401106696 - سید کیمیا پیدایی

تمرین تئوری سوم درس یادگیری ماشین

سؤال اول

۱. (۲۵ نمره) صحیح و غلط بودن هر مورد را مشخص کنید. در موارد (الف)، (و) و (ی) نیز با استدلال پاسخ دهید.

الف) می‌دانیم که $۲^{۲^n}$ تابع boolean متمایز بر روی n ورودی وجود دارند. در نتیجه برای ۲ ورودی، ۱۶ تابع boolean متمایز وجود دارد. چند تا از این ۱۶ تابع را می‌توان توسط یک perceptron نشان داد؟ چرا؟

ب) تابع فعال‌سازی tanh به دلیل صفرمحور بودن، برخلاف تابع sigmoid مشکل ناپدیدکردن گرادیان را ندارد.

ج) هر تابع منطقی را می‌توان با یک شبکه دولایه (یک لایه مخفی) و یک تابع فعال‌سازی ReLU نمایش داد.

د) افزایش عمق و عرض یک شبکه عصبی همیشه عملکرد آن را بهبود می‌بخشد.

ه) یک شبکه عصبی سطحی و بسیار پهن می‌تواند همیشه به اندازه یک شبکه عصبی عمیق هر تابعی را به‌صورت کارآمد تقریب بزند.

و) مشکل ناپدیدشدن گرادیان در آموزش شبکه‌های عصبی چیست و چگونه توابع فعال‌سازی مانند ReLU به کاهش آن کمک می‌کنند؟

ز) Stochastic Gradient Descent (SGD) با اندازه batch کوچک به دلیل فرکانس بالای به‌روزرسانی تضمین می‌کند سریع‌تر از mini-batch gradient descent همگرا شود.

ی) با استفاده از Stochastic Gradient Descent با ممان اول، از نقطه اولیه $x_. = -۲/۸$، با نرخ یادگیری $\gamma = ۰/۰۵$، ضریب ممان $\mu = ۰/۷$ و تعداد تکرار $i = ۲$ کمینه کنید. مقادیر g_i, m_i, x_i و y_i و محاسبات هر مرحله باید ارائه شوند. (منظور از g_i گرادیان در مرحله i و m_i مقدار ممان در مرحله i می‌باشد) نیازی به اصلاح بایاس نیست. $y = ۰/۳x^۴ - ۰/۱x^۳ - ۲x^۲ - ۰/۸x$

الف) یک perceptron تنها می‌تواند داده‌ها را طبقه‌بندی کند که به صورت خطی جداپذیرند. دو ورودی چهار حالت می‌توانند داشته باشند به‌صورت ((0,0), (0,1), (1,0) و (1,1)). نمی‌توان دو تابع XOR و XNOR را با یک خط و در نتیجه با یک perceptron نشان داد. بنابراین تنها 14 تابع از 16 تابع را می‌توان با استفاده از یک perceptron نشان داد.

ب) نادرست. تابع $\tanh(x)$ در مقادیر زیاد و کم x دارای گرادیان بسیار کمی می‌باشد و در نتیجه دارای مشکل ناپدید شدن گرادیان خواهد بود.

ج) درست. با توجه به قضیهٔ Universal Approximation Theorem می‌دانیم که با استفاده از یک شبکهٔ عصبی که دارای یک لایهٔ hidden می‌باشد که دارای تابع activation غیرخطی می‌باشد، می‌توان هر تابع پیوسته‌ای را تقریب زد. به طور کلی توابع منطقی را می‌توان به صورت جمع تعدادی ضرب (Sum of Products) نوشت.

د) نادرست. این کار لزوما باعث افزایش دقت نمی‌شود و ممکن است باعث ایجاد مواردی همچون overfitting بر روی داده‌ی train بشود. همچنین بسته به نوع activation function ها می‌تواند باعث به وجود آمدن vanishing gradients بشود.

ه) درست. می‌توان شبکه‌های عصبی سطحی و پهن با تعداد زیادی نورون را استفاده کرد تا هر تابعی را تقریب زد؛ اما در اینجا نسبت به شبکه‌های عصبی عمیق تعداد زیادی نورون نیاز داریم زیرا شبکه‌های عصبی عمیق پیچیدگی‌های توابع را با افزایش عمق بهتر تقریب می‌زنند اما شبکه‌ها عصبی سطحی با افزایش پهنا.

و) درصورت وجود مشکل ناپدید شدن گرادیان، دیگر نمی‌توانیم وزن‌ها و bias ها را در شبکه‌ی عصبی update کنیم و درنتیجه دقت مدل افزایش نمی‌یابد.

تابع فعال‌سازی $ReLU(x)$ به صورت $\max(0,x)$ تعریف می‌شود و بنابراین مشتق آن به صورت $\begin{cases} 1 & \text{if } x>0 \\ 0 & \text{if } x<0 \end{cases}$ تعریف می‌شود و باتوجه به اینکه عموما با مقادیر مثبت کار داریم، به مشکل ناپدید شدن گرادیان برنخواهیم خورد.

ز) درست. در SGD با batch size کم به ازای تعداد نقاط کمتری به روزرسانی صورت می‌گیرد و درنتیجه نوسان بالاتری در به روزرسانی مقادیر خواهیم داشت که convergence سریع‌تری را نسبت به mini-batch gradient descent به ما نتیجه می‌دهد؛ اما این روش در نزدیکی نقطه(نقاط) optimal دارای oscillation می‌باشد (به این معنا که دارای تغییرات جزئی در آن نواحی می‌باشد).

ی)

SGD with momentum update rule $\begin{cases} m_{t+1} = \mu m_t + (1-\mu)\nabla J(x) \\ x_{t+1} = x_t - \gamma m_{t+1} \end{cases}$

$\mu = 0.7,\ \gamma = 0.05,\ x_0 = -2.8,\ m_0 = 0 \quad , \nabla J(x) = \frac{\partial}{\partial x}\left[0.3x^4 - 0.1x^3 - 2x^2 - 0.8x\right] = 1.2x^3 - 0.3x^2 - 4x - 0.8$

$i=1: \quad g_1 = 1.2x_0^3 - 0.3x_0^2 - 4x_0 - 0.8 = -18.2944 \longrightarrow m_1 = 0.7m_0 + 0.3g_1 = -5.48832$

$\longrightarrow x_1 = x_0 - 0.05m_1 = -2.525584 \longrightarrow y_1 = 0.3x_1^4 - 0.1x_1^3 - 2x_1^2 - 0.8x_1 = 3.08014$

$i=2: \quad g_2 = 1.2x_1^3 - 0.3x_1^2 - 4x_1 - 0.8 = -11.94279 \longrightarrow m_2 = 0.7m_1 + 0.3g_2 = -7.42466$

$\longrightarrow x_2 = x_1 - 0.05m_2 = -2.15435 \longrightarrow y_2 = 0.3x_2^4 - 0.1x_2^3 - 2x_2^2 - 0.8x_2 = -0.09679$

سؤال دوم:

۲. (۱۵ نمره) تابع $y = f(x_۱, x_۲, x_۳)$ به‌صورت زیر تعریف شده است:

$$y = ۶z_۱ + ۲z_۲ + ۴$$

$$z_۱ = \exp\left(-\frac{\max\{|x_۱ - ۳|, |x_۲ - ۵|, |x_۳ + ۲|\}^۲}{۴}\right)$$

$$z_۲ = \begin{cases} ۰, & \text{if} \sqrt{(x_۱ + ۳)^۲ + (x_۳ - ۴)^۲} > ۴ \\ ۱, & \text{otherwise} \end{cases}$$

الف) شبکه عصبی تولیدکننده این تابع را با در نظر گرفتن مدل نورون عمومی رسم کنید. برای هر نورون، تابع ورودی شبکه، تابع فعال‌سازی و تابع خروجی را برحسب وزن‌های ورودی و مقادیر پارامتری مشخص کنید. وزن‌ها و مقادیر آستانه را تعیین کنید. (نکته: جواب این سؤال یکتا نیست)

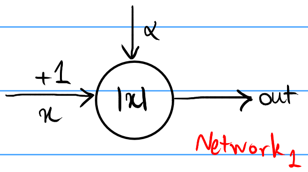
ب) با یک شبکه عصبی متشکل از نورون‌های عمومی، تابع زیر را بسازید (فرض کنید f_{net} فقط می‌تواند جمع وزن‌دار ورودی‌ها باشد و f_{out} هم فقط می‌تواند تابع همانی باشد).

$$f(x_۱, x_۲, x_۳) = \max\{x_۳, x_۱^{x_۲}\}$$

الف) ابتدا متغیرها و توابع کوچک‌تر را می‌سازیم:

در هر نورون تابع activation اش را نوشته‌ام

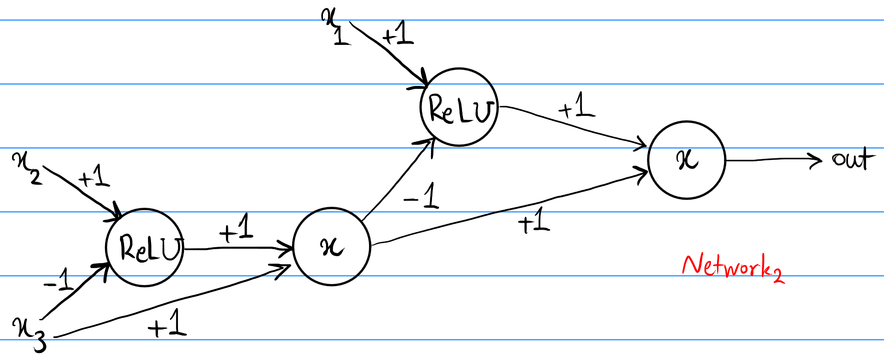
$|x+\alpha|$:



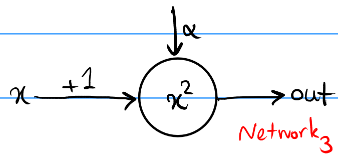
$\max\{x_1, x_2\} = ReLU(x_1 - x_2) + x_2 \longrightarrow \max\{x_1, x_2, x_3\} = \max\{x_1, \max\{x_2, x_3\}\} = \max\{x_1, ReLU(x_2 - x_3) + x_3\}$

$= ReLU(x_1 - ReLU(x_2 - x_3) - x_3) + ReLU(x_2 - x_3) + x_3$

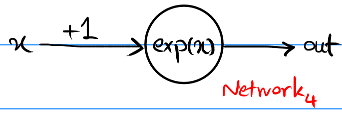
$\longrightarrow \max\{x_1, x_2, x_3\}$:



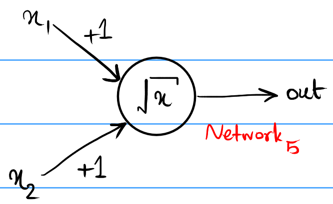
$(x+\alpha)^2$:



$\exp(x)$:

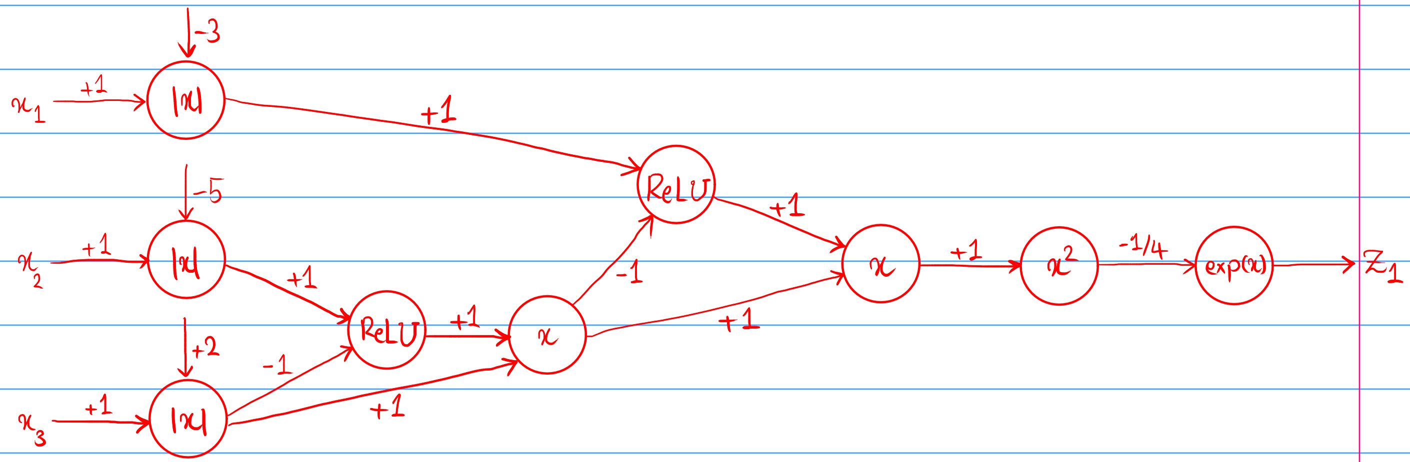


$\sqrt{x_1 + x_2}$:

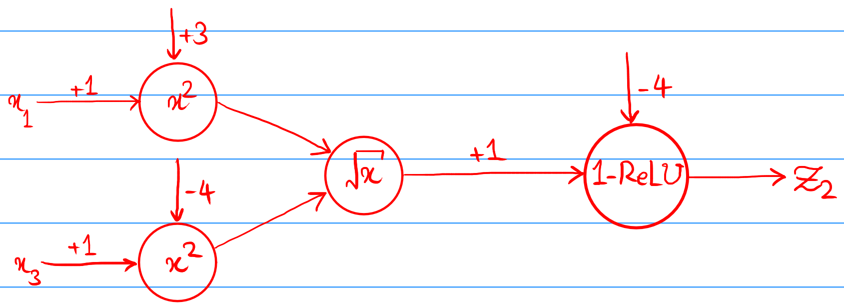


حال با استفاده از Networkهای داده شده با پارامترها، $\underline{z_1}$ و $\underline{z_2}$ را می‌سازیم:

Creating $\underline{z_1}$ using networks no. 1, 2, 3 and 4:



Creating z_2 using network no. 3, 5:

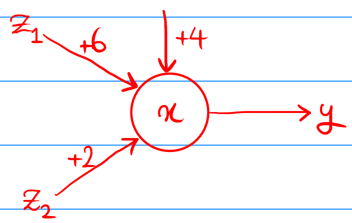


در اینجا تابع 1-ReLU به صورت زیر است:

$$1\text{-}ReLU(x)=1-\begin{cases}0 & \text{if } x\leqslant 0\\ 1 & \text{if } x>0\end{cases}=\begin{cases}1 & \text{if } x\leqslant 0\\ 0 & \text{if } x>0\end{cases}$$

حال که $\underline{z_1}$ و $\underline{z_2}$ را به دست آوردیم، می‌توانیم $\underline{y}$ را هم به دست آوریم:

$y=6z_1+2z_2+4$:

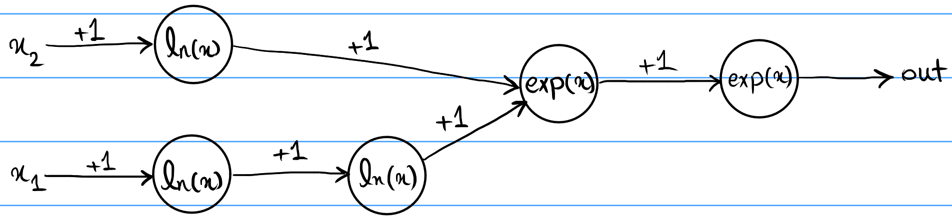


ب) در بخش قبل دیدیم که $\underline{\max\{x_1,x_2\}=ReLU(x_1-x_2)+x_2}$. حال عبارت $\underline{x_1^{x_2}}$ را به دست می‌آوریم:

$$x_1^{x_2}=(x_1)^{x_2}=(e^{\ln(x_1)})^{x_2}=e^{x_2\ln(x_1)}\ ,\ x_2\ln(x_1)=e^{\ln(x_2)}e^{\ln(\ln(x_1))}=\exp[\ln(x_2)+\ln(\ln(x_1))]$$

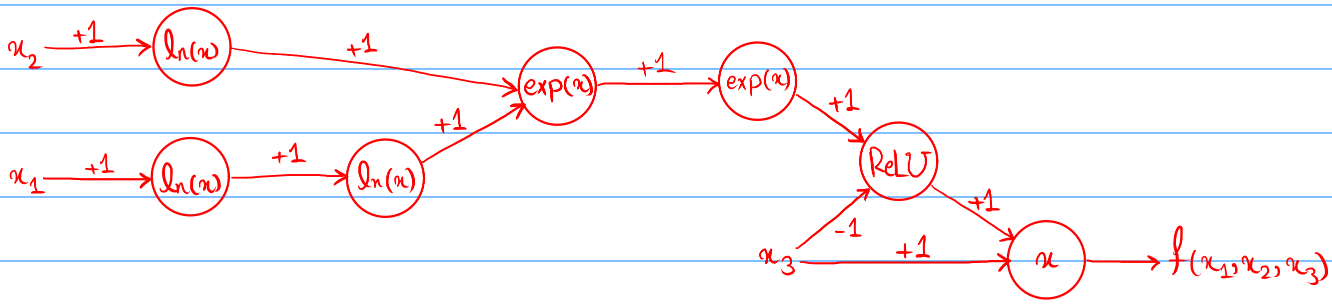
حال شبکهٔ عصبی را می‌سازیم:

$x_1^{x_2}$:



حال شبکه‌ی عصبی کلی را می‌سازیم:

$$f(x_1, x_2, x_3) = \max\{x_3, x_1^{x_2}\} = ReLU(x_1^{x_2} - x_3) + x_3$$



سؤال سوم :

۳. (۱۵ نمره) فرض کنید که یک شبکه عصبی دولایه داریم که به‌صورت زیر تعریف شده است:

$$z_1 = W_1 x^{(i)} + b_1$$

$$a_1 = \mathrm{ReLU}(z_1)$$

$$z_2 = W_2 a_1 + b_2$$

$$\hat{y}^{(i)} = \sigma(z_2)$$

$$L^{(i)} = y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

$$J = -\frac{1}{m} \sum_{i=1}^{m} L^{(i)}$$

توجه کنید که $x^{(i)}$ یک نمونه ورودی منفرد است و شکل آن $D_x \times 1$ است. همچنین $y^{(i)}$ یک برچسب خروجی است و یک اسکالر محسوب می‌شود. در دیتاست ما m نمونه وجود دارد. ما از تعداد D_{a_1} نود در لایه مخفی استفاده می‌کنیم، بنابراین شکل z_1 برابر $D_{a_1} \times 1$ است.

۱. ابعاد W_1، b_1، W_2 و b_2 چیست؟ اگر بخواهیم شبکه را روی چندین نمونه vectorize کنیم، شکل وزن‌ها و بایاس‌ها چه تغییری خواهد کرد؟ اگر بخواهیم شبکه را روی چندین نمونه ورودی vectorize کنیم، شکل‌های X و Y چگونه خواهد بود؟

۲. مقدار $\frac{\partial J}{\partial \hat{y}^{(i)}}$ چیست؟ این مقدار را $\delta_1^{(i)}$ نامگذاری کنید. با استفاده از این نتیجه، مقدار $\frac{\partial J}{\partial \hat{y}}$ چیست؟

۳. مقدار $\frac{\partial \hat{y}^{(i)}}{\partial z_2}$ چیست؟ این مقدار را $\delta_2^{(i)}$ نامگذاری کنید.

۴. مقدار $\frac{\partial z_2}{\partial a_1}$ چیست؟ این مقدار را $\delta_3^{(i)}$ بنامید.

۵. مقدار $\frac{\partial a_1}{\partial z_1}$ چیست؟ این مقدار را $\delta_4^{(i)}$ بنامید.

۶. مقدار $\frac{\partial z_1}{\partial W_1}$ چیست؟ این مقدار را $\delta_5^{(i)}$ بنامید.

۷. مقدار $\frac{\partial J}{\partial W_1}$ چیست؟ می‌توانید از نتایج قبلی استفاده کنید.

.1

$z_1 = W_1 x^{(i)} + b_1 \longrightarrow \text{shape}(z) = \text{shape}(W_1 x^{(i)} + b_1) = D_{a_1} \times 1 \longrightarrow \text{shap}(b_1) = D_{a_1} \times 1$

$\longrightarrow \text{shape}(W_1 x^{(i)}) = D_{a_1} \times 1 \ , \ \text{shap}(x^{(i)}) = D_x \times 1 \longrightarrow \text{shape}(W_1) = D_{a_1} \times D_x$

$a_1 = \mathrm{ReLU}(z_1) \longrightarrow \text{shape}(a_1) = \text{shape}(z_1) = D_{a_1} \times 1$

$\hat{y}^{(i)} = \sigma(z_2) \longrightarrow \text{shape}(z_2) = \text{shape}(\hat{y}^{(i)}) = 1 \times 1 \quad (\hat{y}^{(i)} \text{ is a scalar})$

$, \ z_2 = W_2 a_1 + b_2 \longrightarrow \text{shape}(W_2 a_1 + b_2) = \text{shape}(z_2) = 1 \times 1 \longrightarrow \text{shape}(b_2) = 1 \times 1$

$\longrightarrow \text{shape}(W_2 a_1) = 1 \times 1, \ \text{shape}(a_1) = D_{a_1} \times 1 \longrightarrow \text{shape}(W_2) = 1 \times D_{a_1}$

فرض کنید درحال vectorize کردن شبکهٔ عصبی بر روی $\underline{N}$ داده هستیم.

$z_1 = W_1 X + b_1 \ , \ \text{shape}(X) = D_x \times N \xrightarrow{W_1\text{'s shape is independent from } X\text{'s second dimension}} \text{shape}(W_1) = D_{a_1} \times D_x$

$\longrightarrow \text{shape}(Z_1) = \text{shape}(W_1 X + b_1) = \text{shape}(W_1 X) = D_{a_1} \times N, \ \text{shape}(Z_1) = \text{shape}(b_1)$

$\longrightarrow \text{shape}(b_1) = D_{a_1} \times N \ , \ \text{shape}(a_1) = \text{shape}(z_1) = D_{a_1} \times N \ , \ \text{shape}(z_2) = 1 \times N \ (z_2 \text{ is a scalar})$

$\xrightarrow{W_2\text{'s shape is independent from } a_1\text{'s second dimension}} \text{shape}(W_2) = 1 \times D_{a_1} \ , \ \text{shape}(b_2) = \text{shape}(z_2)$

$\longrightarrow \text{shape}(b_2) = 1 \times N \qquad \text{shape}(\hat{Y}) = \text{shape}(z_2) \longrightarrow \text{shape}(\hat{Y}) = 1 \times N$

.2

$$\delta_1^{(i)} = \frac{\partial J}{\partial \hat{y}^{(i)}} = \frac{\partial}{\partial \hat{y}^{(i)}}\left[-\frac{1}{m}\sum_{j=1}^{m}\left(y^{(j)}\log(\hat{y}^{(j)}) + (1-y^{(j)})\log(1-\hat{y}^{(j)})\right)\right] = -\frac{1}{m}\sum_{j=1}^{m}\frac{\partial}{\partial \hat{y}^{(i)}}\left[y^{(j)}\log(\hat{y}^{(j)}) + (1-y^{(j)})\log(1-\hat{y}^{(j)})\right]$$

$$= -\frac{1}{m}\frac{\partial}{\partial \hat{y}^{(i)}}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right] = -\frac{1}{m}\left[\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1-y^{(i)}}{1-\hat{y}^{(i)}}\right]$$

$$\longrightarrow \delta_1^{(i)} = \frac{\partial J}{\partial \hat{y}^{(i)}} = -\frac{1}{m}\left(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1-y^{(i)}}{1-\hat{y}^{(i)}}\right)$$

$$\longrightarrow \frac{\partial J}{\partial \hat{Y}} = \begin{bmatrix} \frac{\partial J}{\partial \hat{y}^{(1)}} & \frac{\partial J}{\partial \hat{y}^{(2)}} & \cdots & \frac{\partial J}{\partial \hat{y}^{(m)}} \end{bmatrix} \longrightarrow \frac{\partial J}{\partial \hat{Y}} = -\frac{1}{m}\begin{bmatrix} \frac{y^{(1)}}{\hat{y}^{(1)}} - \frac{1-y^{(1)}}{1-\hat{y}^{(1)}} & \cdots & \frac{y^{(m)}}{\hat{y}^{(m)}} - \frac{1-y^{(m)}}{1-\hat{y}^{(m)}} \end{bmatrix}$$

.3

$$\delta_2^{(i)} = \frac{\partial \hat{y}^{(i)}}{\partial z_2} = \frac{\partial}{\partial z_2}\sigma(z_2) = \sigma(z_2)(1-\sigma(z_2)) \longrightarrow \delta_2^{(i)} = \frac{\partial \hat{y}^{(i)}}{\partial z_2} = \sigma(z_2)\left(1-\sigma(z_2)\right) = \hat{y}^{(i)}(1-\hat{y}^{(i)})$$

.4

$$\delta_3^{(i)} = \frac{\partial z_2}{\partial a_1} = \frac{\partial}{\partial a_1}\left[W_2 a_1 + b_2\right] = W_2 \longrightarrow \delta_3^{(i)} = \frac{\partial z_2}{\partial a_1} = W_2$$

.5

$$\delta_4^{(i)} = \frac{\partial a_1}{\partial z_1} = \frac{\partial}{\partial z_1}\mathrm{ReLU}(z_1) = \begin{cases} \partial z_1/\partial z_1 = 1 & \text{if } z_1 > 0 \\ \partial/\partial z_1\, 0 = 0 & \text{o.w.} \end{cases} \longrightarrow \delta_4^{(i)} = \frac{\partial a_1}{\partial z_1} = \begin{cases} 1 & \text{if } z_1 > 0 \\ 0 & \text{o.w.} \end{cases}$$

.6

$$\delta_5^{(i)} = \frac{\partial z_1}{\partial W_1} = \frac{\partial}{\partial W_1}\left[W_1 x^{(i)} + b_1\right] = x^{(i)T} \longrightarrow \delta_5^{(i)} = \frac{\partial z_1}{\partial W_1} = x^{(i)}$$

.7

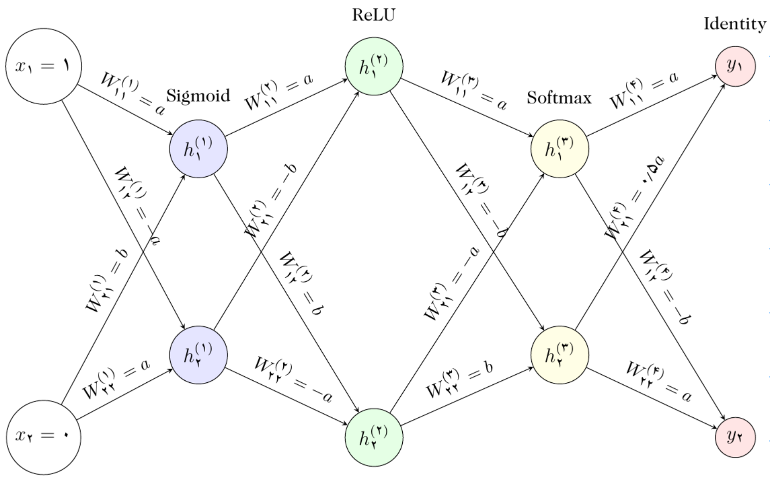
$$\frac{\partial J}{\partial W_1} = \frac{\partial J}{\partial \hat{y}^{(i)}}\frac{\partial \hat{y}^{(i)}}{\partial z_2}\frac{\partial z_2}{\partial a_1}\frac{\partial a_1}{\partial z_1}\frac{\partial z_1}{\partial W_1} = \delta_1^{(i)}\delta_2^{(i)}\delta_3^{(i)}\delta_4^{(i)}\delta_5^{(i)}$$

$$= -\frac{1}{m}\left(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1-y^{(i)}}{1-\hat{y}^{(i)}}\right)\hat{y}^{(i)}(1-\hat{y}^{(i)})\, W_2 x^{(i)}\, \delta_4^{(i)} = -\frac{1}{m}\left[y^{(i)}(1-\hat{y}^{(i)}) - \hat{y}^{(i)}(1-y^{(i)})\right] W_2 x^{(i)}\, \delta_4^{(i)}$$

$$\longrightarrow \frac{\partial J}{\partial W_1} = \begin{cases} \dfrac{\hat{y}^{(i)}(1-y^{(i)}) - y^{(i)}(1-\hat{y}^{(i)})}{m} W_2 x^{(i)} & \text{if } W_1 x^{(i)} + b_1 > 0 \\ 0 & \text{o.w.} \end{cases}$$

سؤال چهارم:

۴. (۳۰ نمره) یک شبکه عصبی feedforward با معماری زیر را در نظر بگیرید:



- **لایه ورودی:** ۲ نورون
- **لایه مخفی ۱:** ۲ نورون، با استفاده از تابع فعال‌سازی Sigmoid
- **لایه مخفی ۲:** ۲ نورون، با استفاده از تابع فعال‌سازی ReLU
- **لایه مخفی ۳:** ۲ نورون، با استفاده از تابع فعال‌سازی Softmax
- **لایه خروجی:** ۲ نورون، با استفاده از تابع فعال‌سازی Identity

وزن‌ها و بایاس‌های اولیه شبکه به صورت زیر است:

وزن‌ها و بایاس‌ها از لایه ورودی به لایه مخفی ۱ (لایه ۱):

$$W^{(1)} = \begin{bmatrix} a & -a \\ b & a \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

وزن‌ها و بایاس‌ها از لایه مخفی ۱ به لایه مخفی ۲ (لایه ۲):

$$W^{(2)} = \begin{bmatrix} a & b \\ -b & -a \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

وزن‌ها و بایاس‌ها از لایه مخفی ۲ به لایه مخفی ۳ (لایه ۳):

$$W^{(3)} = \begin{bmatrix} a & -b \\ -a & b \end{bmatrix}, \quad b^{(3)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

وزن‌ها و بایاس‌ها از لایه مخفی ۳ به لایه خروجی (لایه ۴):

$$W^{(4)} = \begin{bmatrix} a & -b \\ 0.5a & a \end{bmatrix}, \quad b^{(4)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

بردار ورودی x و خروجی هدف t به صورت زیر است:

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad t = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

با استفاده از نرخ یادگیری (η)، یک بار فرآیند **backpropagation** را انجام دهید تا **وزن‌ها** و **بایاس‌های** شبکه به‌روزرسانی شوند. از توابع فعال‌سازی زیر در هر لایه استفاده کنید:

- **لایه ۱ (Sigmoid):** $\sigma(x) = \frac{1}{1 + e^{-x}}$
- **لایه ۲ (ReLU):** $\mathrm{ReLU}(x) = \max(0, x)$
- **لایه ۳ (Softmax):** $\mathrm{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$
- **لایه ۴ (Identity):** $f(x) = x$

فرض کنید تابع خطا به صورت میانگین مربعات (MSE) تعریف شده است:

$$E = \frac{1}{2} \sum_{k=1}^{2} (a_k^{(4)} - t_k)^2$$

که در آن $a^{(4)}$ خروجی شبکه است.

سوال:

الف) یک forward pass از شبکه انجام دهید تا خروجی $a^{(4)}$ را محاسبه کنید.

ب) خطای E را با استفاده از تابع خطای MSE محاسبه کنید.

ج) یک backward pass انجام دهید تا گرادیان‌های خطا نسبت به تمام وزن‌ها و بایاس‌ها را محاسبه کنید.

د) وزن‌ها و بایاس‌ها را با استفاده از gradient descent و نرخ یادگیری η به‌روزرسانی کنید.

الف) ورودی را از تمامی لایه‌ها می‌گذرانیم تا forward pass انجام شود:

$$z^{(1)} = W^{(1)}x + b^{(1)} = \begin{bmatrix} a & -a \\ b & a \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \rightarrow a^{(1)} = \sigma(z^{(1)}) = \begin{bmatrix} \frac{1}{1+e^{-a}} \\ \frac{1}{1+e^{-b}} \end{bmatrix}$$

$$\rightarrow z^{(2)} = W^{(2)}a^{(1)} + b^{(2)} = \begin{bmatrix} a & b \\ -b & -a \end{bmatrix}\begin{bmatrix} \frac{1}{1+e^{-a}} \\ \frac{1}{1+e^{-b}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{a}{1+e^{-a}} + \frac{b}{1+e^{-b}} \\ \frac{-b}{1+e^{-a}} - \frac{a}{1+e^{-b}} \end{bmatrix}$$

$$\rightarrow a^{(2)} = \mathrm{ReLU}(z^{(2)}) = \begin{bmatrix} \max\left(0, \frac{a}{1+e^{-a}} + \frac{b}{1+e^{-b}}\right) \\ \max\left(0, \frac{-b}{1+e^{-a}} - \frac{a}{1+e^{-b}}\right) \end{bmatrix} = \begin{bmatrix} \frac{a}{1+e^{-a}} + \frac{b}{1+e^{-b}} \\ 0 \end{bmatrix}$$

$$\longrightarrow z^{(3)} = W^{(3)} a^{(2)} + b^{(3)} = \begin{bmatrix} a & -b \\ -a & b \end{bmatrix} \begin{bmatrix} \frac{a}{1+e^{-a}} + \frac{b}{1+e^{-b}} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}} \\ -\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}} \end{bmatrix}$$

$$\longrightarrow a^{(3)} = \mathrm{softmax}(z^{(3)}) = \frac{1}{\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) + \exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right)} \begin{bmatrix} \exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) \\ \exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right) \end{bmatrix}$$

$$\longrightarrow z^{(4)} = W^{(4)} a^{(3)} + b^{(4)} = \begin{bmatrix} a & -b \\ 0.5a & a \end{bmatrix} \begin{bmatrix} \exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) \\ \exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right) \end{bmatrix} \frac{1}{\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) + \exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right)} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$= \frac{1}{\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) + \exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right)} \begin{bmatrix} a.\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) - b.\exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right) \\ \frac{a}{2}\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) + a.\exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right) \end{bmatrix}$$

$$\longrightarrow a^{(4)} = \mathrm{Identity}(z^{(4)}) = z^{(4)}$$

$$= \frac{1}{\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) + \exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right)} \begin{bmatrix} a.\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) - b.\exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right) \\ \frac{a}{2}\exp\left(\frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}\right) + a.\exp\left(-\frac{a^2}{1+e^{-a}} - \frac{ab}{1+e^{-b}}\right) \end{bmatrix}$$

$$c := \frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}$$

(ب

$$\longrightarrow E = \frac{1}{2}\sum_{k=1}^{2}\left(a_k^{(4)} - t_k\right)^2 = \frac{1}{2}\left[\left(a_1^{(4)} - t_1\right)^2 + \left(a_2^{(4)} - t_2\right)^2\right] = \frac{1}{2}\left[a_1^{(4)2} + \left(a_2^{(4)} - 1\right)^2\right]$$

$$= \frac{1}{2}\,\frac{1}{(e^c + e^{-c})^2}\left[\left(ae^c - be^{-c}\right)^2 + \left(\frac{a}{2}e^c + ae^{-c} - 1\right)^2\right]$$

$$= \frac{1}{2(e^{2c} + e^{-2c} + 2)}\left[a^2e^{2c} + b^2e^{-2c} - 2ab + \frac{a^2}{4}e^{2c} + a^2e^{-2c} + a^2 + 1 - ae^c - 2ae^{-c}\right]$$

$$\longrightarrow E = \frac{1}{2(e^{2c} + e^{-2c} + 2)}\left[\frac{5}{4}a^2e^{2c} + (a^2+b^2)e^{-2c} - (e^c + 2e^{-c})a + (a^2 - 2ab + 1)\right], \quad c = \frac{a^2}{1+e^{-a}} + \frac{ab}{1+e^{-b}}$$

ج) با همفکری با آقای آرین خسروی

$$E = \frac{1}{2}\|a^{(4)} - t\|_2^2 \rightarrow \frac{\partial E}{\partial a^{(4)}} = a^{(4)} - t$$

$$\rightarrow \frac{\partial E}{\partial W^{(4)}} = \frac{\partial E}{\partial a^{(4)}}\frac{\partial a^{(4)}}{\partial W^{(4)}} = \frac{\partial E}{\partial a^{(4)}}\frac{\partial a^{(4)}}{\partial z^{(4)}}\frac{\partial z^{(4)}}{\partial W^{(4)}} = (a^{(4)} - t)a^{(3)^T}$$

$$, \frac{\partial E}{\partial b^{(4)}} = \frac{\partial E}{\partial a^{(4)}}\frac{\partial a^{(4)}}{\partial z^{(4)}}\frac{\partial z^{(4)}}{\partial b^{(4)}} = a^{(4)} - t$$

$$\rightarrow \frac{\partial E}{\partial a^{(3)}} = \frac{\partial E}{\partial a^{(4)}}\frac{\partial a^{(4)}}{\partial a^{(3)}} = \frac{\partial E}{\partial a^{(4)}}\frac{\partial a^{(4)}}{\partial z^{(4)}}\frac{\partial z^{(4)}}{\partial a^{(3)}} = W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial z^{(3)}} = \frac{\partial E}{\partial a^{(3)}}\frac{\partial a^{(3)}}{\partial z^{(3)}} = \left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial W^{(3)}} = \frac{\partial E}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial W^{(3)}} = \left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)\,a^{(2)^T}$$

$$, \frac{\partial E}{\partial b^{(3)}} = \frac{\partial E}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial b^{(3)}} = \left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial a^{(2)}} = \frac{\partial E}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial a^{(2)}} = W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial z^{(2)}} = \frac{\partial E}{\partial a^{(2)}}\frac{\partial a^{(2)}}{\partial z^{(2)}} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial W^{(2)}} = \frac{\partial E}{\partial z^{(2)}}\frac{\partial z^{(2)}}{\partial W^{(2)}} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)\,a^{(1)^T}$$

$$, \frac{\partial E}{\partial b^{(2)}} = \frac{\partial E}{\partial z^{(2)}}\frac{\partial z^{(2)}}{\partial b^{(2)}} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial a^{(1)}} = \frac{\partial E}{\partial z^{(2)}}\frac{\partial z^{(2)}}{\partial a^{(1)}} = W^{(2)^T}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial z^{(1)}} = \frac{\partial E}{\partial a^{(1)}}\frac{\partial a^{(1)}}{\partial z^{(1)}} = \mathrm{diag}\left(a^{(1)}(1 - a^{(1)})\right) W^{(2)^T}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

$$\rightarrow \frac{\partial E}{\partial W^{(1)}} = \frac{\partial E}{\partial z^{(1)}}\frac{\partial z^{(1)}}{\partial W^{(1)}} = \mathrm{diag}\left(a^{(1)}(1 - a^{(1)})\right) W^{(2)^T}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)\,x^T$$

$$, \frac{\partial E}{\partial b^{(1)}} = \frac{\partial E}{\partial z^{(1)}}\frac{\partial z^{(1)}}{\partial b^{(1)}} = \mathrm{diag}\left(a^{(1)}(1 - a^{(1)})\right) W^{(2)^T}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} W^{(3)^T}\left(\mathrm{diag}(a^{(3)}) - a^{(3)}a^{(3)^T}\right)W^{(4)^T}(a^{(4)} - t)$$

د) برای به روزرسانی صرفا کافیست که هر پارامتر را با استفاده از مشتقاتی که در بالا به دست آوردیم، کاهش دهیم:

$$W^{(1)^{(t+1)}} = W^{(1)^{(t)}} - \eta\left(\frac{\partial E}{\partial W^{(1)}}\right)^{(t)} , \quad b^{(1)^{(t+1)}} = b^{(1)^{(t)}} - \eta\left(\frac{\partial E}{\partial b^{(1)}}\right)^{(t)}$$

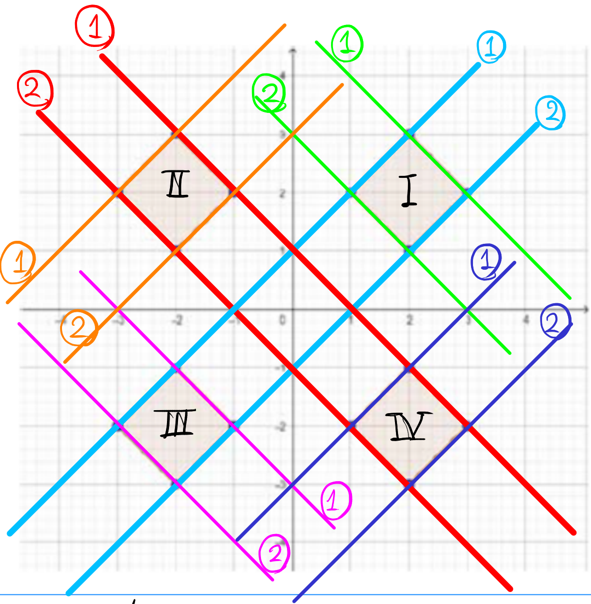
$$W^{(2)^{(t+1)}} = W^{(2)^{(t)}} - \eta\left(\frac{\partial E}{\partial W^{(2)}}\right)^{(t)} , \quad b^{(2)^{(t+1)}} = b^{(2)^{(t)}} - \eta\left(\frac{\partial E}{\partial b^{(2)}}\right)^{(t)}$$

$$W^{(3)^{(t+1)}} = W^{(3)^{(t)}} - \eta\left(\frac{\partial E}{\partial W^{(3)}}\right)^{(t)} , \quad b^{(3)^{(t+1)}} = b^{(3)^{(t)}} - \eta\left(\frac{\partial E}{\partial b^{(3)}}\right)^{(t)}$$

$$W^{(4)^{(t+1)}} = W^{(4)^{(t)}} - \eta\left(\frac{\partial E}{\partial W^{(4)}}\right)^{(t)} , \quad b^{(4)^{(t+1)}} = b^{(4)^{(t)}} - \eta\left(\frac{\partial E}{\partial b^{(4)}}\right)^{(t)}$$

سؤال پنجم:

الف) یک شبکه با حداقل تعداد TLU به گونه‌ای طراحی کنید که برای ورودی‌های داخل ناحیه‌های شکل زیر خروجی یک بدهد و برای خارج آن صفر. ساختار شبکه به همراه وزن‌ها و بایاس‌های آن را مشخص کنید.



ابتدا معادله خطوطی که مشخص شده‌اند را به‌دست می‌آوریم و سپس به بررسی اینکه آیا نقطه مربوطه در محدوده مشخص می‌باشد یا خیر می‌پردازیم:

$\begin{cases}1: y=x+1\\2: y=x-1\end{cases}$ $\begin{cases}1: y=-x+5\\2: y=-x+3\end{cases}$ $\begin{cases}1: y=-x+1\\2: y=-x-1\end{cases}$ $\begin{cases}1: y=x+5\\2: y=x+3\end{cases}$ $\begin{cases}1: y=-x-3\\2: y=-x-5\end{cases}$ $\begin{cases}1: y=x-3\\2: y=x-5\end{cases}$

حال شروط لازم برای قرارگیری در هر یک از نواحی را به‌دست می‌آوریم:

$\mathrm{I}:\begin{cases}y\leqslant x+1\\y\geqslant x-1\\y\leqslant -x+5\\y\geqslant -x+3\end{cases} \longrightarrow \mathrm{I}:\begin{cases}3\leqslant y+x\leqslant 5\\\boxed{-1\leqslant y-x\leqslant 1}\end{cases}$

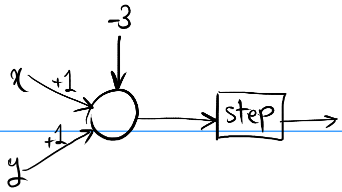
$\mathrm{II}:\begin{cases}y\leqslant -x+1\\y\geqslant -x-1\\y\leqslant x+5\\y\geqslant x+3\end{cases} \longrightarrow \mathrm{II}:\begin{cases}\boxed{-1\leqslant y+x\leqslant 1}\\3\leqslant y-x\leqslant 5\end{cases}$

$\mathrm{III}:\begin{cases}y\leqslant x+1\\y\geqslant x-1\\y\leqslant -x-3\\y\geqslant -x-5\end{cases} \longrightarrow \mathrm{III}:\begin{cases}-5\leqslant y+x\leqslant -3\\\boxed{-1\leqslant y-x\leqslant 1}\end{cases}$

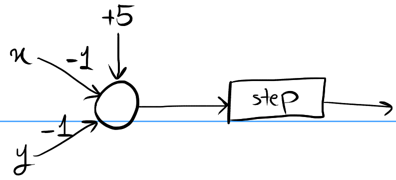
$\mathrm{IV}:\begin{cases}y\leqslant -x+1\\y\geqslant -x-1\\y\leqslant x-3\\y\geqslant x-5\end{cases} \longrightarrow \mathrm{IV}:\begin{cases}\boxed{-1\leqslant y+x\leqslant 1}\\-5\leqslant y-x\leqslant -3\end{cases}$

بنابراین تعداد شروطی که باید بررسی کنیم برابر با 12 تا می‌باشد.

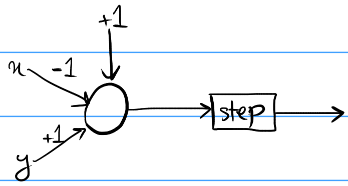
$3 \leqslant y+x$:



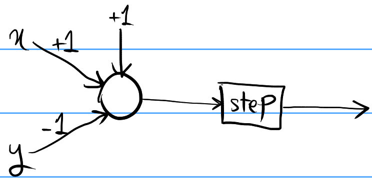
$x+y \leqslant 5$:



$-1 \leqslant y-x$:



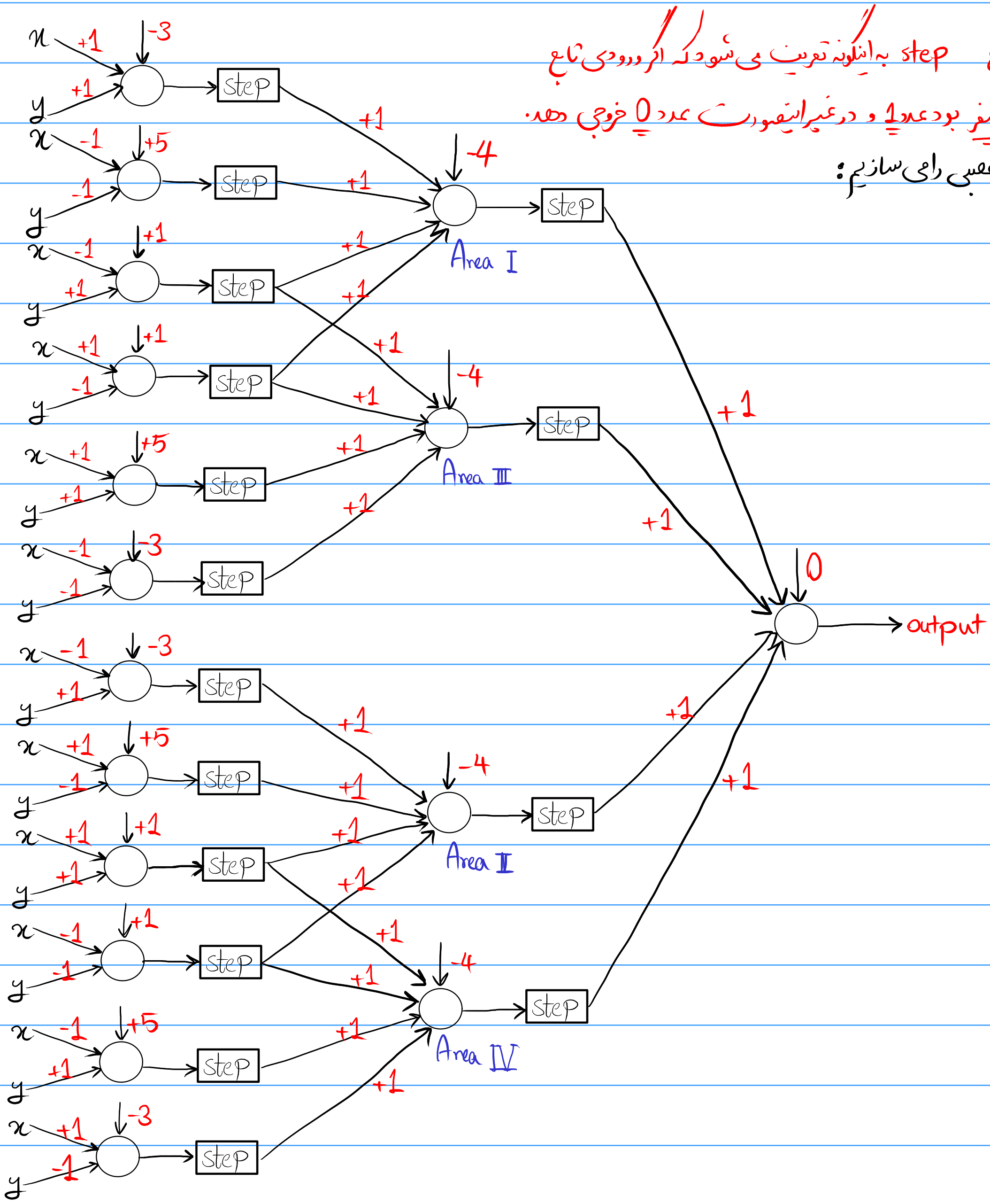
$y-x \leqslant 1$:



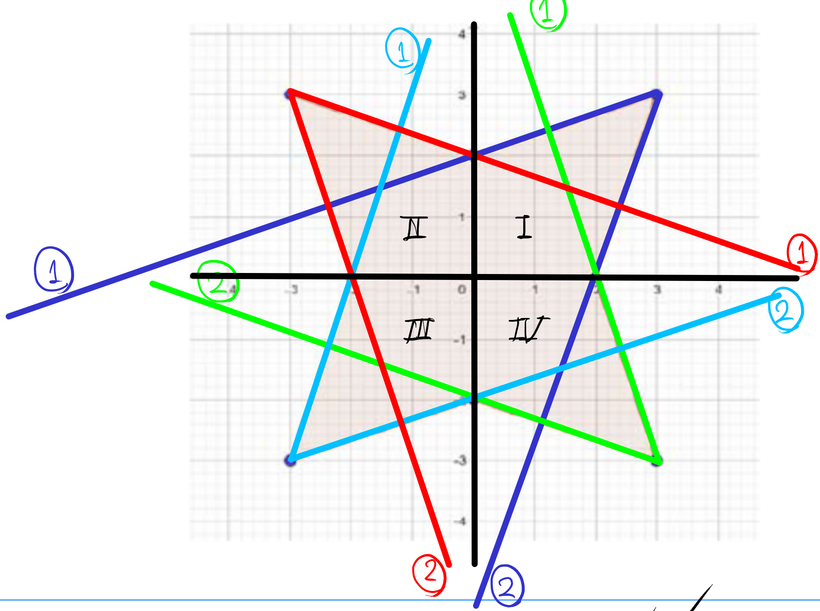
و ...

در طول حل این بخش تابع step به اینگونه تعریف می‌شود که اگر ورودی تابع بزرگتر یا مساوی با صفر بود عدد 1 و در غیر اینصورت عدد 0 خروجی دهد.

به همین ترتیب شبکهٔ عصبی را می‌سازیم:



ب) آیا می‌توان با استفاده از شبکه طراحی شده در قسمت (الف) و صرفا با تغییر پارامترهای آن، طبقه‌بندی برای شکل زیر بدست آورد؟ توضیح دهید.



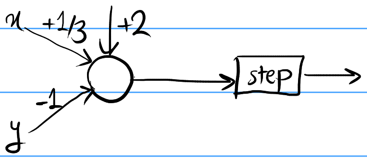
بله می‌توان. به این صورت عمل می‌کنیم که برای هر محدوده دوباره کارهای قبلی را انجام می‌دهیم:

مثلاً برای بخش I داریم:

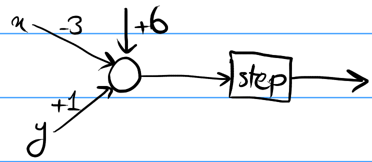
$$\begin{cases} 1: y = \frac{1}{3}x + 2 \\ 2: y = 3x - 6 \end{cases} \longrightarrow I: \begin{cases} y \leqslant \frac{1}{3}x + 2 \\ y \geqslant 3x - 6 \\ x \geqslant 0 \\ y \geqslant 0 \end{cases}$$

حال نورون‌های مربوط به این چهار شرط را می‌سازیم:

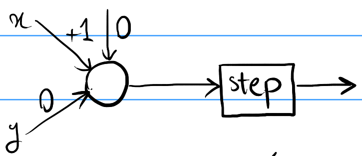
$y \leqslant \frac{1}{3}x + 2$:



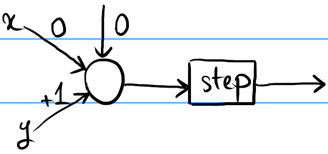
$y \geqslant 3x - 6$:



$x \geqslant 0$:



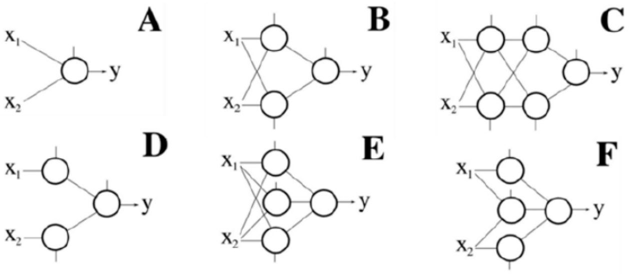
$y \geqslant 0$:

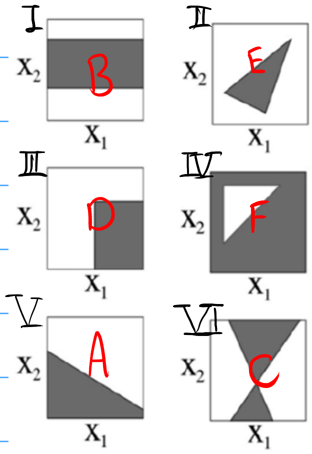


به همین ترتیب بقیه‌ی نورون‌ها را هم می‌سازیم. نیاز داریم که 12 شرط را بررسی کنیم (مانند بخش قبل)
پس در نتیجه می‌توان با تغییر وزن‌های شبکه‌ی عصبی قبلی هم این شکل را شناسایی کرد.

ج) شبکه‌های TLU زیر را در نظر بگیرید. ابتدا مشخص کنید هرکدام از پاسخ‌ها توسط کدام شبکه تولید شده است و سپس علت را توضیح دهید. فرض کنید که هرکدام از شبکه‌ها فقط یک پاسخ را ایجاد کرده است.

در شکل صفحه‌ی بعد مشخص کرده‌ایم که هر شکل توسط کدام شبکه‌ی عصبی ساخته شده است.





شکل I توسط شبکه‌ی عصبی B ایجاد می‌شود زیرا در لایه‌ی اول دو نورون داریم که برروی X_2، threshold می‌گذارند و در لایه‌ی دوم از خروجی آنها AND می‌گیریم که به ما شکل مربوطه را خروجی می‌دهد.

شکل II توسط شبکه‌ی عصبی E ایجاد می‌شود زیرا در لایه‌ی اول سه نورون دارد که برروی ترکیب خطی‌های X_1 و X_2، threshold می‌گذارند و در لایه‌ی دوم از خروجی این سه نورون، AND می‌گیریم که یک مثلث با سه ضلع مورب می‌باشد.

شکل III توسط شبکه‌ی عصبی D ایجاد می‌شود زیرا در لایه‌ی اول دو نورون داریم که هر یک برروی یکی از X_1 و X_2، threshold می‌گذارند و در لایه‌ی دوم از خروجی آنها، AND می‌گیریم که یک مستطیل را به ما نتیجه می‌دهد.

شکل IV توسط شبکه‌ی عصبی F ایجاد می‌شود زیرا در لایه‌ی اول یک نورون دارد که صرفا بر X_1 و یک نورون دیگر دارد که برروی X_2، threshold می‌گذارد و یک نورون دیگر برروی ترکیب خطی X_1 و X_2، threshold می‌گذارد. در لایه‌ی سوم از خروجی نورون‌های 2 قبلی، AND می‌گیریم که یک مثلث با یک ضلع افقی، یک ضلع عمودی و یک ضلع مورب می‌باشد.

شکل V توسط شبکه‌ی عصبی A ایجاد می‌شود زیرا صرفا یک نورون دارد که خروجی‌اش ترکیب خطی‌ای از X_1 و X_2 می‌باشد.

شکل VI توسط شبکه‌ی عصبی C ایجاد می‌شود زیرا در لایه‌ی اول دو نورون دارد که خروجی هرکدام ترکیب خطی‌ای از X_1 و X_2 می‌باشد و در لایه‌ی آخر اشکال ورودی داده شده در هر دو نیم صفحه‌ی مربوط به نورون‌ها می‌باشد بررسی می‌شود.

سؤال ششم:

۶. (امتیازی_ ۱۵ نمره) شبکه عصبی با معماری A را در نظر بگیرید. این شبکه معادل با تابع $F_A(x;W)$ است که در آن x ورودی و W تمام پارامترهای شبکه می‌باشد. فرض کنید تابع activation این شبکه ReLU باشد. آنگاه برای شبکه مفهومی به نام activation pattern تعریف می‌کنیم. activation pattern یک رشته از ۰ و ۱ به طول تعداد نورون‌های شبکه است. ۰ بودن به‌معنای غیرفعال بودن نورون و ۱ بودن به‌معنای فعال بودن است که در تابع ReLU معادل صفر یا غیرصفر بودن مقدار هر نورون است. تعداد این pattern ها را با $\mathcal{A}(F_A(x;W))$ نشان می‌دهیم.

الف) شبکه عصبی $F_A(\mathbb{R}^m;W)$ را با ReLU به‌عنوان تابع activation در نظر بگیرید. نشان دهید فضای ورودی با استفاده از این شبکه به polytope (اشکال هندسی با سطوح صاف)های convex تقسیم می‌شود که $F_A(\mathbb{R}^m;W)$ متناظر با یک تابع خطی در نواحی مختلف این فضا می‌باشد.

ب) معماری $A_{(n,k)}$ را در نظر بگیرید که معادل یک شبکه عصبی fully connected با n لایه با عرض k می‌باشد. ثابت کنید $\mathcal{A}(F_{A_{(n,k)}}(\mathbb{R}^m;W))$ دارای حد بالای $\mathcal{O}(k^{mn})$ با در نظر گرفتن تابع ReLU به‌عنوان activation می‌باشد.

راهنمایی: فرض کنید k ابرصفحه در فضای $\mathbb{R}^m$ وجود دارند که هر کدام متناظر با یک معادله به فرم $a_i^T x = b_i$ هستند. اگر تعداد نواحی تشکیل‌شده در فضا $r(k,m)$ باشد، خواهیم داشت:

$$r(k,m) \leqslant \sum_{i=۰}^{m} \binom{k}{i}$$

باهم‌فکری با آقای علی نجار

الف) هر لایه‌ی شبکه‌ی عصبی دارای خروجی‌ای به فرم $a^{(l)} = ReLU(z^{(l)})$ می‌باشد که l شماره‌ی لایه می‌باشد. همچنین خروجی هر لایه را بر اساس لایه‌ی قبل می‌توان به صورت $z^{(l+1)} = W^{(l)} \hat{a}^{(l)}$ نوشت که $\hat{a}^{(l)} = \begin{bmatrix} 1 \\ a^{(l)} \end{bmatrix}$.

حال با استفاده از activation pattern می‌توانیم فعال بودن نورون‌های شبکه‌ی عصبی را بررسی کنیم ($s_i^{(l)}$ نشان دهنده‌ی فعال بودن یا نبودن نورون iاُم در لایه‌ی lاُم می‌باشد).

$$\rightsquigarrow a_i^{(l)} = ReLU(z_i^{(l)}) = \begin{cases} z_i^{(l)} & \text{if } z_i^{(l)} \geqslant 0 \\ 0 & \text{o.w.} \end{cases} = \begin{cases} z_i^{(l)} & \text{if } s_i^{(l)} = 1 \\ 0 & \text{o.w.} \end{cases} \quad (\text{I})$$

فرض کنید حالت (status) تمامی نورون‌های لایه‌ی lاُم را در $s^{(l)}$ قرار دهیم. داریم:

$$s^{(l)} = \begin{bmatrix} s_1^{(l)} \\ s_2^{(l)} \\ \vdots \\ s_n^{(l)} \end{bmatrix}$$

$$\rightsquigarrow a^{(l)} = ReLU(z^{(l)}) = \begin{bmatrix} ReLU(z_1^{(l)}) \\ ReLU(z_2^{(l)}) \\ \vdots \\ ReLU(z_n^{(l)}) \end{bmatrix} = \begin{bmatrix} s_1^{(l)} z_1^{(l)} \\ s_2^{(l)} z_2^{(l)} \\ \vdots \\ s_n^{(l)} z_2^{(l)} \end{bmatrix}$$

$$\longrightarrow \hat{z}^{(l+1)} = W^{(l)}\hat{a}^{(l)} = W^{(l)}\begin{bmatrix} 1 \\ s_1^{(l)} z_1^{(l)} \\ s_2^{(l)} z_2^{(l)} \\ \vdots \\ s_n^{(l)} z_n^{(l)} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ b_i^{(l)} & w_{1i}^{(l)} & \cdots & w_{ni}^{(l)} \\ | & | & & | \end{bmatrix}\begin{bmatrix} 1 \\ s_1^{(l)} z_1^{(l)} \\ s_2^{(l)} z_2^{(l)} \\ \vdots \\ s_n^{(l)} z_n^{(l)} \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} | & | & & | \\ b_i^{(l)} & s_1^{(l)} w_{1i}^{(l)} & \cdots & s_n^{(l)} w_{ni}^{(l)} \\ | & | & & | \end{bmatrix}}_{:= \hat{W}^{(l)}} \hat{z}^{(l)} = \hat{W}^{(l)}\hat{z}^{(l)}$$

حال می‌توان به‌طور بازگشتی نتیجه گرفت که :

$$\hat{z}^{(l+1)} = \hat{W}^{(l)}\hat{z}^{(l)} = \hat{W}^{(l)}\hat{W}^{(l-1)}\hat{z}^{(l-2)} = \cdots = \left(\prod_{i=l}^{1} W^{(i)}\right)\hat{z}^{(1)}, \quad \hat{z}^{(1)} = \hat{W}^{(0)}\hat{x}$$

$$\longrightarrow \hat{z}^{(l+1)} = \left(\prod_{i=l}^{0} \hat{W}^{(i)}\right)\hat{x}$$

از طرفی می‌دانیم که رابطه‌ی یک به یک زیر برای ورودی نورون‌ها و حالت‌ها داده شده وجود دارد :

$$(I) \longrightarrow \begin{cases} z_i^{(l)} \geqslant 0 \longleftrightarrow s_i^{(l)} = 1 \\ z_i^{(l)} < 0 \longleftrightarrow s_i^{(l)} = 0 \end{cases}$$

بنابراین برای نورون i ام از لایه‌ی l ام رابطه‌ی زیر برقرار است :

$$\left(2s_i^{(l)} - 1\right) z_i^{(l)} = \begin{cases} z_i^{(l)} & \text{if } z_i^{(l)} \geqslant 0 \\ -z_i^{(l)} & \text{if } z_i^{(l)} < 0 \end{cases} \geqslant 0 \longrightarrow \left(2s^{(l)} - 1_n\right) \odot z^{(l)} \succeq 0$$

$$\longrightarrow \left(2s^{(l+1)} - 1_n\right) \odot z^{(l+1)} \succeq 0 \longrightarrow \left(2s^{(l+1)} - 1_n\right) \odot \left(\prod_{i=l}^{0} \hat{W}^{(i)}\right)\hat{x} \succeq 0$$

این عبارت نشان دهنده‌ی یک convex polyhedron می‌باشد.

Definition of a convex polyhedron: $P = \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n \mid a_i^T x \leqslant b_i\}$ (A polyhedron is also a polytope)

ب) دو حالت داریم :

I) $k < m$: $r(k,m) \leqslant \sum_{i=0}^{m} \binom{k}{i} = \sum_{i=0}^{k} \binom{k}{i} = 2^k \leqslant k^k < k^m \longrightarrow r(k,m)_I \leqslant O(k^m)$

II) $k \geqslant m$: $\quad r(k,m) \leqslant \sum_{i=0}^{m}\binom{k}{i} \leqslant \sum_{i=0}^{k}\binom{k}{i} \leqslant \sum_{i=0}^{k}\binom{k}{i}\left(\frac{k}{m}\right)^{m-i} = \sum_{i=0}^{k}\binom{k}{i}\left(\frac{k}{m}\right)^{m}\left(\frac{k}{m}\right)^{-i} = \left(\frac{k}{m}\right)^{m}\sum_{i=0}^{k}\binom{k}{i}\left(\frac{k}{m}\right)^{-i}$

$$= \left(\frac{k}{m}\right)^{m}\sum_{i=0}^{k}\binom{k}{i}\left(\frac{m}{k}\right)^{i} = \left(\frac{k}{m}\right)^{m}\left(1+\frac{m}{k}\right)^{k}$$

Taylor: $e^{-x} = 1-x+\frac{x^2}{2!}-\frac{x^3}{3!}+\dots \longrightarrow 1-x \leqslant e^{-x}$

$$\longrightarrow r(k,m) \leqslant \left(\frac{k}{m}\right)^{m}\left(1+\frac{m}{k}\right)^{k} \leqslant \left(\frac{k}{m}\right)^{m} e^{m/k} \leqslant (k/m)^m e^{m} = (e/m)^m k^m \leqslant O(k^m)$$

$$\longrightarrow r(k,m)_{\text{II}} \leqslant O(k^m) \xrightarrow{(\text{I}),(\text{II})} r(k,m) \leqslant O(k^m)$$

باتوجه به بخش الف می‌دانیم که هرلایه یک convex polytope تشکیل می‌دهد (با اضافه کردن $O(k^m)$ خط). بنابراین
با n لایه $O(k^{mn})$ خط اضافه می‌شود:

$$\underset{l=0}{1} \longrightarrow \underset{l=1}{O(k^m)} \longrightarrow \underset{l=2}{O(k^{2m})} \longrightarrow \underset{\cdots}{\cdots} \longrightarrow \underset{l=n}{O(k^{mn})}$$