به نام خدا

تمرین چهارم درس هوش مصنوعی

استاد: دکتر رهبان

نویسنده: سید علیرضا میررکنی

شماره دانشجویی: ۴۰۱۱۰۶۶۱۷

دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف – بهار ۱۴۰۳

سوال ۱: در هر مرحله با استفاده از مفهوم Information Gain، بهترین feature یا ویژگی (feature با بیشترین Information Gain) را برای split کردن داده ها انتخاب می کنیم.

دقت کنید که در هر مرحله، در ابتدا $H(Y)$ و سپس $H(Y|X)$ را به ازای هر ویژگی X محاسبه می کنیم و از تفاضل این دو مقدار، Information Gain را برای آن ویژگی به دست می آوریم. دقت کنید که این دو مقدار از رابطه های زیر به دست می آیند:

$$H(Y) = -\sum_{i=0}^{N} P(Y = y_i) \log_2 P(Y = y_i)$$

$$H(Y|X) = -\sum_{j=0}^{M} P(X = x_j) \sum_{i=0}^{N} P(Y = y_i|X = x_i) \log_2 P(Y = y_i|X = x_i)$$

دقت کنید که feature های x_1 و x_3، feature های عددی پیوسته هستند و در نتیجه باید حالات مختلف threshold برای تقسیم بندی نمونه ها با استفاده از این دو feature را در نظر بگیریم و از هر کدام که IG بیشتری داشت استفاده بکنیم. در ادامه threshold برای ویژگی x_1 را با t_1 و threshold برای ویژگی x_3 را با t_2 نمایش می دهیم. و بنابراین خواهیم داشت:

$$\begin{aligned} H(Y|x_1) = &-P(x_1 \leq t_1)P(Y|x_1 \leq t_1) \log_2 P(Y|x_1 \leq t_1) \\ &- P(x_1 > t_1)P(Y|x_1 > t_1) \log_2 P(Y|x_1 > t_1) \\ H(Y|x_3) = &-P(x_3 \leq t_3)P(Y|x_3 \leq t_3) \log_2 P(Y|x_3 \leq t_3) \\ &- P(x_3 > t_3)P(Y|x_3 > t_3) \log_2 P(Y|x_3 > t_3) \end{aligned}$$

توجه کنید که عملا نیازی به محاسبه $H(Y)$ نمی باشد و کافی است تنها مقادیر $H(Y|X)$ را محاسبه کنیم و از بین آن ها، ویژگی ای که کمترین آنتروپی را دارا می باشد، برای split کردن راسی که در آن هستیم انتخاب نماییم؛ چرا که قطعا این ویژگی دارای بیشترین IG خواهد بود.

در ادامه با توجه به مطالب بالا، درخت تصمیم را به ترتیب برای دروس Computer Architecture و Machine Learning رسم می کنیم.

:Computer Architecture

عمق اول:

دقت کنید که در اینجا t_1 می تواند برابر 5، 6 یا 7 و t_3 می تواند برابر 0.1، 0.3 یا 0.4 باشد.

$$t_1 = 5 \Longrightarrow H(Y|x_1) = -\frac{1}{5}(\log_2 1) - \frac{4}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.8$$

$$t_1 = 6 \Longrightarrow H(Y|x_1) = -\frac{2}{5}(\log_2 1) - \frac{3}{5}\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.551$$

$$t_1 = 7 \Longrightarrow H(Y|x_1) = -\frac{1}{5}(\log_2 1) - \frac{4}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.8$$

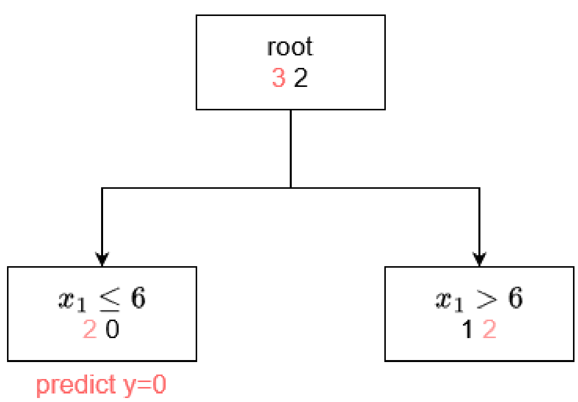
$$H(Y|x_2) = -\frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.951$$

$$t_3 = 0.1 \Longrightarrow H(Y|x_3) = -\frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{3}{5}\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.951$$

$$t_3 = 0.3 \Longrightarrow H(Y|x_3) = -\frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{2}{5}(\log_2 1) = 0.551$$

$$t_3 = 0.4 \Longrightarrow H(Y|x_3) = -\frac{4}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{1}{5}(\log_2 1) = 0.8$$

بنابراین در این عمق، ویژگی x_1 با threshold 6 دارای بیشترین IG (کمترین آنتروپی) خواهد بود و با استفاده از آن درخت را باز می کنیم. در نتیجه، درخت به شکل زیر می شود:



پس باید برگ سمت راست را مجددا split کنیم.

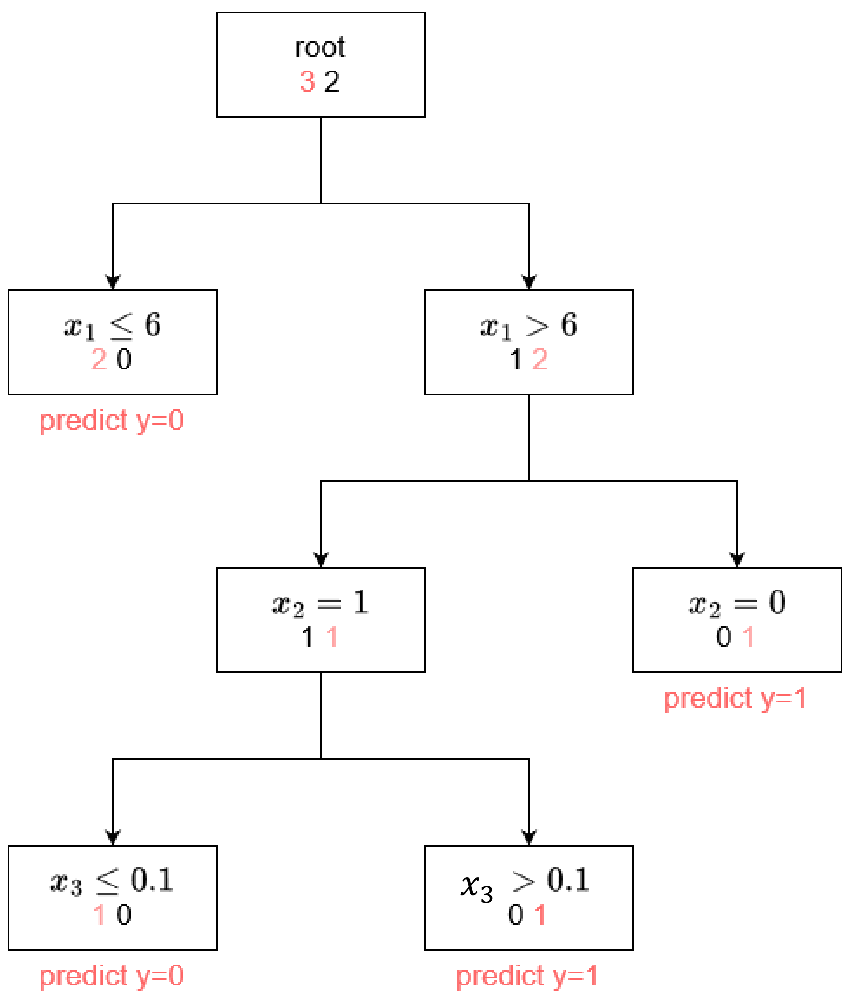
عمق دوم و سوم:

دقت کنید که در اینجا t_3 می تواند برابر 0.1 باشد.

$$H(Y|x_2) = -\frac{1}{3}(\log_2 1) - \frac{2}{3}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.67$$

$$t_3 = 0.1 \Longrightarrow H(Y|x_3) = -\frac{2}{3}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{1}{3}(\log_2 1) = 0.67$$

بنابراین در این عمق، ویژگی x_2 با دارای بیشترین IG (کمترین آنتروپی) خواهد بود و با استفاده از آن درخت را باز می کنیم. در نهایت برگی که خالص نشده است (هم نمونه ای با برچسب ۱ و هم نمونه ای با برچسب ۰ دارد) را با استفاده از ویژگی x_3 با threshold 0.1 باز می کنیم و درخت تصمیم نهایی به شکل زیر می شود (به منظور سادگی بیشتر، درخت تا عمق ۲ را رسم نمی کنیم و مستقیم درخت نهایی را رسم می نماییم).



:Machine Learning

عمق اول:

دقت کنید که در اینجا t_1 می تواند برابر 5، 6 یا 7 و t_3 می تواند برابر 0.4 یا 0.5 باشد.

$$t_1 = 5 \Longrightarrow H(Y|x_1) = -\frac{1}{5}(\log_2 1) - \frac{4}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.8$$

$$t_1 = 6 \Longrightarrow H(Y|x_1) = -\frac{2}{5}(\log_2 1) - \frac{3}{5}\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.551$$

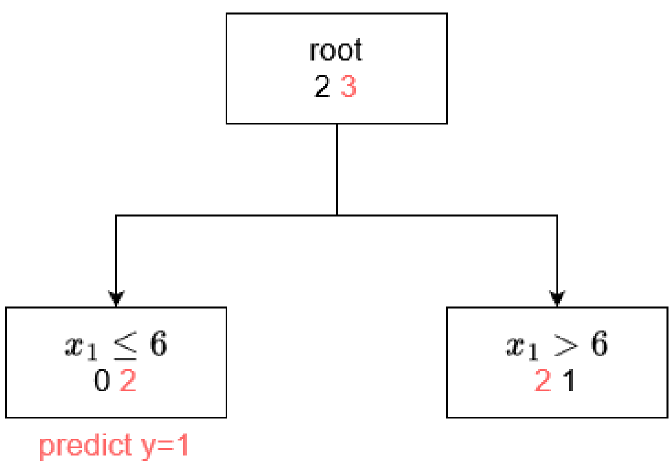
$$t_1 = 7 \Longrightarrow H(Y|x_1) = -\frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.951$$

$$H(Y|x_2) = -\frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.951$$

$$t_3 = 0.4 \Longrightarrow H(Y|x_3) = -\frac{1}{5}\ (\log_2 1) - \frac{4}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.8$$

$$t_3 = 0.5 \Longrightarrow H(Y|x_3) = -\frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.951$$

بنابراین در این عمق، ویژگی x_1 با threshold 6 دارای بیشترین IG (کمترین آنتروپی) خواهد بود و با استفاده از آن درخت را باز می کنیم. در نتیجه، درخت به شکل زیر می شود:



پس باید برگ سمت راست را مجددا split کنیم.

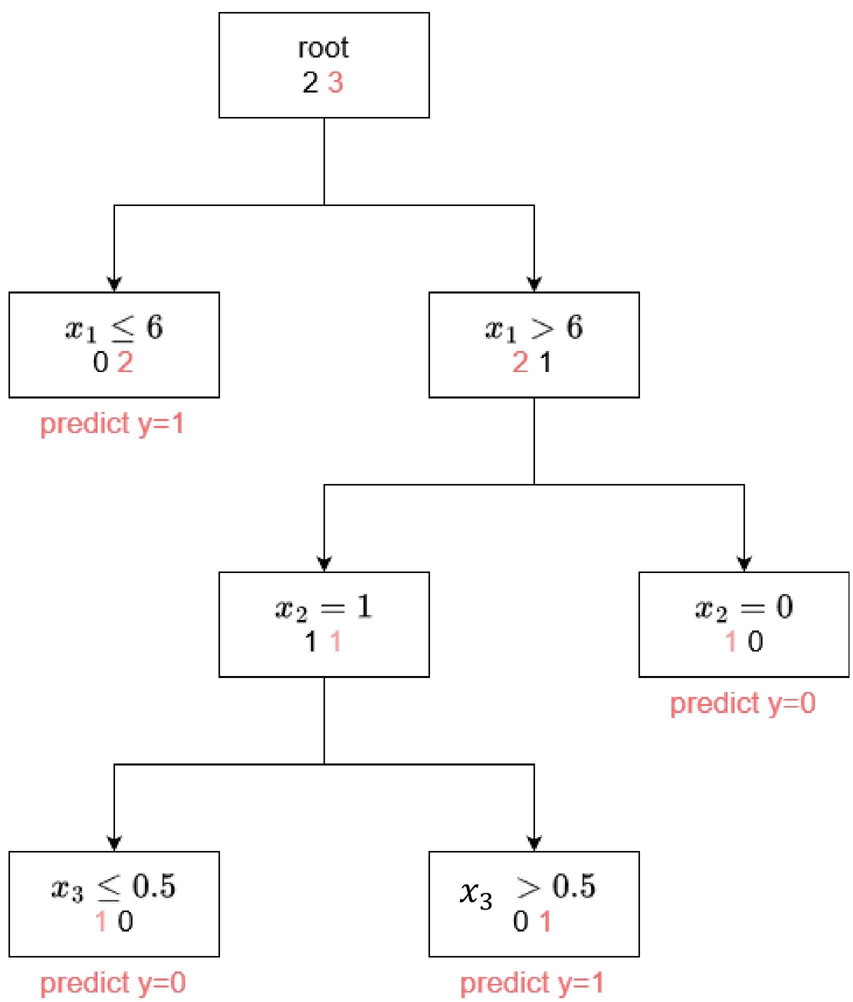
عمق دوم و سوم:

دقت کنید که در اینجا t_3 می تواند برابر 0.5 باشد.

$$H(Y|x_2) = -\frac{1}{3}(\log_2 1) - \frac{2}{3}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.67$$

$$t_3 = 0.5 \Longrightarrow H(Y|x_3) = -\frac{1}{3}(\log_2 1) - \frac{2}{3}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.67$$

بنابراین در این عمق، ویژگی x_2 با دارای بیشترین IG (کمترین آنتروپی) خواهد بود و با استفاده از آن درخت را باز می کنیم. در نهایت برگی که خالص نشده است (هم نمونه ای با برچسب ۱ و هم نمونه ای با برچسب ۰ دارد) را با استفاده از ویژگی x_3 با threshold 0.5 باز می کنیم و درخت تصمیم نهایی به شکل زیر می شود (به منظور سادگی بیشتر، درخت تا عمق ۲ را رسم نمی کنیم و مستقیم درخت نهایی را رسم می نماییم



در انتها، با توجه به ویژگی های دو نمونه داده شده، تمدید شدن هر کدام از دو درس را تخمین می زنیم.

برای درس معماری کامپیوتر، $x_1 > 6$ ، $x_2 = 1$ و $x_3 > 0.1$ می باشد و در نتیجه با توجه به درخت تصمیم به دست آمده، درس معماری کامپیوتر تمدید می شود.

برای درس یادگیری ماشین، $x_1 > 6$ ، $x_2 = 0$ و $x_3 > 0.5$ می باشد و در نتیجه با توجه به درخت تصمیم به دست آمده، درس یادگیری ماشین تمدید نمی شود.

بنابراین دانشجو باید تمرین درس یادگیری ماشین را زودتر بزند، چرا که تمرین این درس (احتمالا) تمدید نمی شود، در حالی که تمرین درس معماری (احتمالا) تمدید خواهد شد.

سوال ۲:

الف) می دانیم که با فرض مدل بیز ساده لوحانه، هر دو feature دلخواه به شرط برچسب Y از یکدیگر مستقل هستند. بنابراین در این مسئله، با فرض مدل بیز ساده لوحانه، علاوه بر استقلال های ذکر شده در صورت سوال feature های X_2 و X_3 نیز به شرط برچسب Y مستقل از یکدیگر خواهند بود.

حال دقت کنید که با این فرض، می توان نوشت (دقت کنید که چون در نهایت می خواهیم مقادیر زیر را با هم مقایسه کنیم، نیازی به normalize کردن احتمالات نداریم):

$$P(Y=1|X_1=1,X_2=0,X_3=0)$$
$$\propto P(Y=1)P(X_1=1|Y=1)P(X_2=0|Y=1)P(X_3=0|Y=1)$$
$$\Longrightarrow P(Y=1|X_1=1,X_2=0,X_3=0)\propto 0.5\times p\times q^2=\frac{pq^2}{2}$$
$$P(Y=0|X_1=1,X_2=0,X_3=0)$$
$$\propto P(Y=0)P(X_1=1|Y=0)P(X_2=0|Y=0)P(X_3=0|Y=0)$$
$$\Longrightarrow P(Y=0|X_1=1,X_2=0,X_3=0)\propto 0.5\times(1-p)\times(1-q)^2=\frac{(1-p)(1-q)^2}{2}$$

حال برای به دست آوردن قاعده تصمیم گیری به ازای $Y=1$، داریم (برای انتخاب شدن $Y=1$، باید احتمال اول بزرگ تر مساوی احتمال دوم باشد):

$$P(Y=1|X_1=1,X_2=0,X_3=0)\geq P(Y=0|X_1=1,X_2=0,X_3=0)$$
$$\Longrightarrow \frac{pq^2}{2}\geq\frac{(1-p)(1-q)^2}{2}\Longrightarrow pq^2\geq(1-p)(1-q)^2$$

پس نامساوی بالا که با رنگ قرمز مشخص شده است، قاعده تصمیم گیری به ازای $Y=1$ بر حسب p و q می باشد.

ب) در این قسمت، با توجه به اینکه می دانیم feature های X_2 و X_3 کاملا یکسان هستند و feature های X_1 و X_2 به شرط برچسب Y از یکدیگر مستقل می باشند، می توانیم ویژگی X_3 را از بین ویژگی ها حذف کنیم (چرا که نسبت به ویژگی های دیگر، اطلاعات اضافه ای به ما نمی دهد) و تنها از ویژگی های X_1 و X_2 استفاده کنیم.

بنابراین، خواهیم داشت (دقت کنید که چون در نهایت می خواهیم مقادیر زیر را با هم مقایسه کنیم، نیازی به normalize کردن احتمالات نداریم):

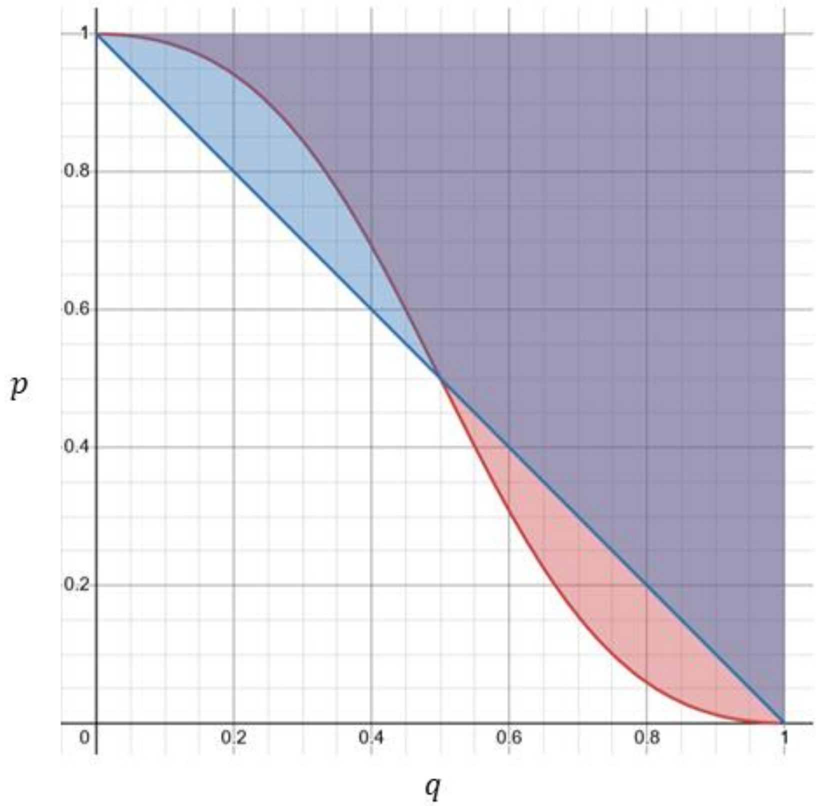
$$P(Y=1|X_1=1,X_2=0) \propto P(Y=1)P(X_1=1|Y=1)P(X_2=0|Y=1)$$
$$\Longrightarrow P(Y=1|X_1=1,X_2=0) \propto 0.5\times p\times q=\frac{pq}{2}$$
$$P(Y=0|X_1=1,X_2=0) \propto P(Y=0)P(X_1=1|Y=0)P(X_2=0|Y=0)$$
$$\Longrightarrow P(Y=0|X_1=1,X_2=0) \propto 0.5\times (1-p)\times (1-q)=\frac{(1-p)(1-q)}{2}$$

حال برای به دست آوردن قاعده تصمیم گیری به ازای $Y=1$، داریم (برای انتخاب شدن $Y=1$، باید احتمال اول بزرگ تر مساوی احتمال دوم باشد):

$$P(Y=1|X_1=1,X_2=0) \geq P(Y=0|X_1=1,X_2=0)$$
$$\Longrightarrow \frac{pq}{2}\geq\frac{(1-p)(1-q)}{2}\Longrightarrow pq\geq(1-p)(1-q)\Longrightarrow pq\geq 1-p-q+pq$$
$$\Longrightarrow p+q\geq 1$$

پس نامساوی بالا که با رنگ قرمز مشخص شده است، قاعده تصمیم گیری به ازای $Y=1$ بر حسب p و q می باشد.

ج) در شکل زیر می توانید مرز تصمیم گیری را برای قسمت های الف و ب مشاهده کنید. خط قرمز رنگ، مرز تصمیم گیری را برای قسمت الف و خط آبی رنگ، مرز تصمیم گیری را برای قسمت ب نمایش می دهد. هر نقطه روی و یا در بالای مرز های تصمیم گیری، نشان دهنده حالتی می باشد که مقدار $Y = 1$ را در آن پیش بینی خواهیم نمود.



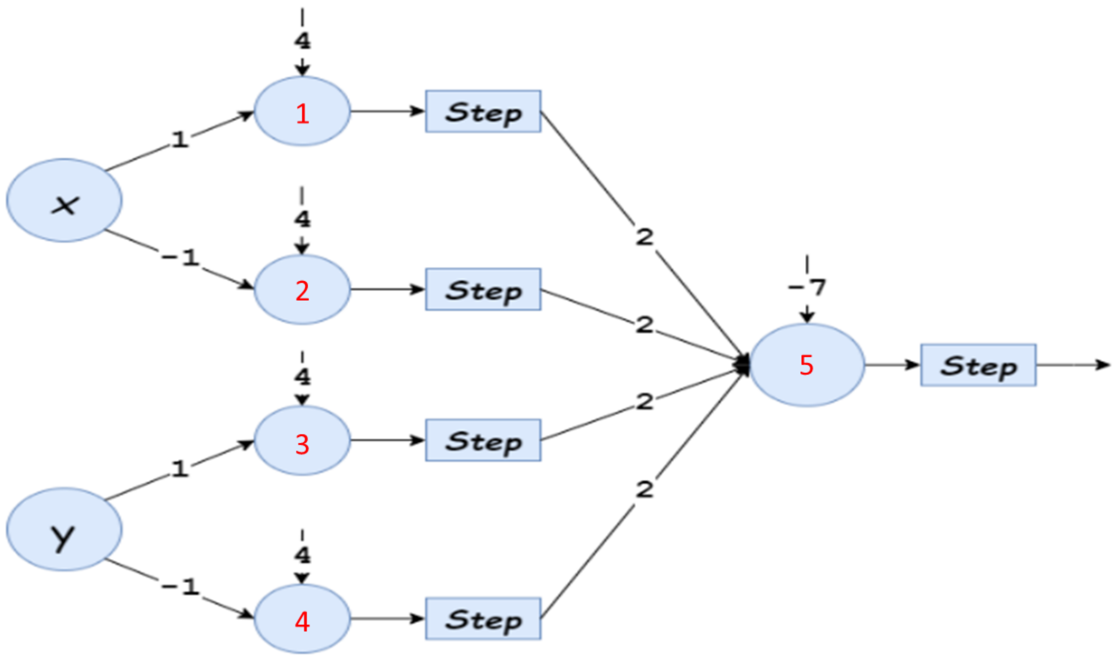
همانطور که مشاهده می شود، ناحیه $(0,1) \times (0,1)$ به ۴ قسمت تقسیم شده است:

- نواحی سفید رنگ
- نواحی فقط قرمز رنگ
- نواحی فقط آبی رنگ
- نواحی قرمزآبی

در نواحی که فقط قرمز رنگ و یا فقط آبی رنگ هستند، مدل بیز ساده لوحانه نسبت به قاعده تصمیم گیری بهینه دچار خطا می شود. در نواحی قرمز رنگ، مدل بیز ساده لوحانه مقدار $Y = 1$ و قاعده تصمیم گیری بهینه مقدار $Y = 0$ را پیش بینی می کند و در نواحی آبی رنگ، مدل بیز ساده لوحانه مقدار $Y = 0$ و قاعده تصمیم گیری بهینه مقدار $Y = 0$ را پیش بینی خواهد کرد.

سوال ۳:

الف) در ابتدا نورون های موجود در شبکه عصبی را به شکل زیر شماره گذاری می کنیم.



حال مقدار نوشته شده در داخل هر کدام از نورون ها را به دست می آوریم. دقت کنید که مقدار به دست آمده در راس شماره i را با o_i نمایش می دهیم.

$$o_1 = x + 4$$

$$o_2 = -x + 4$$

$$o_3 = y + 4$$

$$o_4 = -y + 4$$

$$\begin{aligned} o_5 &= 2step(o_1) + 2step(o_2) + 2step(o_3) + 2step(o_4) - 7 \\ &= 2step(x + 4) + 2step(-x + 4) + 2step(y + 4) + 2step(-y + 4) - 7 \end{aligned}$$

حال دقت کنید که خروجی نهایی شبکه عصبی که آن را با o نمایش می دهیم، به شکل زیر محاسبه می شود:

$$\begin{aligned} o &= step(o_5) \\ &= step(2step(x + 4) + 2step(-x + 4) + 2step(y + 4) + 2step(-y + 4) - 7) \end{aligned}$$

در ادامه توجه کنید که تابع $step$ به شکل زیر تعریف می شود:

$$step(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

بنابراین می توان نوشت:

$$\begin{aligned} o = 1 &\Leftrightarrow step(2step(x+4) + 2step(-x+4) + 2step(y+4) + 2step(-y+4) - 7) = 0 \\ &\Leftrightarrow 2step(x+4) + 2step(-x+4) + 2step(y+4) + 2step(-y+4) - 7 \geq 0 \\ &\Leftrightarrow 2step(x+4) + 2step(-x+4) + 2step(y+4) + 2step(-y+4) \geq 7 \end{aligned}$$

حال دقت کنید که مقدار عبارت سمت چپ نامساوی بالا، اگر حتی خروجی یکی از توابع $step$ برابر ۰ شود، کوچک تر مساوی ۶ خواهد بود. بنابراین برای اینکه نامساوی بالا برقرار شود، باید خروجی تمامی توابع $step$ برابر ۱ شود که برای این منظور باید داشته باشیم:

$$step(x+4) = 1 \Rightarrow x + 4 \geq 0 \Rightarrow x \geq -4$$
$$step(-x+4) = 1 \Rightarrow -x + 4 \geq 0 \Rightarrow x \leq 4$$
$$step(y+4) = 1 \Rightarrow y + 4 \geq 0 \Rightarrow y \geq -4$$
$$step(-y+4) = 1 \Rightarrow -y + 4 \geq 0 \Rightarrow y \leq 4$$

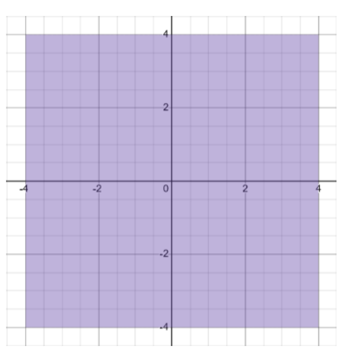
بنابراین، مقدار o برابر ۱ می شود اگر و تنها اگر تمامی نامساوی های بالا برقرار باشند:

$$o = 1 \Leftrightarrow x \geq -4 \ \wedge \ x \leq 4 \ \wedge \ y \geq -4 \ \wedge \ y \leq 4 \Leftrightarrow |x| \leq 4 \wedge |y| \leq 4$$

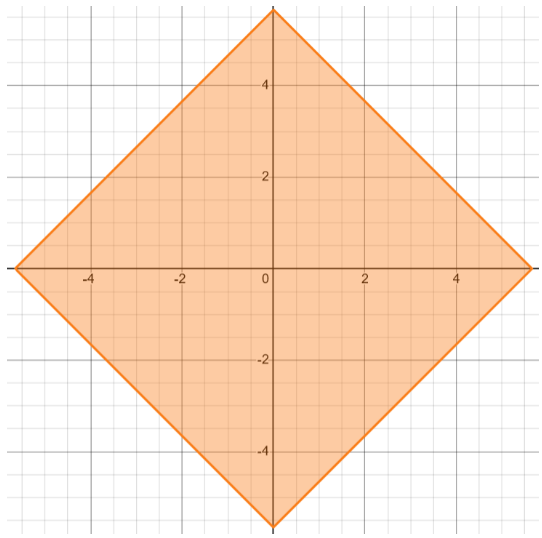
پس می توان خروجی شبکه عصبی را به شکل زیر نمایش داد:

$$o = \begin{cases} 1 & |x| \leq 4 \wedge |y| \leq 4 \\ 0 & otherwise \end{cases}$$

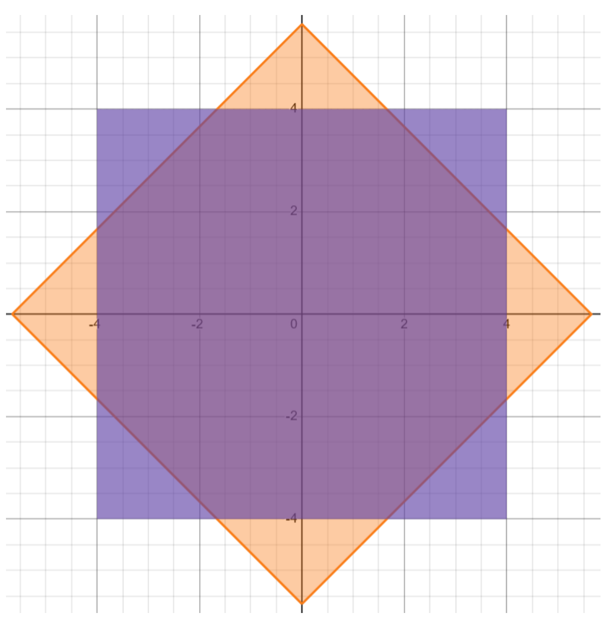
به عبارت دیگر این شبکه عصبی، تعیین می کند که آیا تیر به هدفی به شکل زیر خورده است یا خیر:



ب) برای به دست آوردن شبکه عصبی مطلوب، از ایده قسمت قبل استفاده می کنیم؛ به این شکل که در ابتدا یک شبکه عصبی طراحی می کنیم تا تشخیص دهد که آیا تیر به هدفی به شکل زیر خورده است یا خیر:



هدف بالا یک مربع به طول ضلع ۸ (و در نتیجه طول قطر $8\sqrt{2}$ می باشد). در نهایت خروجی شبکه عصبی به دست آمده را با خروجی شبکه عصبی قسمت قبل or می کنیم تا تعیین کنیم آیا تیر به هدف اصلی (هدفی که شکل آن در صورت سوال آمده است) خورده است یا خیر. علت درست بودن این کار، این است که همانطور که در شکل زیر مشاهده می شود هدف اصلی از روی هم قرار گرفتن هدف بالا و هدف قسمت الف به دست می آید.



حال معادله خطوط هدف جدید را به دست می آوریم. دقت کنید که معادله مربع هدف به شکل زیر می باشد:

$$|x| + |y| \leq 4\sqrt{2}$$

برقراری نامعادله بالا، معادل با برقراری همزمان ۴ نامعادله زیر می باشد:

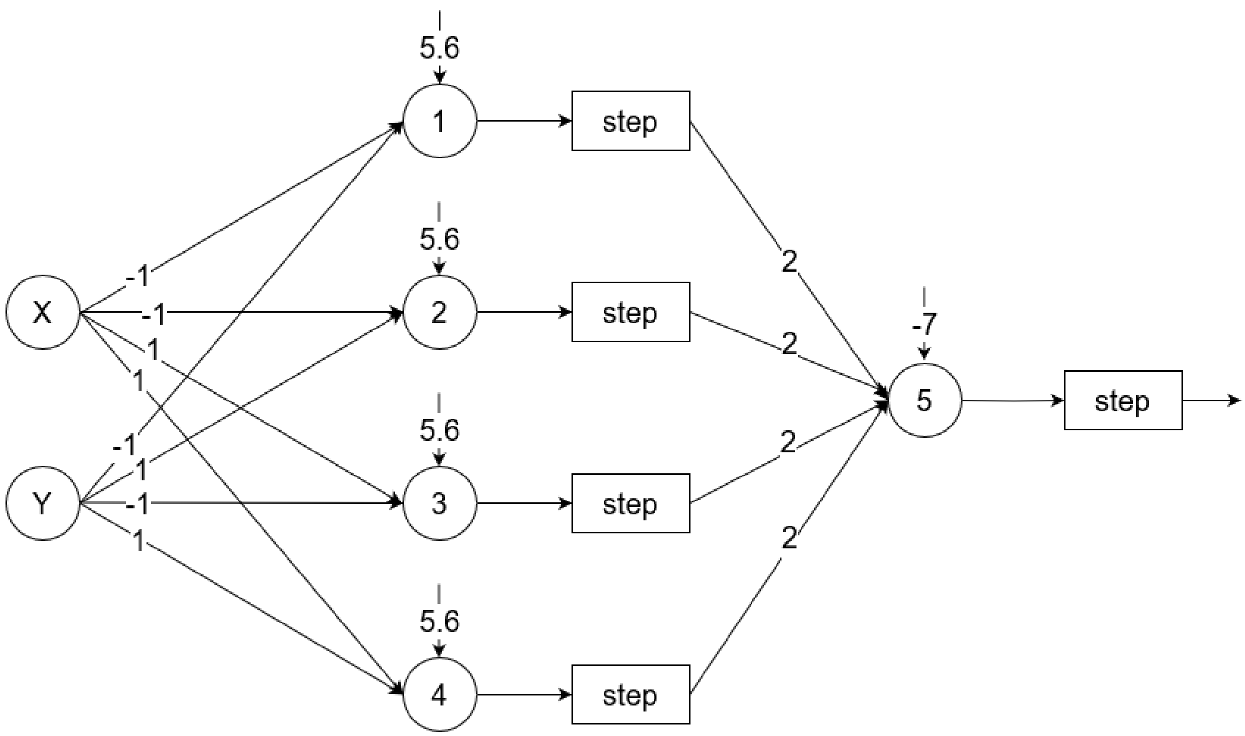
$$x + y \leq 4\sqrt{2} \xRightarrow{\sqrt{2}=1.4} 0 \leq 5.6 - x - y$$

$$x - y \leq 4\sqrt{2} \xRightarrow{\sqrt{2}=1.4} 0 \leq 5.6 - x + y$$

$$-x + y \leq 4\sqrt{2} \xRightarrow{\sqrt{2}=1.4} 0 \leq 5.6 + x - y$$

$$-x - y \leq 4\sqrt{2} \xRightarrow{\sqrt{2}=1.4} 0 \leq 5.6 + x + y$$

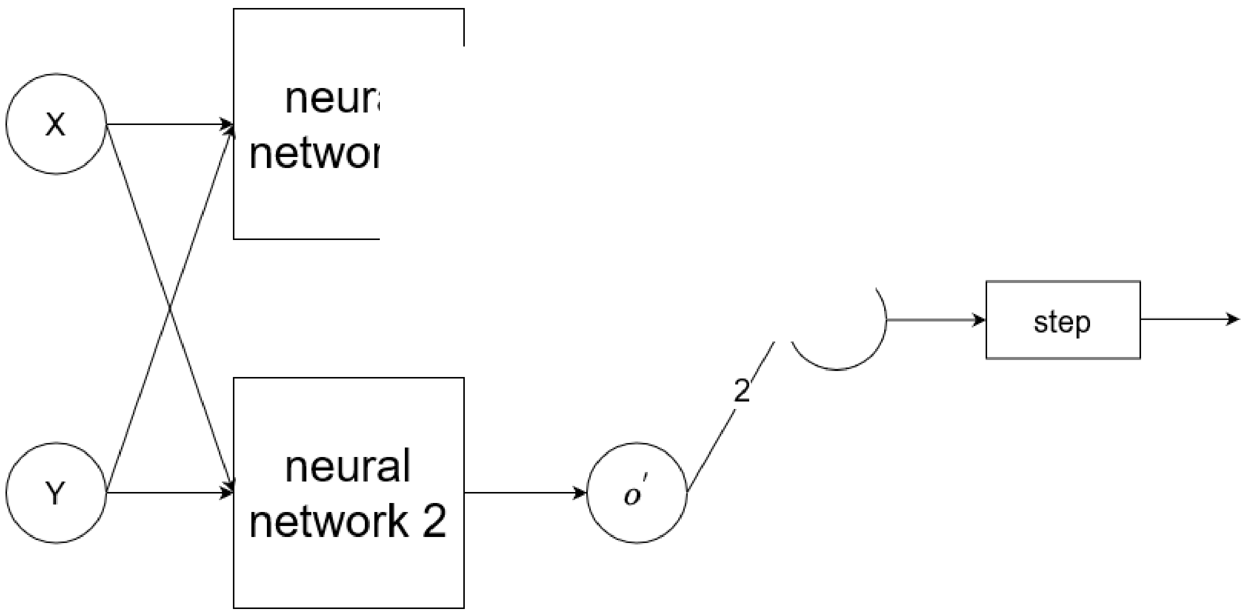
حال دقت کنید که خروجی شبکه عصبی زیر در صورتی که هر ۴ نامعادله بالا برقرار شوند برابر ۱ خواهد بود:



به عبارت دیگر، اگر خروجی شبکه عصبی بالا را با o' نمایش دهیم، داریم:

$$o' = \begin{cases} 1 & |x| + |y| \leq 5.6 \\ 0 & otherwise \end{cases}$$

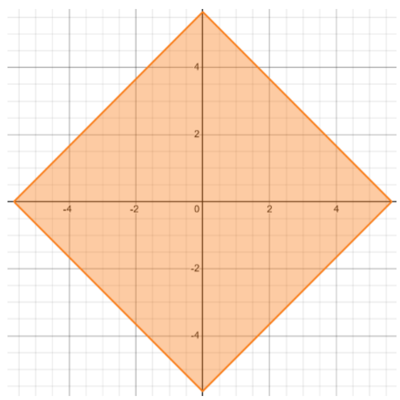
حال شبکه عصبی نهایی را به شکل زیر می سازیم:



به طوری که در شکل بالا، $neural\ network\ 1$ شبکه عصبی مربوط به قسمت الف و $neural\ network\ 2$ شبکه عصبی مربوط به همین قسمت می باشد.

دقت کنید که خروجی شبکه عصبی بالا، or خروجی شبکه های عصبی $neural\ network\ 1$ و $neural\ network\ 2$ می باشد.

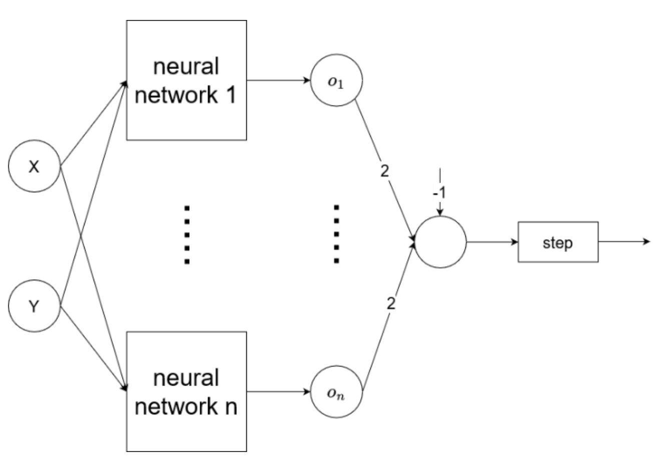
ج) به منظور تبدیل هدف به یک دایره، تعداد مربع های مورب را افزایش می دهیم. به عبارت دیگر به جای استفاده از تنها دو مربع، از n مربع استفاده می کنیم؛ به طوری که مربع اول به شکل زیر بوده و برای $2 \leq i \leq n$، i امین مربع از دوران $\theta_i = \frac{(i-1)\pi}{2n}$ درجه مربع اول در جهت پاد ساعتگرد حول مبدا مختصات حاصل می شود.



حال دقت کنید که ناحیه مشخص کننده i امین مربع (برای $1 \leq i \leq n$)، در نامعادله زیر صدق می کند (با استفاده از ضرب کردن مختصات نقاط در ماتریس دوران می توان به نامعادله زیر دست یافت) :

$$|x \cos \theta_i - y \sin \theta_i| + |x \sin \theta_i + y \cos \theta_i| \leq 4\sqrt{2}$$
$$\Rightarrow \left|x \cos \frac{(i-1)\pi}{2n} - y \sin \frac{(i-1)\pi}{2n}\right| + \left|x \sin \frac{(i-1)\pi}{2n} + y \cos \frac{(i-1)\pi}{2n}\right| \leq 4\sqrt{2}$$

حال مشابه قسمت های قبل، شبکه عصبی متناظر به هر کدام از این اهداف مربعی را می سازیم و خروجی آن ها را با هم or می کنیم. اگر خروجی شبکه عصبی متناظر به مربع i ام را با o_i نمایش دهیم، شبکه عصبی نهایی به شکل زیر خواهد بود. دقت کنید که حالت خاص $n = 2$ منجر به ایجاد هدفی مشابه هدف صورت سوال می شود.



سوال ۴:

الف) شکل ماتریسی مسئله رگرسیون خطی، به صورت زیر می باشد.

در ابتدا ماتریس $X_{n\times(m+1)}$ را به شکل زیر تعریف کرده و آن را ماتریس ویژگی ها می نامیم:

$$X = \begin{bmatrix} x_{10} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{nm} \end{bmatrix}$$

به عبارت دیگر، درایه واقع در تقاطع سطر i ام و ستون j ام ماتریس X به طوری که $1 \le i \le n$ و $0 \le j \le m$، همان x_{ij} می باشد. دقت کنید که برای هر $1 \le i \le n$، تعریف می کنیم $x_{i0} = 1$.

به طور مفهومی، در سطر i ام ماتریس ویژگی ها، ویژگی های نمونه i ام نوشته شده اند.

در ادامه، بردار $\hat{\beta}$ را به شکل زیر تعریف کرده و آن را بردار ضرایب می نامیم:

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix}$$

در نهایت، بردار y را به شکل زیر تعریف کرده و آن را بردار برچسب ها می نامیم:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

حال، با تعریف ماتریس و بردار های بالا، فرم ماتریسی مسئله رگرسیون خطی به شکل زیر خواهد بود:

$$\begin{bmatrix} x_{10} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \Longrightarrow X\beta = y$$

به طوری که در آن، هدف کمینه کردن تابع لاس است که به شکل زیر بازنویسی می شود:

$$L = \|X\beta - y\|^2$$

ب) در ابتدا تابع L را با توجه به تعریف نرم با استفاده از ضرب داخلی، به شکل زیر بازنویسی می کنیم:

$$\begin{aligned} L = \|X\beta - y\|^2 &= (X\beta - y)^T(X\beta - y) = ((X\beta)^T - y^T)(X\beta - y) \\ &= (\beta^T X^T - y^T)(X\beta - y) = \beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y \\ &= \beta^T X^T X\beta - 2\beta^T X^T y + y^T y \end{aligned}$$

دقت کنید که در مراحل بالا، از تساوی زیر استفاده نمودیم:

$$y^T X\beta = y \cdot X\beta = X\beta \cdot y = (X\beta)^T y = \beta^T X^T y$$

در ادامه، برای کمینه کردن تابع L، از آن (نسبت به β) مشتق می گیریم و مشتق را برابر صفر قرار می دهیم تا نقطه بحرانی تابع را پیدا کنیم. دقت کنید که تابع L یک تابع محدب می باشد، بنابراین دارای دقیقا یک کمینه سراسری بوده و هر کمینه موضعی آن، کمینه سراسری می باشد. این بدین معنی است که نقطه بحرانی به دست آمده از برابر قرار دادن مشتق L با صفر، همان نقطه ای است که L در آن کمینه می شود.

$$\frac{\partial L}{\partial \beta}(\hat{\beta}) = 2X^T X\hat{\beta} - 2X^T y = 0 \Longrightarrow X^T X\hat{\beta} = X^T y$$

حال طرفین تساوی بالا را در $(X^T X)^{-1}$ ضرب می کنیم (فرض می کنیم ستون های ماتریس X مستقل خطی هستند و در نتیجه $X^T X$ وارون پذیر می باشد):

$$(X^T X)^{-1}\left(X^T X\hat{\beta}\right) = \left((X^T X)^{-1}(X^T X)\right)\hat{\beta} = (X^T X)^{-1} X^T y \Longrightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

که این همان حکم مسئله می باشد.

ج) تابع L را می توان به شکل زیر بازنویسی کرد (با توجه به شکل ماتریسی مسئله رگرسیون خطی):

$$L = \|X\beta - y\|^2 + \lambda\|\beta\|^2$$

تابع L را با توجه به تعریف نرم با استفاده از ضرب داخلی و همچنین با استفاده از قسمت قبل، به شکل زیر بازنویسی می کنیم:

$$L = \beta^T X^T X\beta - 2\beta^T X^T y + y^T y + \lambda\beta^T\beta$$

مجددا با توجه به محدب بودن تابع لاس، با مشتق گرفتن از آن نسبت به β و برابر صفر قرار دادن مشتق، نقطه ای را می یابیم که L در آن کمینه می شود.

$$\frac{\partial L}{\partial\beta}\left(\hat{\beta}\right) = 2X^T X\hat{\beta} - 2X^T y + 2\lambda\hat{\beta} = 0 \Longrightarrow (X^T X + \lambda I)\hat{\beta} = X^T y$$

حال طرفین تساوی بالا را در $(X^T X + \lambda I)^{-1}$ ضرب می کنیم (فرض می کنیم ستون های ماتریس $X^T X + \lambda I$ مستقل خطی هستند و در نتیجه $X^T X + \lambda I$ وارون پذیر می باشد):

$$(X^T X + \lambda I)^{-1}\left((X^T X + \lambda)\hat{\beta}\right) = \left((X^T X + \lambda I)^{-1}(X^T X + \lambda I)\right)\hat{\beta} = (X^T X + \lambda I)^{-1}X^T y$$
$$\Longrightarrow \hat{\beta} = (X^T X + \lambda I)^{-1}X^T y$$

که این همان حکم مسئله می باشد.

دقت کنید که در عبارات بالا از تساوی زیر استفاده کردیم:

$$\beta^T\beta = \beta^T I\beta \Longrightarrow \frac{\partial\beta^T\beta}{\partial\beta} = (I + I^T)\beta = 2I\beta = 2\beta$$

سوال ۵:

الف) در ابتدا تعریف می کنیم:

$$w^* = argmin_w\{\|w\| : \forall i \in [m], y_i\langle w, x_i\rangle \geq 1\} \Longrightarrow \|w^*\| = B$$

حال دقت کنید که با توجه به سودوکد نوشته شده برای الگوریتم و همچنین با توجه به تعریف w^*، می توان نوشت:

$$\langle w^*, w^{(t+1)}\rangle - \langle w^*, w^{(t)}\rangle = \langle w^*, w^{(t+1)} - w^{(t)}\rangle = \langle w^*, y_i x_i\rangle = y_i\langle w^*, x_i\rangle \geq 1$$

همچنین داریم:

$$w^{(1)} = (0, \dots, 0) \Longrightarrow \langle w^*, w^{(1)}\rangle = \langle w^*, (0, \dots, 0)\rangle = 0$$

$$\Longrightarrow \langle w^*, w^{(T+1))}\rangle = \langle w^*, w^{(T+1)}\rangle - \langle w^*, w^{(1)}\rangle = \sum_{t=1}^{T}\langle w^*, w^{(t+1)}\rangle - \langle w^*, w^{(t)}\rangle \geq \sum_{t=1}^{T} 1 = T$$

از طرف دیگر، دقت کنید که می توان نوشت:

$$\left\|w^{(t+1)}\right\|^2 = \left\|w^{(t)} + y_i x_i\right\|^2 = \left\|w^{(t)}\right\|^2 + y_i^2\|x_i\|^2 + 2y_i\langle w^{(t)}, x_i\rangle$$

حال دقت کنید که در عبارت بالا، $y_i\langle w^{(t)}, x_i\rangle \leq 0$ می باشد چرا که update انجام داده ایم. بنابراین با توجه به تعریف R خواهیم داشت:

$$\left\|w^{(t+1)}\right\|^2 \leq \left\|w^{(t)}\right\|^2 + y_i^2\|x_i\|^2 \leq \left\|w^{(t)}\right\|^2 + R^2 \Longrightarrow \left\|w^{(t+1)}\right\|^2 \leq tR^2 + \left\|w^{(1)}\right\|^2 = tR^2$$

$$\Longrightarrow \left\|w^{(T+1)}\right\|^2 \leq TR^2 \Longrightarrow \left\|w^{(T+1)}\right\| \leq R\sqrt{T}$$

حال دقت کنید که با ترکیب نامساوی هایی که با رنگ آبی مشخص شده اند و نامساوی کوشی شوارتز، داریم:

$$T \leq \langle w^*, w^{(T+1))}\rangle \leq \|w^*\|\left\|w^{(T+1)}\right\| \leq BR\sqrt{T} \Longrightarrow \sqrt{T} \leq BR \Longrightarrow T \leq (BR)^2$$

که این همان قسمت اول حکم مسئله می باشد و درستی آن را نشان دادیم. همچنین به وضوح اجرای الگوریتم زمانی پایان می یابد که عبارت شرطی if برابر false شود، یا عبارت دیگر نقیض آن True شود و داشته باشیم $\forall i\ ; y_i\langle w^*, w^{(T+1))}\rangle > 0$ که این درستی ادامه حکم را نشان می دهد.

ب) با توجه به راهنمایی سوال، قرار می دهیم $d = m$ و دنباله داده زیر را معرفی می کنیم:

$$\{(x_i = e_i, y_i = 1)\}_{i=1}^{n}$$

به طوری که e_i، i امین بردار یکه استاندارد در فضای $\mathbb{R}^m$ می باشد. به وضوح برای این دنباله داده داریم:

$$R = \max_i \|x_i\| = \max_i \|e_i\| = 1 \leq 1$$

لم: در t امین تکرار اجرای الگوریتم PTA بر روی دنباله داده معرفی شده، بردار وزن ها به شکل زیر خواهد بود:

$$w^{(t+1)} = e_1 + e_2 + \cdots + e_t = \sum_{i=1}^{t} e_i$$

اثبات با استفاده از استقرا – پایه: به وضوح در اولین تکرار اجرای الگوریتم (حالت $t = 1$)، $w^{(1)} = (0, \dots, 0)$ بوده و در نتیجه داریم $y_1 \langle w^{(1)}, e_1 \rangle = 0$. پس می توان نوشت:

$$w^{(2)} = w^{(1)} + y_1 e_1 = w^{(1)} + 1 \times e_1 = (0, \dots, 0) + e_1 = e_1$$

بنابراین پایه برقرار می باشد.

گام: فرض کنید حکم برای $t = k$ برقرار باشد. درستی حکم را برای $t = k + 1$ نشان می دهیم.

دقت کنید که بنابه فرض استقرا داریم:

$$w^{(k)} = \sum_{i=1}^{k-1} e_i \Longrightarrow y_k \langle w^{(k)}, e_k \rangle = \langle \sum_{i=1}^{k-1} e_i \, , e_k \rangle = \sum_{i=1}^{k-1} \langle e_i, e_k \rangle = \sum_{i=1}^{k-1} 0 = 0$$

بنابراین، با توجه به اینکه $y_k \langle w^{(k)}, e_k \rangle \leq 0$ می باشد، بردار w به شکل زیر بروزرسانی می شود:

$$w^{(k+1)} = w^{(k)} + y_k e_k = \sum_{i=1}^{k-1} e_i + 1 \times e_k = \sum_{i=1}^{k} e_i$$

که این همان حکم برای $t = k + 1$ می باشد. پس گام نیز ثابت شد و حکم برقرار می باشد.

حال دقت کنید که با توجه به لم بالا، پس از m بار اجرای الگوریتم خواهیم داشت:

$$w^{(m+1)} = \sum_{i=1}^{m} e_i = (1, \dots, 1)$$

بنابراین پس از m بار اجرای الگوریتم، داریم:

$$\left\|w^{(m+1)}\right\|^2 = m \wedge \forall 1 \leq i \leq m \,; y_i \langle w^{(m+1)}, x_i \rangle = \langle w^{(m+1)}, e_i \rangle = 1 > 0$$

و این بدین معنی است که اجرای الگوریتم به اتمام می رسد و تمامی شروط مطلوب مسئله ارضا شده اند.

در نهایت، توجه کنید که در اینجا داریم:

$$B \leq \sqrt{m} \wedge R \leq 1 \Longrightarrow (RB)^2 \leq \left(1 \times \sqrt{m}\right)^2 = m$$

همچنین با توجه به قسمت الف داریم (تعداد تکرار ها برابر $T = m$ می باشد):

$$T = m \leq (BR)^2$$

از دو نامساوی بالا نتیجه می گیریم:

$$m = (RB)^2$$

به عبارت دیگر کران $(RB)^2$ در این مسئله، یک کران شارپ می باشد.