w_i در بازی f را با $P(Wins_i)$ نمایش دهیم داریم :

$$P(TruePredict) = P(TruePredict|Wins_1)P(Wins_1) + P(TruePredict|Wins_2)P(Wins_2)$$

که با جایگذاری $P(Wins_i) = \frac{1}{2}$ و $P(TruePredict|Wins_i) = \frac{1}{2^{i+1}}$ داریم :

$$P(TruePredict) = \frac{1}{2^{i+1}} \times \frac{1}{2} + \frac{1}{2^{i+1}} \times \frac{1}{2} = \frac{1}{2^{i+1}}$$

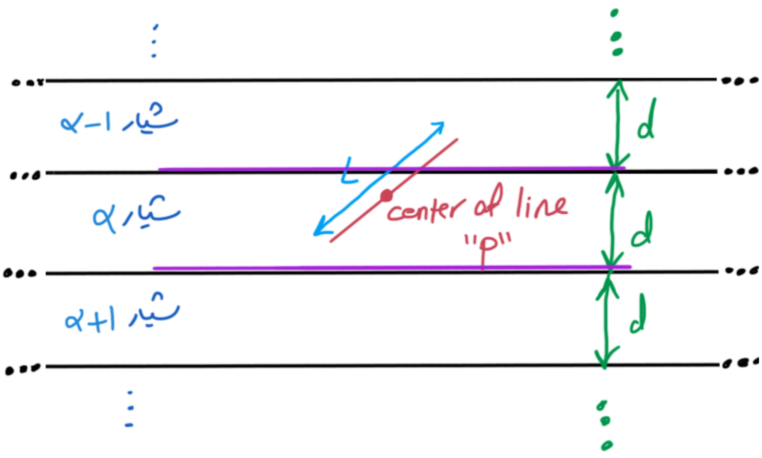
پس حکم استقرا اثبات شد.

حال کافی است میزان سود $Expected$ پیش‌بینی‌کننده را محاسبه کنیم.
از آنجایی که در مرحله‌ی i ام $\frac{64}{2^i}$ بازی انجام می‌شود (n_i) و میزان سود متوسط از هر بازی در دور i ام برابر $\frac{1}{2^i}$ به دست آمد (e_i). پس میزان سود عایدی پیش‌بینی‌کننده در نهایت برابر است با :

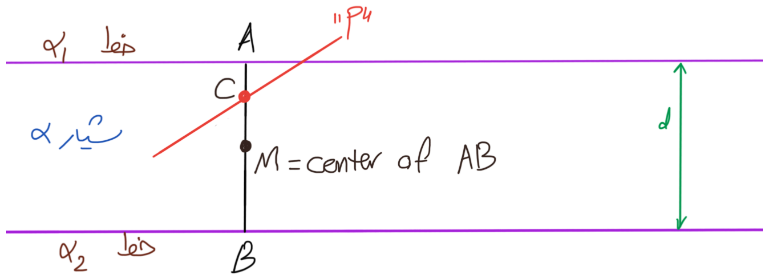
$$E = \sum_{i=1}^{6} n_i p_i = \sum_{i=1}^{6} \frac{64}{2^i} \times \frac{1}{2^i} = 64 \sum_{i=1}^{6} \frac{1}{2^{i+1}} = 64(\frac{1}{2}(1 - \frac{1}{64})) = 31.5$$

۲ سوال دوم

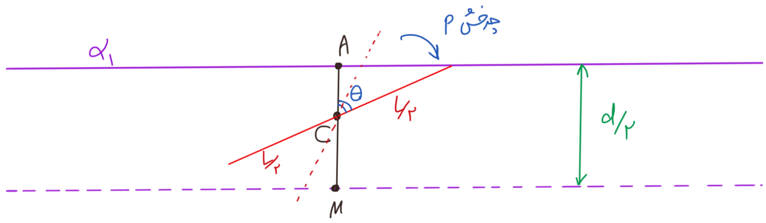
برای پاسخ به این سوال ابتدا سوال را ساده سازی می‌کنیم تا به مسئله‌ی ساده تری برسیم و آن را حل کنیم. صفحه را به شکل شیارهایی به ارتفاع d در نظر می‌گیریم که مرز بین هر دو شیار، خطوط موازی به فاصله d باشند. همچنین سوزن به طول L را معادل خط p در نظر می‌گیریم که مرکز آن در شیار α قرار دارد. به شکل زیر :



حال اگر p یکی از خطوط موازی روی کاغذ را قطع کند، معادل این است که یکی از دو خط موازی شیاری که مرکز p در آن دارد را قطع نماید. یعنی اگر خط p خطی از صفحه را قطع کند، یکی از خطوط شیار α که با رنگ بنفش مشخص شده‌اند را نیز قطع می‌کند و برعکس. در نتیجه به جای اینکه احتمال این را حساب کنیم که خط p خطی از صفحه را قطع کند، روی یک شیار α متمرکز می‌شویم و احتمال قطع کردن یکی از دو خط بنفش رنگ توسط خط p را محاسبه می‌کنیم.
حال از آنجایی که خطوط افقی را می‌توانیم تا بی نهایت امتداد دهیم محل افقی مرکز خط p در احتمال خواسته شده هیچ تاثیری ندارد و هر دو نقطه ای که روی یک خط افقی در شیار α باشند، احتمال یکسانی را تولید می‌کنند. به این خاطر می‌توانیم دوباره مسئله را ساده سازی کنیم و به یک خط عمودی روی شیار فکر کنیم:



در نهایت یک مرحله دیگر ساده سازی وجود دارد. احتمال آنکه خط p به مرکز C خطی از دو خط موازی تشکیل دهنده‌ی شیار α را قطع کند، معادل این احتمال است که خط p به مرکز C یکی از دو خط موازی تشکیل دهنده‌ی شیار α را قطع کند که به C نزدیک تر است. یعنی وقتی p خطی از شیار را قطع می‌کند، قطعا خط نزدیک تر را نیز قطع می‌کند و برعکس وقتی خط نزدیک تر را قطع می‌کند، خطی از شیار را قطع کرده است.
پس به جای یک شیار روی نصف شیار متمرکز می‌شویم و تحلیل نهایی را روی نیمه ای از خط AB به طول $\frac{d}{2}$ انجام می‌دهیم که مرکز خط p روی آن قرار گرفته. در اینجا مرکز p روی AM قرار گرفته است. پس روی آن متمرکز می‌شویم.
دو حالت وجود دارد که در این دو حالت پاسخ را به دست می‌آوریم:
حالت اول : $L < d$
در این حالت می‌خواهیم احتمال آنکه خط به مرکز C با فاصله‌ی x از A، خط α_1 را قطع کند را حساب کنیم.
تحلیل زیر را انجام می‌دهیم. برای حالتی که $x = |AC| < L/2$ ، با شروع از خط AC فرض کنیم بزرگترین زاویه‌ای که خط p خط α_1 را قطع کند، برابر θ باشد



که :

$$cos(\theta) = \frac{|AC|}{L/2} = \frac{x}{L/2} = \frac{2x}{L}$$

پس :

$$\theta = \cos^{-1}\frac{2x}{L}$$

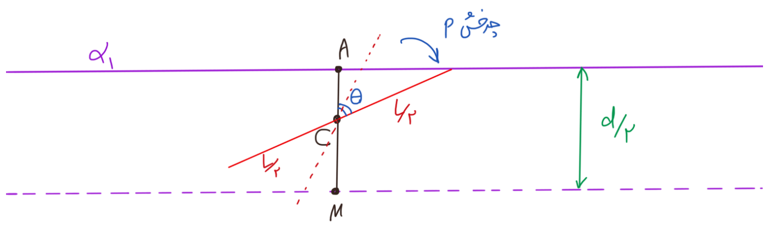
در اینصورت احتمال اینکه خط p خط α_1 را قطع کند برابر است با :

$$\frac{\theta}{\frac{\pi}{2}} = \frac{2\cos^{-1}\frac{2x}{L}}{\pi}$$

همچنین برای حالتی که $x = |AC| > L/2$ باشد چنین زاویه‌ای وجود ندارد و خط p امکان این را ندارد که α_1 را قطع کند. در نتیجه در $x > L/2$ این احتمال برابر صفر است. پس در کل احتمال اینکه خط p که مرکز آن روی AM است خط α_1 را قطع کند برابر است با :

$$\begin{aligned}
\frac{1}{\frac{d}{2}}\int_0^{L/2}\frac{2\cos^{-1}\frac{2x}{L}}{\pi}dx &= \frac{4}{d}\times\frac{1}{\pi}\int_0^{L/2}\cos^{-1}\frac{2x}{L}dx \\
&= \frac{4}{d}\times\frac{1}{\pi}\left(x\cos^{-1}(\frac{2x}{L}) - \frac{\sqrt{1-\frac{4}{L^2}x^2}}{\frac{2}{L}}\right)\Big|_0^{\frac{L}{2}} \\
&= \frac{4}{d}\times\frac{1}{\pi}(\frac{1}{\frac{2}{L}}) \\
&= \frac{2L}{\pi d}
\end{aligned}$$

حالت دوم : $L \geq d$ در این حالت می‌خواهیم احتمال آنکه خط به مرکز C با فاصله‌ی x از A، خط α_1 را قطع کند را حساب کنیم.



تحلیل زیر را انجام می‌دهیم. با شروع از خط CA فرض کنیم بزرگترین زاویه‌ای که خط p خط α_1 را قطع کند، برابر θ باشد که :

$$cos(\theta) = \frac{|AC|}{L/2} = \frac{x}{L/2} = \frac{2x}{L}$$

پس :

$$\theta = \cos^{-1}\frac{2x}{L}$$

در اینصورت احتمال اینکه خط p خط α_1 را قطع کند برابر است با :

$$\frac{\theta}{\frac{\pi}{2}} = \frac{2\cos^{-1}\frac{2x}{L}}{\pi}$$

پس احتمال اینکه خط p که مرکز آن روی AM است خط α_1 را قطع کند برابر است با :

$$\begin{aligned}
\frac{1}{\frac{d}{2}}\int_0^{d/2}\frac{2\cos^{-1}\frac{2x}{L}}{\pi}dx &= \frac{4}{d}\times\frac{1}{\pi}\int_0^{d/2}\cos^{-1}\frac{2x}{L}dx \\
&= \frac{4}{d}\times\frac{1}{\pi}\Big(x\cos^{-1}(\frac{2x}{L}) - \frac{\sqrt{1-\frac{4}{L^2}x^2}}{\frac{2}{L}}\Big)\Big|_0^{\frac{d}{2}} \\
&= \frac{4}{d}\times\frac{1}{\pi}\big[\big(\frac{d}{2}\cos^{-1}(\frac{d}{L}) - \frac{\sqrt{1-\frac{d^2}{L^2}}}{\frac{2}{L}}\big) - \big(-\frac{1}{\frac{2}{L}}\big)\big] \\
&= \frac{4}{d}\times\frac{1}{\pi}\big[\big(\frac{d}{2}\cos^{-1}(\frac{d}{L}) - \frac{\sqrt{L^2-d^2}}{2}\big) + \big(\frac{L}{2}\big)\big] \\
&= \frac{2}{\pi}\cos^{-1}(\frac{d}{L}) - 2\frac{\sqrt{L^2-d^2}}{d\pi} + \frac{2L}{d\pi}
\end{aligned}$$

پس در هر دو حالت، پاسخ نهایی به دست آمد.

۳ سوال سوم

۱.۳ قسمت آ

فرض کنید X یک متغیر تصادفی باشد که از یک توزیع با چگالی $f_X(x)$ و توزیع تجمعی $F_X(x)$ آمده باشد. برای اینکه اثبات کنیم $P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$. با فرض $\alpha > 0$ کافی است ثابت کنیم $\alpha P(X \geq \alpha) \leq E(X)$.
از سمت چپ معادله شروع می‌کنیم. داریم :

$$\begin{aligned}
\alpha P(X \geq \alpha) &= \alpha \int_{\alpha}^{\infty} f_X(x)dx && (۱)\\
&= \int_{\alpha}^{\infty} \alpha f_X(x)dx && (۲)\\
&\leq \int_{\alpha}^{\infty} x f_X(x)dx && (۳)\\
&\leq \int_{-\infty}^{\infty} x f_X(x)dx && (۴)\\
&= E(X) && (۵)
\end{aligned}$$

در ۱ از تعریف $P(X \geq \alpha)$ استفاده شد. برای رسیدن از ۱ به ۲ از خاصیت خطی بودن انتگرال استفاده شد. برای رسیدن از ۲ به ۳ از این ویژگی استفاده شد که x بین α و ∞ است. و در نهایت برای رسیدن از ۳ به ۴ از ویژگی مثبت بودن متغیر تصادفی X استفاده شد که عبارت ۴ نیز تعریف Expected Value است.پس با کنار هم قرار دادن این معادلات داریم :

$$\alpha P(X \geq \alpha) \leq E(X) \rightarrow P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$$

لطفا توجه شود که از فرض $\alpha > 0$ استفاده شد در حالی که این فرض جزو مفروضات سوال نبود. برای این مورد اگر فرض کنیم $\alpha < 0$ در آن صورت خواهیم داشت $P(X \geq \alpha) = 1$ و به دلیل مثبت بودن متغیر تصادفی X خواهیم داشت $\frac{E(X)}{\alpha} < 0$. پس در این حالت نامساوی برقرار نمی‌باشد.

نتیجه اینکه باید شرط $\alpha > 0$ نیز به شروط مسئله اضافه گردد.

۲.۳ قسمت ب

برای متغیر تصادفی دلخواه Z ، متغیر تصادفی $(Z-\mu)^2$ یک متغیر تصادفی نامنفی است. اگر در نامساوی مارکوف قرار دهیم $X = (Z-\mu)^2$ و $\alpha = \epsilon^2$ ، در آن صورت هر دو شرط نامساوی مارکوف (مثبت بودن X و α) برقرار است و داریم :

$$P((Z-\mu) \geq \epsilon^2) \leq \frac{E[(Z-\mu)^2]}{\epsilon^2} \tag{۶}$$

ابتدا طبق تعریف واریانس :

$$\sigma^2 = Var(Z) = E[(Z-E[Z])^2] = E[(Z-\mu)^2]$$

اگر این عبارت را در معادله ۶ قرار دهیم داریم معادله‌ی ۶ به شکل زیر ساده می‌شود :

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

همچنین عبارت $(Z-\mu)^2 \geq \epsilon^2$ فقط در صورتی برقرار است که $Z-\mu \geq \epsilon$ و یا $-(Z-\mu) \leq \epsilon$ که مجموع این دو عبارت معادل است با اینکه :

$$|Z-\mu| \geq \epsilon$$

پس با بازنویسی معادله طبق شرط ساده شده داریم :

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

که همان نامساوی چبیشف است.
در این قسمت هم باید نکته ای اشاره شود. در بخشی که از توان ۲ به قدر مطلق ساده سازی کردم این فرض باید وجود میداشت که $\epsilon > 0$ زیرا در غیر اینصورت اگر $\epsilon < 0$ می‌بود، در آن صورت $P((Z-\mu) \geq \epsilon) = 1$ است و در نتیجه نامساوی چبیشف به $1 \leq \frac{\sigma^2}{\epsilon^2}$ تبدیل می‌شود. در حالی که این نامساوی همیشه برقرار نیست.

نتیجه اینکه باید $\epsilon > 0$ جزو شروط مساله آورده می‌شد و در غیر اینصورت نامساوی برقرار نیست.

۳.۳ قسمت پ

۱.۳.۳ قسمت پ ۱

برای این قسمت از سوال متغیر تصادفی Z را به شکل زیر در نظر می‌گیریم

$$Z = \begin{cases} 1 & \text{if selected point be in the circle} \\ 0 & \text{if selected point be out of the circle} \end{cases}$$

از آنجایی که احتمال اینکه یک نقطه درون دایره بیفتد برابر تقسیم مساحت دایره بر مساحت مربع است پس هر نقطه به احتمال $p = \frac{\pi}{4}$ درون دایره می‌افتد و به احتمال $q = 1-p = 1-\frac{\pi}{4}$ بیرون آن. در نتیجه متغیر تصادفی Z به احتمال p مقدار 1 و به احتمال $q = 1-p$ مقدار 0 می‌گیرد. پس Z یک متغیر تصادفی برنولی است. در مورد متغیر تصادفی برنولی می‌دانیم :

$$E(Z) = p = \frac{\pi}{4}$$
$$Var(Z) = p(1-p) = \frac{\pi}{4}(1-\frac{\pi}{4})$$

ما به دنبال تخمین مساحت دایره هستیم. از آنجایی که مساحت دایره برابر احتمال انتخاب نقطه درون دایره ضرب در مساحت مربع است و چون مساحت مربع 4 است، پس ما به دنبال تخمین متغیر تصادفی $X = 4Z$ هستیم. طبق خواص Expected Value و Variance داریم :

$$E(X) = 4E(Z) = 4p = \pi$$
$$Var(X) = 4^2 Var(Z) = 16p(1-p) = \pi(4-\pi)$$

حال m سمپل از X برمی‌داریم و آن ها را $X_1, X_2, ..., X_m$ می‌نامیم. متغیر تصادفی $\bar{X}$ را به شکل زیر تعریف می‌کنیم :

$$\bar{X} = \frac{\sum_{i=1}^{m} X_i}{m} = \frac{X_1 + X_2 + ... + X_m}{m}$$

طبق خواص Expected Value و Variance داریم :

$$E(\bar{X}) = E(X) = \pi$$
$$Var(\bar{X}) = \frac{1}{m} Var(X) = \frac{\pi}{m}(4-\pi)$$

می‌خواهیم با استفاده از نامساوی چبیشف با خطای 0.01 و با قاطعیت 0.95 مقدار $\bar{X}$ را تخمین بزنیم. طبق این نامساوی داریم :

$$P(|\bar{X} - E(\bar{X})| \geq \epsilon) \leq \frac{Var(\bar{X})}{\epsilon^2}$$

با جایگذاری خطای تخمین مورد نظر ($\epsilon = 0.01$) داریم :

$$P(|\bar{X} - E(\bar{X})| \geq 10^{-2}) \leq \frac{Var(\bar{X})}{10^{-4}}$$

حال $E(\bar{X})$ و $Var(\bar{X})$ را جایگذاری می‌کنیم.

$$P(|\bar{X} - \pi| \geq 10^{-2}) \leq \frac{\pi(4-\pi)}{10^{-4}m} = 10^4 \frac{\pi(4-\pi)}{m}$$

از آنجایی که می‌خواهیم به قاطعیت ۹۵ درصد برسیم پس می‌خواهیم احتمال Bad Event حداکثر ۵ درصد شود. پس سمت راست این نامساوی را کمتر از 0.05 قرار می‌دهیم. داریم :

$$10^4 \frac{\pi(4-\pi)}{m} \leq 5 \times 10^{-2}$$

پس

$$m \geq \frac{1}{5} \times 10^6 \pi(4-\pi) = 539353.2426...$$

در نتیجه برای رسیدن به این قاطعیت و دقت باید حداقل ۵۳۹،۳۵۴ نمونه طبق نامساوی چبیشف داشته باشیم. البته این نامساوی باند خوبی ارائه نمیدهد و با تعداد کمتری نمونه نیز می‌توان به همین دقت رسید که توسط نامساوی هافدینگ به دست خواهد آمد.

۲.۳.۳ قسمت پ ۲

با همان فرمولە بندی ارائه شده در بخش قبلی پیش می‌رویم. **(طبق برگه ارائه شده مرور نامساوی های احتمالاتی)** نامساوی هافدینگ به شکل زیر تعریف می‌شود :
اگر $Z_1, Z_2, ..., Z_m$ دنباله‌ای از متغیرهای تصادفی $i.i.d$ باشند و $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ ، اگر $E[\bar{Z}] = \mu$ و $P(a \leq Z_i \leq b) = 1$ در اینصورت برای هر $\epsilon > 0$ داریم :

$$P[|\bar{Z} - \mu| > \epsilon] \leq 2\exp(\frac{-2M\epsilon^2}{(b-a)^2}) \qquad (۷)$$

حال طبق تعریف متغیر تصادفی X مقدار آن همواره یا 0 و یا 4 است. پس داریم :

$$P(0 \leq X_i \leq 4) = 1$$

با جایگذاری a و b و $E(\bar{X})$ در نامساوی ۷ داریم :

$$P[|\bar{X} - \pi| > \epsilon] \leq 2\exp(\frac{-2M\epsilon^2}{(4-0)^2}) = 2\exp(\frac{-M\epsilon^2}{8})$$

حال چون می‌خواهیم دقت تخمین ۱ درصد باشد قرار می‌دهیم $\epsilon = 0.01$ و داریم :

$$P[|\bar{X} - \pi| > 0.01] \leq 2\exp(\frac{-M \times 10^{-4}}{8})$$

چون می‌خواهیم احتمال Bad Event حداکثر ۵ درصد باشد قرار می‌دهیم :

$$2\exp(\frac{-M \times 10^{-4}}{8}) \leq 0.05$$

یعنی :

$$\exp(\frac{-M \times 10^{-4}}{8}) \leq 0.025$$

برای حل این نامساوی از دو طرف ln می‌گیریم :

$$\frac{-M \times 10^{-4}}{8} \leq \log_e(0.025)$$

پس :

$$\frac{M \times 10^{-4}}{8} \geq -\log_e(0.025)$$

در نتیجه :

$$M \geq -\log_e(0.025) \times 8 \times 10^4 = 295110.3563...$$

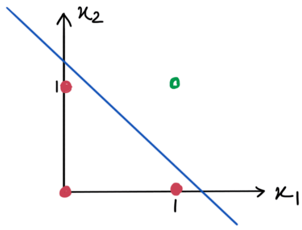
پس در این حالت نیز حداقل ۲۹۵،۱۱۱ داده برای رسیدن به چنین دقت و قاطعیتی نیاز است.

۴ سوال چهارم

۱.۴ قسمت آ

برای توابع AND و OR، فرض کنیم خروجی 1 دارای برچسب $+1$ و خروجی 0 دارای برچسب -1 باشد. برای تابع AND دو ورودی، مدلسازی به شکل زیر خواهد بود (نقاط با برچسب -1 با رنگ قرمز و نقاط با برچسب $+1$ با رنگ سبز در شکل آمده اند): مدلسازی پرسپترون به شکل زیر انجام می‌شود

x_1	x_2	y
0	0	-1
0	1	-1
1	0	-1
1	1	+1



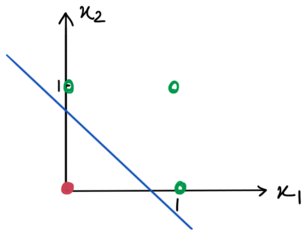
$$h(x) = sign(w_0 + w_1x_1 + w_2x_2)$$

با قرار دادن چهار سطر جدول در این رابطه داریم :

$$\begin{aligned}
-1 &= sign(w_0) && \Rightarrow w_0 < 0 \\
-1 &= sign(w_0 + w_2) && \Rightarrow w_0 + w_2 < 0 \\
-1 &= sign(w_0 + w_1) && \Rightarrow w_0 + w_1 < 0 \\
+1 &= sign(w_0 + w_1 + w_2) && \Rightarrow w_0 + w_1 + w_2 > 0
\end{aligned}$$

دستگاه معادلات بالا دارای بی‌شمار پاسخ است. به عنوان مثال $(w_0, w_1, w_2) = (-1, 0.9, 0.9)$ یک پاسخ ممکن است که در شکل بالا با رنگ آبی مشخص شده است.
برای تابع OR، مدلسازی به شکل زیر خواهد بود :

x_1	x_2	y
0	0	-1
0	1	+1
1	0	+1
1	1	+1



اگر سطرهای جدول را در معادله‌ی پرسپترون که در بالا آمد، قرار دهیم داریم :

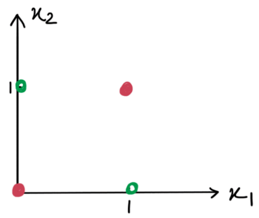
$$\begin{aligned}
-1 &= sign(w_0) && \Rightarrow w_0 < 0 \\
+1 &= sign(w_0 + w_2) && \Rightarrow w_0 + w_2 > 0 \\
+1 &= sign(w_0 + w_1) && \Rightarrow w_0 + w_1 > 0 \\
+1 &= sign(w_0 + w_1 + w_2) && \Rightarrow w_0 + w_1 + w_2 > 0
\end{aligned}$$

دستگاه معادلات بالا دارای بی‌شمار پاسخ است. به عنوان مثال $(w_0, w_1, w_2) = (-1, 2, 2)$ یک پاسخ ممکن است که در شکل بالا با رنگ آبی مشخص شده است.

۲.۴ قسمت ب

تابع XOR به شکل زیر مدلسازی می‌شود

x_1	x_2	y
0	0	-1
0	1	+1
1	0	+1
1	1	-1



اگر پرسپترون بتواند با موفقیت دسته‌بندی را انجام دهد، فرض کنید در نقطه‌ای که به این موفقیت رسیده است دارای بردار وزن (w_0, w_1, w_2) باشد. رابطه‌ی پرسپترون به شکل زیر است

$$h(x) = sign(w_0 + w_1x_1 + w_2x_2)$$

با جایگذاری سطر های جدول در این رابطه داریم :

$$-1 = sign(w_0) \quad \Rightarrow w_0 < 0 \tag{۸}$$
$$+1 = sign(w_0 + w_2) \quad \Rightarrow w_0 + w_2 > 0 \tag{۹}$$
$$+1 = sign(w_0 + w_1) \quad \Rightarrow w_0 + w_1 > 0 \tag{۱۰}$$
$$-1 = sign(w_0 + w_1 + w_2) \quad \Rightarrow w_0 + w_1 + w_2 < 0 \tag{۱۱}$$

از ۸ و ۹ نتیجه می‌شود :

$$w_2 > 0 \tag{۱۲}$$

از ۱۲ و ۱۰ هم نتیجه می‌شود :

$$w_0 + w_1 + w_2 > 0$$

که با ۱۱ در تناقض است. پس مدلسازی پرسپترون با خطای صفر برای این تابع وجود ندارد.

۳.۴ قسمت پ

فرض کنیم بردار وزن اولیه‌ی پرسپترون $w_0 = 0$ باشد. همچنین فرض کنیم در هر مرحله داده‌ی با اندیس α_i به اشتباه دسته‌بندی شود. یعنی در مرحله‌ی اول داده‌ی α_1 ام دچار misclassify می‌شود و بردار وزن به شکل $w_1 = w_0 + x_{\alpha_1}y_{\alpha_1}$ درمی‌آید. سپس در مرحله‌ی دوم داده‌ی α_2 ام دچار misclassify می‌شود و بردار وزن به شکل $w_2 = w_1 + x_{\alpha_2}y_{\alpha_2}$ درمی‌آید و به همین شکل $w_k = w_{k-1} + x_{\alpha_k}y_{\alpha_k}$. ابتدا سعی میکنیم کران بالایی برای اندازه بردار وزن در مرحله‌ی k پیدا کنیم.
از جبر خطی می‌دانیم اگر β و η دو بردار به طول n باشند، داریم :

$$\begin{aligned}\|\beta + \eta\|^2 &= (\beta_1 + \eta_1)^2 + ... + (\beta_n + \eta_n)^2 \\ &= (\beta_1^2 + ... + \beta_n^2) + (\eta_1^2 + ... + \eta_n^2) + 2(\beta_1\eta_1 + ... + \beta_n\eta_n) \\ &= \|\beta\|^2 + \|\eta\|^2 + 2\langle\beta, \eta\rangle\end{aligned}$$

پس داریم

$$\|w_k\|^2 = \|w_{k-1} + x_{\alpha_k} y_{\alpha_k}\|^2 \quad (۱۳)$$
$$= \|w_{k-1}\|^2 + \|x_{\alpha_k} y_{\alpha_k}\|^2 + 2\langle w_{k-1}, x_{\alpha_k} y_{\alpha_k}\rangle \quad (۱۴)$$
$$= \|w_{k-1}\|^2 + \|x_{\alpha_k}\|^2 + 2y_{\alpha_k}\langle w_{k-1}, x_{\alpha_k}\rangle \quad (۱۵)$$
$$\leq \|w_{k-1}\|^2 + \|x_{\alpha_k}\|^2 \quad (۱۶)$$
$$\leq \|w_{k-1}\|^2 + r^2 \quad (۱۷)$$

چون $y_{\alpha_k} \in \{-1, +1\}$ از ۱۴ به ۱۵ رسیدیم. برای اینکه بتوانیم از ۱۵ به ۱۶ برسیم، توجه داشته باشد که سیستم روی داده‌ی α_k ام دچار خطای دسته‌بندی شده است پس عبارت $2y_{\alpha_k}\langle w_{k-1}, x_{\alpha_k}\rangle$ منفی است. برای رسیدن از ۱۶ به ۱۷ از فرض سوال استفاده کردیم.
حال اگر این نامساوی را تا صفر ادامه دهیم داریم :

$$\|w_k\|^2 \leq \|w_0\|^2 + kr^2 = kr^2 \quad (۱۸)$$

حال به سراغ یافتن کران پایینی برای ضرب داخلی w^* و w_k می‌رویم. داریم:

$$\langle w^*, w_k\rangle = \langle w^*, w_{k-1} + x_{\alpha_k} y_{\alpha_k}\rangle \quad (۱۹)$$
$$= \langle w^*, w_{k-1}\rangle + \langle w^*, x_{\alpha_k} y_{\alpha_k}\rangle \quad (۲۰)$$
$$= \langle w^*, w_{k-1}\rangle + y_{\alpha_k}\langle w^*, x_{\alpha_k}\rangle \quad (۲۱)$$
$$\geq \langle w^*, w_{k-1}\rangle + \rho \quad (۲۲)$$

برای رسیدن از ۱۹ به ۲۰ از خاصیت پخشی ضرب داخلی استفاده می‌کنیم. برای رسیدن از ۲۰ به ۲۱ از مولفه‌ی دوم ضرب داخلی، عنصر y_{α_k} را استخراج کرده ایم. در نهایت برای رسیدن از ۲۱ به ۲۲ از فرض سوال استفاده کرده‌ایم. نتیجه اینکه اگر این نامساوی را تا رسیدن به $k = 0$ ادامه دهیم خواهیم داشت :

$$\langle w^*, w_k\rangle \geq \langle w^*, w_0\rangle + k\rho \quad (۲۳)$$

از جبر خطی می‌دانیم اگر β و η دو بردار باشند، داریم :

$$\langle \beta, \eta\rangle \leq \|\beta\|\|\eta\| \quad (۲۴)$$

داریم :

$$\langle w^*, w_k\rangle \leq \|w^*\|\|w_k\| \quad (۲۵)$$
$$= \|w_k\| \quad (۲۶)$$

که برای رسیدن از ۲۵ به ۲۶ از این فرض استفاده شد که $\|w^*\| = 1$
از ترکیب ۲۳ و ۲۶ داریم :

$$\|w_k\| \geq \langle w^*, w_0\rangle + k\rho$$

چون $w_0 = 0$ پس :

$$\|w_k\| \geq k\rho \quad (۲۷)$$

از جمع بندی ۱۸ و ۲۷ داریم :

$$(k\rho)^2 \leq \|w_k\|^2 \leq kr^2 \Rightarrow k \leq (\frac{r}{\rho})^2$$

۵ سوال پنجم

۱.۵ قسمت آ

خیر . به عنوان مثال اگر ما بدشانس بوده باشیم و همه‌ی نمونه‌های درون $\mathcal{D}$ برچسب -1 داشته باشند، در اینصورت A_1 فرضیه‌ی h_2 را انتخاب می‌کند که کمترین خطا را روی $\mathcal{D}$ دارد. دقت آن عبارتست از :

$$\begin{aligned} P(f(x) = h_2(x)) &= P(f(x) = +1, h_2(x) = +1) + P(f(x) = -1, h_2(x) = -1) \\ &= P(f(x) = +1)P(h_2(x) = +1) + P(f(x) = -1)P(h_2(x) = -1) \\ &= p \times (0) + (1-p) \times (1) \\ &= 1 - p \end{aligned}$$

اگر h فرضیه‌ای باشد که انتخاب تصادفی دارد. دقت h عبارتست از :

$$\begin{aligned} P(f(x) = h(x)) &= P(f(x) = +1, h(x) = +1) + P(f(x) = -1, h(x) = -1) \\ &= P(f(x) = +1)P(h(x) = +1) + P(f(x) = -1)P(h(x) = -1) \\ &= p \times \frac{1}{2} + (1-p) \times \frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

که برای به دست آورن این احتمالات از استقلال h و f استفاده شده است.
همانطور که مشخص است A_1 نمی‌تواند تضمین دهد که از انتخاب تصادفی عملکرد بهتری دارد. زیرا اگر $p > 1/2$ باشد، حالت تصادفی دقت بالاتری خواهد داشت.

از این قسمت به بعد به دلیل اینکه همه‌ی عناصر $\mathcal{D}$ دارای برچسب $+1$ هستند، پس الگوریتم A_1 فرضیه‌ی h_1 و الگوریتم A_2 فرضیه‌ی h_2 را انتخاب می‌کنند.

۲.۵ قسمت ب

بله ممکن است A_2 بهتر از A_1 باشد. اگر ما بدشانس بوده باشیم و داده‌های خارج از $\mathcal{D}$ بیشتر دارای برچسب -1 باشند، در اینصورت A_2 فرضیه‌ی بهتری تولید می‌کند.
در حالت کلی دقت A_1 برابر است با :

$$\begin{aligned} P(f(x) = h_1(x)) &= P(f(x) = +1, h_1(x) = +1) + P(f(x) = -1, h_1(x) = -1) \\ &= P(f(x) = +1)P(h_1(x) = +1) + P(f(x) = -1)P(h_1(x) = -1) \\ &= p \times 1 + (1-p) \times 0 \\ &= p \end{aligned}$$

و دقت A_2 برابر است با :

$$\begin{aligned} P(f(x) = h_2(x)) &= P(f(x) = +1, h_2(x) = +1) + P(f(x) = -1, h_2(x) = -1) \\ &= P(f(x) = +1)P(h_2(x) = +1) + P(f(x) = -1)P(h_2(x) = -1) \\ &= p \times (0) + (1-p) \times (1) \\ &= 1 - p \end{aligned}$$

حال اگر $p < 0.5$ در آن صورت A_2 الگوریتم بهتری است.

۳.۵ قسمت پ

در این بخش فرض کنیم n نمونه‌ی تست داریم. به شکل $Expected$ برچسب داده‌ها برابر $(1)(p)+(-1)(1-p)=2p-1=0.8$ است. حال اگر فقط روی n نمونه بحث کنیم اگر مقدار متوسط برچسب‌های این n نمونه بزرگنر از صفر باشد، نشان دهنده‌ی این است که الگوریتم A_1 با انتخاب h_1 به پاسخ بهتری می‌رسد و اگر مقدار متوسط این برچسب‌ها کمتر از صفر باشد، الگوریتم A_2 با انتخاب h_2 ما را به پاسخ بهتری می‌رساند. اگر برچسب های این n نمونه را $y_1, y_2, \dots, y_n$ بنامیم در این صورت احتمال اینکه این میانگین کمتر از صفر شود طبق نامساوی هافدینگ برابر است با :

$$P(|\frac{1}{n}\sum_{i=1}^{n} y_i - \underbrace{0.8}_{\text{expected value of y}} | > 0.8) \leq 2e^{-2(0.8)^2 n}$$

توجه شود که از آنجایی که $\frac{1}{n}\sum_{i=1}^{n} y_i$ حدکثر برابر مقدار 1 است پس عملا فقط زمانی $BadEvent$ رخ می‌دهد که $\frac{1}{n}\sum_{i=1}^{n} y_i < 0$ که دقیقا همان $BadEvent$ مد نظر ماست و نشان دهنده‌ی احتمال بهتر بودن A_2 نسبت به A_1 است.

حال برای استفاده از نامساوی هافدینگ اگر فرض کنیم $n \to \infty$ در آن صورت $2e^{-2(0.8)^2 n} \to 0$. به عبارتی :

$$\lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n} y_i - 0.8| > 0.8) \leq \lim_{n\to\infty} 2e^{-2(0.8)^2 n} = 0$$

پس برای n های بزرگ A_1 به احتمال ۱ فرضیه‌ی بهتری نسبت به A_2 ایجاد می‌کند.

۴.۵ قسمت ت

برای $p < 0.5$ به احتمال زیادی A_2 فرضیه بهتری نسبت به A_1 تولید می‌کند. همانطور که قبلا محاسبه شد $ExpectedValue$ مقادیر y برابر مقدار زیر است :

$$E = (1)(p) + (-1)(1-p) = 2p - 1$$

حال می‌خواهیم احتمال بهتر بودن فرضیه‌ی A_1 بر A_2 را محاسبه کنیم. در صورتی A_1 فرضیه‌ی بهتری را تولید می‌کند که میانگین مقادیر y_i از صفر بزرگتر باشد. در اینصورت تعداد نمونه‌های مثبت تست از نمونه‌های منفی بیشتر بوده است. پس A_1 با انتخاب $h_1(x) = +1$ فرضیه‌ی بهتری تولید می‌کند. طبق مفاهیم قدر مطلق می‌دانیم :

$$if |p| \geq b, b > 0\, then\, p \leq -b\, or\, p \geq b$$

پس داریم :

$$\begin{aligned} P(|\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1)| \geq 1-2p) &= P(\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1) \geq 1-2p) + P(\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1) \leq 2p-1) \\ &= P(\frac{1}{n}\sum_{i=1}^{n} y_i \geq 0) + P(\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1) \leq 2p-1) \end{aligned}$$

پس نتیجه می‌شود:

$$P(\frac{1}{n}\sum_{i=1}^{n} y_i \geq 0) \leq P(|\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1)| \geq 1-2p) \tag{۲۸}$$

همچنین طبق نامساوی هافدینگ داریم :

$$P(|\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1)| \geq 1-2p) \leq 2e^{-2(1-2p)^2 n} \tag{۲۹}$$

با کنار هم قرار دادن ۲۸ و ۲۹ داریم :

$$P(\frac{1}{n}\sum_{i=1}^{n} y_i \geq 0) \leq P(|\frac{1}{n}\sum_{i=1}^{n} y_i - (2p-1)| \geq 1-2p)$$
$$\leq 2e^{-2(1-2p)^2 n}$$

و به شکل خلاصه تر :

$$P(\frac{1}{n}\sum_{i=1}^{n} y_i \geq 0) \leq 2e^{-2(1-2p)^2 n}$$

از آنجایی که $0 < 1-2p < 1$ داریم :

$$\lim_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^{n} y_i \geq 0) \leq \lim_{n\to\infty} 2e^{-2(1-2p)^2 n} = 0$$

پس برای n های به اندازه کافی بزرگ، الگوریتم A_1 نمیتواند به خوبی A_2 عمل کند و A_2 به احتمال بسیار زیاد فرضیه های بهتری تولید می‌کند.