علیرضا میرزکنی - ۴۰۱۱۰۶۶۱۷

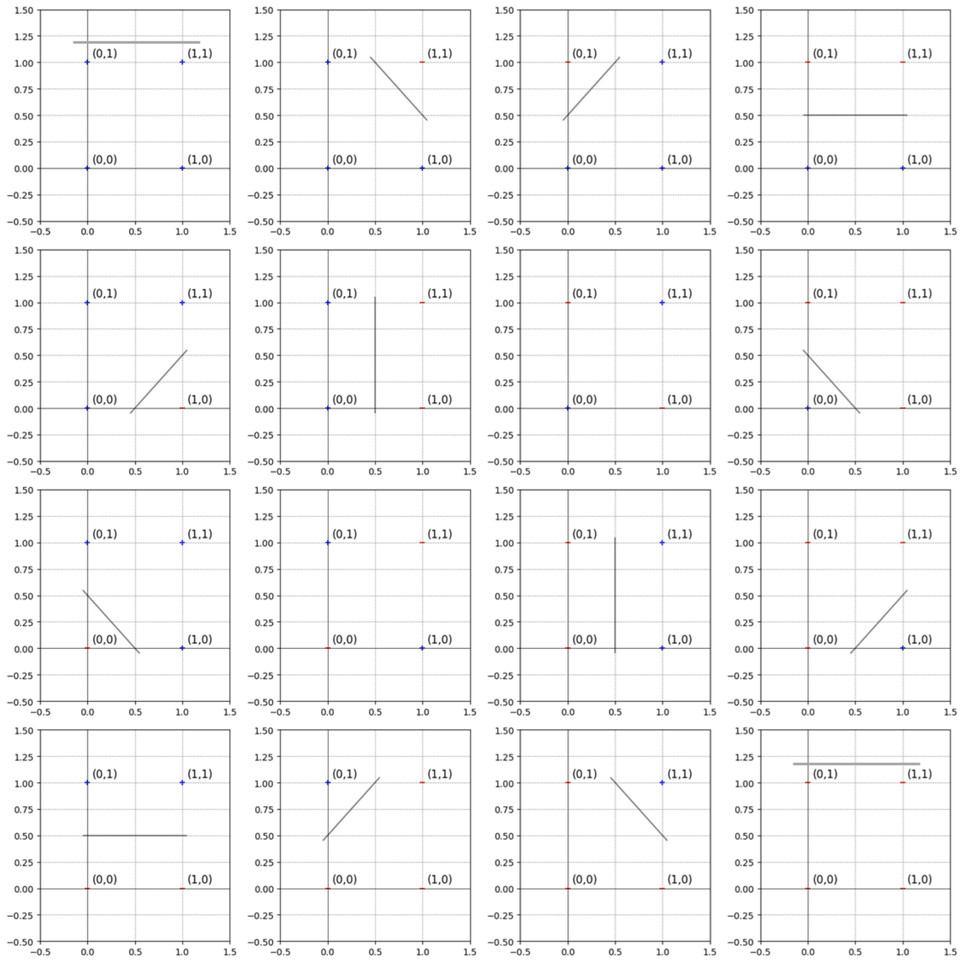
دانشکده‌ی مهندسی کامپیوتر

تمرین سوم

پاسخ مسئله‌ی ۱.

الف) **۱۴ تابع.**

در شکل زیر، تمامی ۱۶ حالت ممکن برای خروجی یک تابع boolean بر روی دو ورودی را مشاهده می کنید. نقاط موجود در صفحه، نشان دهنده ورودی های تابع هستند (به طور مثال نقطه (۰ , ۰) در صفحه نشان دهنده صفر بودن هر دو ورودی تابع است). همچنین اگر یک نقطه با علامت "+" مشخص شده باشد، یعنی خروجی تابع برای آن ورودی برابر یک بوده و اگر با علامت "-" مشخص شده باشد، یعنی خروجی تابع برای آن ورودی برابر صفر است.

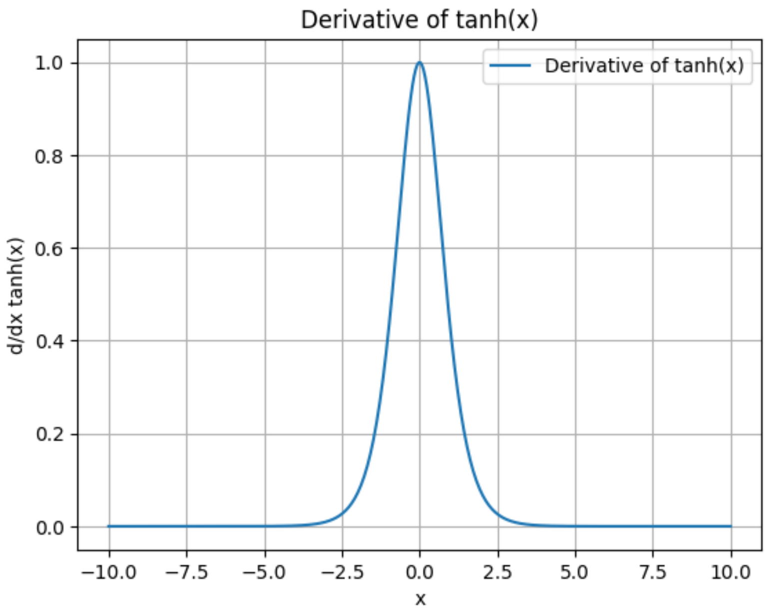


شکل ۱: حالات ممکن برای یک تابع boolean با ۲ ورودی

همانطور که مشاهده می شود، تنها در دو حالتی که کلاس های مختلف به شکل قطری قرار گرفته باشند مجموعه داده حاصل خطی جدایی پذیر نخواهد بود. پس پاسخ مسئله ۱۴ می باشد.

ب) نادرست.

در شکل زیر نمودار $\frac{d(\tanh x)}{dx} = ۱ - \tanh^۲ x$ را مشاهده می کنید. به وضوح خروجی این مشتق همواره بین صفر و یک بوده و به خصوص برای x هایی که قدر مطلقشان از یک بزرگ تر است، بسیار نزدیک به صفر می باشد. بنابراین در صورت استفاده از این تابع فعال سازی در ساختار یک شبکه عصبی، هنگام ضرب شدن مشتقات جزئی در فرآیند پس‌انتشار، اندازه گرادیان بسیار کوچک می شود و ناپدید شدن گرادیان رخ می دهد.



شکل ۲: نمودار $\frac{d(\tanh x)}{dx}$

ج) درست.

براساس قضیه تقریب‌پذیری (Universal Approximation Theorem)، شبکه‌های عصبی با یک لایه مخفی، تعداد کافی نورون و یک تابع فعال سازی غیر خطی (در اینجا RELU) قادر به تقریب‌زنی هر تابع پیوسته‌ای بر روی مجموعه‌های محدود هستند. اگرچه توابع منطقی در اصل گسسته و غیرپیوسته هستند، اما می‌توان آن‌ها را به صورت ترکیبی از توابع پیوسته مدل‌سازی کرد. در واقع هر تابع منطقی را می توان به شکل SOP (تعدادی عبارت or به شکل and تعدادی متغیر دودویی) نوشت، از طرفی می دانیم که برای متغیر های دودویی $x_۱, x_۲$ داریم:

$$x_۱ \wedge x_۲ = x_۱ x_۲$$

$$x_۱ \vee x_۲ = \frac{|x_۱ + x_۲| + |x_۱ - x_۲|}{۲}$$

که عبارات بالا به راحتی قابل تعمیم برای n متغیر دودویی می باشند.

د) نادرست.

عرض و عمق، دو مورد از هایپرپارامتر های یک شبکه عصبی هستند و باید مقدار مناسب آن ها را پیدا کرد (به عبارت دیگر باید tune شوند). افزایش بیش از اندازه این هایپرپارامتر ها، پیچیدگی شبکه عصبی را زیاد می کند و می تواند باعث بیش برازش، افزایش خطای تعمیم (generalization error) و در نتیجه بدتر شدن عملکرد شبکه عصبی شود.

ه) **نادرست.**

شبکه‌های عصبی عمیق دارای چندین لایه مخفی هستند که به آن‌ها امکان می‌دهد ویژگی‌های پیچیده‌تر و انتزاعی‌تری را از داده‌ها استخراج کنند. این ویژگی باعث می‌شود که شبکه‌های عمیق بتوانند توابع پیچیده را با تعداد کمتری نورون نسبت به شبکه‌های سطحی مدل‌سازی کنند.

اگرچه شبکه‌های عصبی سطحی و بسیار پهن قادر به تقریب‌زدن خروجی شبکه‌های عمیق و به دست آوردن عملکرد تقریبا مشابه هستند (با توجه به Universal Approximation Theorem)، اما این امر به معنای کارآمد بودن آنها نیست. در بسیاری از موارد، شبکه‌های سطحی برای مدل‌سازی توابع پیچیده نیاز به تعداد بسیار زیادی نورون دارند که این موضوع از نظر محاسباتی و زمانی غیرعملی است. در مقابل، شبکه‌های عمیق می‌توانند همان توابع را با تعداد کمتری نورون و لایه به طور کارآمدتری مدل‌سازی کنند.

و)

ناپدید شدن گرادیان‌ها زمانی رخ می‌دهد که گرادیان‌های محاسبه شده در فرآیند پس‌انتشار به تدریج کاهش یافته و به صفر نزدیک می‌شوند. این اتفاق باعث می‌شود که وزن‌های لایه‌های اولیه شبکه به‌طور بسیار کند به‌روزرسانی شوند و یا اصلا به‌روزرسانی نشوند و در نتیجه یادگیری شبکه متوقف شود.

توابع فعال‌سازی مانند ReLU که دارای مشتق با اندازه بزرگ‌تر یا مساوی یک در ناحیه مثبت هستند، از کاهش گرادیان‌ها در لایه‌های عمیق جلوگیری می‌کنند (چرا که دیگر گرادیان‌ها در حین فرآیند پس انتشار، مکررا در مشتق های جزئی با اندازه کوچک تر از یک ضرب نمی‌شوند و اندازه‌شان بسیار کوچک نمی‌شود). این ویژگی اجازه می‌دهد که گرادیان‌های قوی‌تر به طور مؤثرتری از لایه‌های بالاتر به پایین‌تر منتقل شوند. در نتیجه، شبکه‌های عصبی با استفاده از ReLU سریع‌تر و کارآمدتر آموزش می‌بینند و مشکل ناپدید شدن گرادیان‌ها کاهش می‌یابد.

ز) **نادرست.**

هنگام استفاده از SGD، گام هایی که برداشته می‌شوند بسیار سریع هستند (فرکانس بروزرسانی زیاد است) اما به دلیل استفاده از تنها یک نمونه آموزشی برای بروزرسانی وزن‌ها، گام‌های برداشته شده دقت بالایی ندارند و حتی ممکن است پارامتر های شبکه را از پارامتر های بهینه دورتر کنند.

در مقابل، Mini-Batch GD دارای فرکانس بروزرسانی کمتری است (گام‌ها کندتر برداشته می‌شوند) اما چون از تعداد بیشتری نمونه آموزشی برای بروزرسانی وزن‌ها استفاده می کند، تابع هزینه حاصل، تقریب دقیق تری از تابع هزینه برای کل نمونه ها می باشد و در نتیجه گام برداشته شده دقیق تر خواهد بود.

بنابراین می توان نتیجه گرفت که معمولا SGD نسبت به Mini-Batch GD، نیاز به تعداد بروزرسانی های بیشتری برای همگرایی دارد و در نتیجه حتی ممکن است دیرتر همگرا شود.

ی) مراحل بروزرسانی در ادامه به ترتیب آورده شده اند.

۱. **محاسبه گرادیان تابع:** گرادیان تابع نسبت به x محاسبه می‌شود:

$$\frac{dy}{dx} = ۱/۲x^۳ - ۰/۳x^۲ - ۴x - ۰/۸$$

۲. **الگوریتم SGD با ممان:** فرمول به‌روزرسانی مقدار x با استفاده از ممان به صورت زیر است:

$$m_i = \mu m_{i-۱} + (۱-\mu) g_{i-۱} \quad , \quad x_i = x_{i-۱} - \gamma m_i$$

۳. محاسبات مرحله‌به‌مرحله:

مرحله ۱: محاسبه $y_{۰}$ **و** $g_{۰}$ **در** $x_{۰}$

$$y_{۰} = ۰/۳(-۲/۸)^{۴} - ۰/۱(-۲/۸)^{۳} - ۲(-۲/۸)^{۲} - ۰/۸(-۲/۸) = ۷/۱۹۴۹$$

$$g_{۰} = ۱/۲(-۲/۸)^{۳} - ۰/۳(-۲/۸)^{۲} - ۴(-۲/۸) - ۰/۸ = -۱۸/۲۹۴۴$$

مرحله ۲: محاسبه $m_{۱}$ **و** $x_{۱}$

فرض می‌کنیم $m_{۰} = ۰$:

$$m_{۱} = \mu m_{۰} + (۱ - \mu) g_{۰} = ۰/۷(۰) + ۰/۳(-۱۸/۲۹۴۴) = -۵/۴۸۸۳$$

$$x_{۱} = x_{۰} - \gamma m_{۱} = -۲/۸ - ۰/۰۵(-۵/۴۸۸۳) = -۲/۵۲۵۶$$

مرحله ۳: محاسبه $y_{۱}$ **و** $g_{۱}$ **در** $x_{۱}$

$$y_{۱} = ۰/۳(-۲/۵۲۵۶)^{۴} - ۰/۱(-۲/۵۲۵۶)^{۳} - ۲(-۲/۵۲۵۶)^{۲} - ۰/۸(-۲/۵۲۵۶) = ۳/۰۸۰۳$$

$$g_{۱} = ۱/۲(-۲/۵۲۵۶)^{۳} - ۰/۳(-۲/۵۲۵۶)^{۲} - ۴(-۲/۵۲۵۶) - ۰/۸ = -۱۱/۹۴۲۸$$

مرحله ۴: محاسبه $m_{۲}$ **و** $x_{۲}$

$$m_{۲} = \mu m_{۱} + (۱ - \mu) g_{۱} = ۰/۷(-۵/۴۸۸۳) + ۰/۳(-۱۱/۹۴۲۸) = -۷/۴۲۴۶$$

$$x_{۲} = x_{۱} - \gamma m_{۲} = -۲/۵۲۵۶ - ۰/۰۵(-۷/۴۲۴۶) = -۲/۱۵۴۳$$

مرحله ۵: محاسبه $y_{۲}$ **در** $x_{۲}$

$$y_{۲} = ۰/۳(-۲/۱۵۴۳)^{۴} - ۰/۱(-۲/۱۵۴۳)^{۳} - ۲(-۲/۱۵۴۳)^{۲} - ۰/۸(-۲/۱۵۴۳) = -۰/۰۹۷۱$$

نتایج نهایی

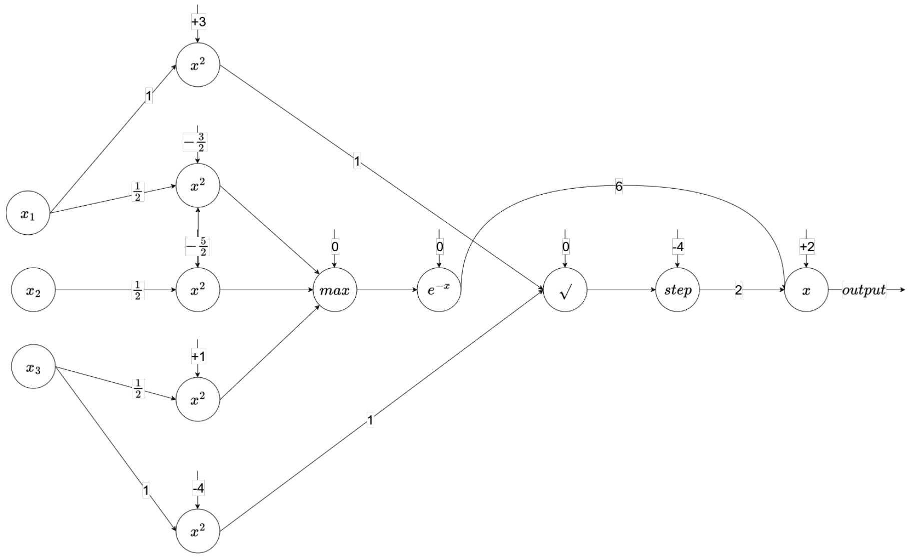
$$x_{۰} = -۲/۸ \quad , \quad m_{۰} = ۰ \quad , \quad g_{۰} = -۱۸/۲۹۴۴ \quad , \quad y_{۰} = ۷/۱۹۴۹$$

$$x_{۱} = -۲/۵۲۵۶ \quad , \quad m_{۱} = -۵/۴۸۸۳ \quad , \quad g_{۱} = -۱۱/۹۴۲۸ \quad , \quad y_{۱} = ۳/۰۸۰۳$$

$$x_{۲} = -۲/۱۵۴۳ \quad , \quad m_{۲} = -۷/۴۲۴۶ \quad , \quad y_{۲} = -۰/۰۹۷۱$$

پاسخ مسئله‌ی ۲.

الف) در شکل زیر یک شبکه عصبی مطلوب را مشاهده می کنید.



شکل ۳: یک شبکه عصبی مطلوب

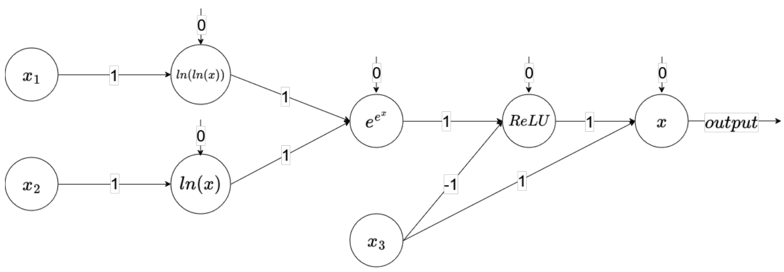
دقت کنید که تابع فعال سازی هر نورون داخل آن و بایاس هر نورون در بالای آن نوشته شده است. همچنین در نورون هایی که یال های ورودی وزن دار دارند، تابع ورودی جمع وزن دار ورودی ها است؛ اما در نورون هایی که یال های ورودیشان بدون وزن است، تابع ورودی همانی است (به عبارت دیگر همه ورودی ها مستقیما به تابع فعال سازی وارد می شوند). همچنین توجه نمایید که می توان بجای نورون با تابع فعال سازی max، از دو نورون با تابع فعال سازی ReLU استفاده کرد که $\max(a, b, c) = \mathrm{ReLU}(\mathrm{ReLU}(a - b) + b - c) + c$ را خروجی می دهند.

ب) دقت کنید که می توان نوشت: (فرض می کنیم که ورودی های $x_۱$ و $x_۲$ مثبت اند، چرا که در غیر این صورت تابع $x_۱^{x_۲}$ ممکن است تعریف نشود. به علاوه، بدون این فرض ساختن آن با پیچیدگی بسیار زیادی همراه خواهد بود)

$$x_۱^{x_۲} = e^{x_۲ \ln x_۱} = e^{e^{ln x_۲ + \ln \ln x_۱}}$$

$$\max(a, b) = \mathrm{ReLU}(a - b) + b$$

با توجه به تساوی های بالا، یک شبکه عصبی مطلوب به شکل زیر می باشد. مشابه قسمت قبل، تابع فعال سازی و بایاس هر نورون به ترتیب درون و بالای آن نوشته شده اند.



شکل ۴: یک شبکه عصبی مطلوب

پاسخ مسئله‌ی ۳.

۱. ابعاد پارامتر های ذکر شده به شکل زیر است: (دقت کنید که چون $z_۲$ به عنوان ورودی به تابع $\sigma(x)$ داده شده است، باید یک عدد اسکالر باشد)

$$\begin{aligned} W_۱ &\to D_{a_۱} \times D_x \\ b_۱ &\to D_{a_۱} \times ۱ \\ W_۲ &\to ۱ \times D_{a_۱} \\ b_۲ &\to ۱ \times ۱ \end{aligned}$$

حال فرض کنید بخواهیم شبکه را بر روی k نمونه vectorize کنیم. در این صورت ابعاد پارامتر های بالا به شکل زیر تغییر می کنند:

$$\begin{aligned} W_۱ &\to D_{a_۱} \times D_x \\ b_۱ &\to D_{a_۱} \times k \\ W_۲ &\to k \times D_{a_۱} \\ b_۲ &\to k \times ۱ \end{aligned}$$

همچنین در این حالت ابعاد X و Y به شکل زیر خواهند بود:

$$\begin{aligned} X &\to D_x \times k \\ Y &\to k \times ۱ \end{aligned}$$

.۲

$$\begin{aligned} \frac{\partial J}{\partial \hat{y}^{(i)}} &= \frac{\partial J}{\partial L^{(i)}} \times \frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} \\ &= \frac{\partial \left(-\frac{۱}{m}\sum_{j=۱}^{m} L^{(j)}\right)}{\partial L^{(i)}} \times \frac{\partial \left(y^{(i)} \log \hat{y}^{(i)} + \left(۱ - y^{(i)}\right)\left(۱ - \log \hat{y}^{(i)}\right)\right)}{\partial \hat{y}^{(i)}} \\ &= \left(-\frac{۱}{m}\right) \times \left(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{۱ - y^{(i)}}{۱ - \hat{y}^{(i)}}\right) \\ &= \frac{\hat{y}^{(i)} - y^{(i)}}{m\hat{y}^{(i)}\left(۱ - \hat{y}^{(i)}\right)} \\ &\Longrightarrow \frac{\partial J}{\partial \hat{y}} = \frac{\hat{y} - y}{m\hat{y} \odot (۱ - \hat{y})} \end{aligned}$$

به طوری که در عبارت بالا $\odot$ نشان دهنده ضرب elementwise بوده و تقسیم نیز به صورت elementwise انجام گرفته است.

.۳

$$\begin{aligned} \frac{\partial \hat{y}^{(i)}}{\partial z_۲} &= \frac{\partial \sigma(z_۲)}{\partial z_۲} \\ &= \sigma(z_۲)\left(۱ - \sigma(z_۲)\right) \\ &= \hat{y}^{(i)}\left(۱ - \hat{y}^{(i)}\right) \end{aligned}$$

.۴

$$\frac{\partial z_۲}{\partial a_۱} = \frac{\partial \left(W_۲ a_۱ + b_۲\right)}{\partial a_۱}$$
$$= W_۲$$

.۵

$$\left(\frac{\partial a_۱}{\partial z_۱}\right)_i = \left(\frac{\partial \,\mathrm{ReLU}(z_۱)}{\partial z_۱}\right)_i$$
$$= \begin{cases} ۰ & z_{۱_i} \leqslant ۰ \\ ۱ & z_{۱_i} > ۰ \end{cases}$$

.۶

$$\frac{\partial z_۱}{\partial W_۱} = \frac{\partial \left(W_۱ x^{(i)} + b_۱\right)}{\partial W_۱}$$
$$= x^{(i)^T}$$

.۷

$$\frac{\partial J}{\partial W_۱} = -\frac{۱}{m}\sum_{i=۱}^{m} \frac{\partial J}{\partial \hat{y}^{(i)}} \cdot \frac{\partial \hat{y}^{(i)}}{\partial z_۲} \cdot \left(\frac{\partial z_۲}{\partial a_۱} \odot \frac{\partial a_۱}{\partial z_۱}\right) \cdot \frac{\partial z_۱}{\partial W_۱}$$
$$= -\frac{۱}{m}\sum_{i=۱}^{m} \delta_۱^{(i)} \cdot \delta_۲^{(i)} \cdot \left(\delta_۳^{(i)} \odot \delta_۴^{(i)}\right) \cdot \delta_۵^{(i)}$$

پاسخ مسئله‌ی ۴.

الف) مقادیر خروجی نورون ها در ادامه آورده شده اند.

لایه مخفی اول:

$$h_1^{(1)} = \mathrm{Sigmoid}(a) = \frac{1}{1+e^{-a}}$$
$$h_2^{(1)} = \mathrm{Sigmoid}(-a) = \frac{1}{1+e^{a}}$$

لایه مخفی دوم:

$$h_1^{(2)} = \mathrm{ReLU}(ah_1^{(1)} - bh_2^{(1)}) = \mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)$$
$$h_2^{(2)} = \mathrm{ReLU}(bh_1^{(1)} - ah_2^{(1)}) = \mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)$$

لایه مخفی سوم:

$$h_1^{(3)} = \mathrm{Softmax}(ah_1^{(2)} - ah_2^{(2)}, bh_2^{(2)} - bh_1^{(2)})_1 = \frac{e^{ah_1^{(2)} - ah_2^{(2)}}}{e^{ah_1^{(2)} - ah_2^{(2)}} + e^{bh_2^{(2)} - bh_1^{(2)}}}$$
$$\Longrightarrow h_1^{(3)} = \frac{e^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)}}{e^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)} + e^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)}}$$
$$h_2^{(3)} = \mathrm{Softmax}(ah_1^{(2)} - ah_2^{(2)}, bh_2^{(2)} - bh_1^{(2)})_2 = \frac{e^{bh_2^{(2)} - bh_1^{(2)}}}{e^{ah_1^{(2)} - ah_2^{(2)}} + e^{bh_2^{(2)} - bh_1^{(2)}}}$$
$$\Longrightarrow h_2^{(3)} = \frac{e^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)}}{e^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)} + e^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)}}$$

لایه خروجی:

$$a_1^{(4)} = ah_1^{(3)} + 0/5ah_2^{(3)}$$
$$\Longrightarrow a_1^{(4)} = \frac{ae^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)} + \frac{a}{2}e^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)}}{e^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)} + e^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)}}$$
$$a_2^{(4)} = ah_2^{(3)} - bh_1^{(3)}$$
$$\Longrightarrow a_2^{(4)} = \frac{ae^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)} - be^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)}}{e^{a\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right) - a\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right)} + e^{b\,\mathrm{ReLU}\left(\frac{b}{1+e^{-a}} - \frac{a}{1+e^{a}}\right) - b\,\mathrm{ReLU}\left(\frac{a}{1+e^{-a}} - \frac{b}{1+e^{a}}\right)}}$$

ب)

$$E = \frac{۱}{۲}\left(\left(a_۱^{(۴)} - t_۱\right)^۲ + \left(a_۲^{(۴)} - t_۲\right)^۲\right) = \frac{۱}{۲}\left(a_۱^{(۴)^۲} + \left(a_۲^{(۴)} - ۱\right)^۲\right)$$

با جایگذاری مقادیر $a_۱^{(۴)}$ و $a_۲^{(۴)}$ (به دست آمده در قسمت الف) در عبارت بالا، مقدار دقیق E به دست می آید.

ج) گرادیان E را نسبت به تمامی پارامتر ها، از لایه آخر به لایه اول به ترتیب محاسبه می کنیم.

لایه خروجی:

$$\frac{\partial E}{\partial a_۱^{(۴)}} = a_۱^{(۴)}$$

$$\frac{\partial E}{\partial a_۲^{(۴)}} = a_۲^{(۴)} - ۱$$

$$\frac{\partial E}{\partial W_{۱۱}^{(۴)}} = \frac{\partial E}{\partial a_۱^{(۴)}} \cdot \frac{\partial a_۱^{(۴)}}{\partial W_{۱۱}^{(۴)}} = a_۱^{(۴)} h_۱^{(۳)}$$

$$\frac{\partial E}{\partial W_{۲۱}^{(۴)}} = \frac{\partial E}{\partial a_۱^{(۴)}} \cdot \frac{\partial a_۱^{(۴)}}{\partial W_{۲۱}^{(۴)}} = a_۱^{(۴)} h_۲^{(۳)}$$

$$\frac{\partial E}{\partial W_{۱۲}^{(۴)}} = \frac{\partial E}{\partial a_۲^{(۴)}} \cdot \frac{\partial a_۲^{(۴)}}{\partial W_{۱۲}^{(۴)}} = \left(a_۲^{(۴)} - ۱\right) h_۱^{(۳)}$$

$$\frac{\partial E}{\partial W_{۲۲}^{(۴)}} = \frac{\partial E}{\partial a_۲^{(۴)}} \cdot \frac{\partial a_۲^{(۴)}}{\partial W_{۲۲}^{(۴)}} = \left(a_۲^{(۴)} - ۱\right) h_۲^{(۳)}$$

$$\frac{\partial E}{b_۱^{(۴)}} = \frac{\partial E}{\partial a_۱^{(۴)}} \cdot \frac{\partial a_۱^{(۴)}}{\partial b_۱^{(۴)}} = a_۱^{(۴)}$$

$$\frac{\partial E}{b_۲^{(۴)}} = \frac{\partial E}{\partial a_۲^{(۴)}} \cdot \frac{\partial a_۲^{(۴)}}{\partial b_۲^{(۴)}} = a_۲^{(۴)} - ۱$$

لایه مخفی سوم:

$$\frac{\partial E}{\partial h_۱^{(۳)}} = \frac{\partial E}{\partial a_۱^{(۴)}} \cdot \frac{\partial a_۱^{(۴)}}{\partial h_۱^{(۳)}} + \frac{\partial E}{\partial a_۲^{(۴)}} \cdot \frac{\partial a_۲^{(۴)}}{\partial h_۱^{(۳)}} = a_۱^{(۴)} a - \left(a_۲^{(۴)} - ۱\right) b$$

$$\frac{\partial E}{\partial h_۲^{(۳)}} = \frac{\partial E}{\partial a_۱^{(۴)}} \cdot \frac{\partial a_۱^{(۴)}}{\partial h_۲^{(۳)}} + \frac{\partial E}{\partial a_۲^{(۴)}} \cdot \frac{\partial a_۲^{(۴)}}{\partial h_۲^{(۳)}} = \frac{۱}{۲} a_۱^{(۴)} a + \left(a_۲^{(۴)} - ۱\right) a$$

$$\begin{aligned}\frac{\partial E}{\partial W_{۱۱}^{(۳)}} = \frac{\partial E}{\partial h_۱^{(۳)}} \cdot \frac{\partial h_۱^{(۳)}}{\partial W_{۱۱}^{(۳)}} + \frac{\partial E}{\partial h_۲^{(۳)}} \cdot \frac{\partial h_۲^{(۳)}}{\partial W_{۱۱}^{(۳)}} &= \left(a_۱^{(۴)} a - \left(a_۲^{(۴)} - ۱\right) b\right) h_۱^{(۳)} (۱ - h_۱^{(۳)}) h_۱^{(۲)} \\ &- \left(\frac{۱}{۲} a_۱^{(۴)} a + \left(a_۲^{(۴)} - ۱\right) a\right) h_۱^{(۳)} h_۲^{(۳)} h_۱^{(۲)}\end{aligned}$$

$$\frac{\partial E}{\partial W_{21}^{(3)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial W_{21}^{(3)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial W_{21}^{(3)}} = \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)}(1 - h_1^{(3)}) h_2^{(2)} \\ - \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_1^{(3)} h_2^{(3)} h_2^{(2)}$$

$$\frac{\partial E}{\partial W_{12}^{(3)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial W_{12}^{(3)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial W_{12}^{(3)}} = \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_2^{(3)}(1 - h_2^{(3)}) h_1^{(2)} \\ - \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)} h_2^{(3)} h_1^{(2)}$$

$$\frac{\partial E}{\partial W_{22}^{(3)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial W_{22}^{(3)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial W_{22}^{(3)}} = \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_2^{(3)}(1 - h_2^{(3)}) h_2^{(2)} \\ - \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)} h_2^{(3)} h_2^{(2)}$$

$$\frac{\partial E}{\partial b_1^{(3)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial b_1^{(3)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial b_1^{(3)}} = \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)}(1 - h_1^{(3)}) \\ - \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_1^{(3)} h_2^{(3)}$$

$$\frac{\partial E}{\partial b_2^{(3)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial b_2^{(3)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial b_2^{(3)}} = \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_2^{(3)}(1 - h_2^{(3)}) \\ - \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)} h_2^{(3)}$$

لایه مخفی دوم:

$$\frac{\partial E}{\partial h_1^{(2)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial h_1^{(2)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial h_1^{(2)}} = \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)}(1 - h_1^{(3)}) a \\ + \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)} h_2^{(3)} b - \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_1^{(3)} h_2^{(3)} a \\ - \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_2^{(3)}(1 - h_2^{(3)}) b$$

$$\frac{\partial E}{\partial h_2^{(2)}} = \frac{\partial E}{\partial h_1^{(3)}} \cdot \frac{\partial h_1^{(3)}}{\partial h_2^{(2)}} + \frac{\partial E}{\partial h_2^{(3)}} \cdot \frac{\partial h_2^{(3)}}{\partial h_2^{(2)}} = -\left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)}(1 - h_1^{(3)}) a \\ - \left(a_1^{(4)} a - \left(a_2^{(4)} - 1\right) b\right) h_1^{(3)} h_2^{(3)} b + \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_1^{(3)} h_2^{(3)} a \\ + \left(\frac{1}{2} a_1^{(4)} a + \left(a_2^{(4)} - 1\right) a\right) h_2^{(3)}(1 - h_2^{(3)}) b$$

$$\frac{\partial E}{\partial W_{11}^{(2)}} = \frac{\partial E}{\partial h_1^{(2)}} \cdot \frac{\partial h_1^{(2)}}{\partial W_{11}^{(2)}} = \frac{\partial E}{\partial h_1^{(2)}} \cdot \mathrm{ReLU}'(a h_1^{(1)} - b h_2^{(1)}) h_1^{(1)}$$

$$\frac{\partial E}{\partial W_{21}^{(2)}} = \frac{\partial E}{\partial h_1^{(2)}} \cdot \frac{\partial h_1^{(2)}}{\partial W_{21}^{(2)}} = \frac{\partial E}{\partial h_1^{(2)}} \cdot \mathrm{ReLU}'(a h_1^{(1)} - b h_2^{(1)}) h_2^{(1)}$$

$$\frac{\partial E}{\partial W^{(۲)}_{۱۲}} = \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \frac{\partial h^{(۲)}_{۲}}{\partial W^{(۲)}_{۱۲}} = \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \mathrm{ReLU}'(bh^{(۱)}_{۱} - ah^{(۱)}_{۲})h^{(۱)}_{۱}$$

$$\frac{\partial E}{\partial W^{(۲)}_{۲۲}} = \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \frac{\partial h^{(۲)}_{۲}}{\partial W^{(۲)}_{۲۲}} = \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \mathrm{ReLU}'(bh^{(۱)}_{۱} - ah^{(۱)}_{۲})h^{(۱)}_{۲}$$

$$\frac{\partial E}{\partial b^{(۲)}_{۱}} = \frac{\partial E}{\partial h^{(۲)}_{۱}} \cdot \frac{\partial h^{(۲)}_{۱}}{\partial b^{(۲)}_{۱}} = \frac{\partial E}{\partial h^{(۲)}_{۱}} \cdot \mathrm{ReLU}'(ah^{(۱)}_{۱} - bh^{(۱)}_{۲})$$

$$\frac{\partial E}{\partial b^{(۲)}_{۲}} = \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \frac{\partial h^{(۲)}_{۲}}{\partial b^{(۲)}_{۲}} = \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \mathrm{ReLU}'(bh^{(۱)}_{۱} - ah^{(۱)}_{۲})$$

لایه مخفی اول

$$\begin{aligned}\frac{\partial E}{\partial h^{(۱)}_{۱}} = \frac{\partial E}{\partial h^{(۲)}_{۱}} \cdot \frac{\partial h^{(۲)}_{۱}}{\partial h^{(۱)}_{۱}} + \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \frac{\partial h^{(۲)}_{۲}}{\partial h^{(۱)}_{۱}} =& \frac{\partial E}{\partial h^{(۲)}_{۱}} \cdot \mathrm{ReLU}'(ah^{(۱)}_{۱} - bh^{(۱)}_{۲})a \\ &+ \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \mathrm{ReLU}'(bh^{(۱)}_{۱} - ah^{(۱)}_{۲})b\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial h^{(۱)}_{۲}} = \frac{\partial E}{\partial h^{(۲)}_{۱}} \cdot \frac{\partial h^{(۲)}_{۱}}{\partial h^{(۱)}_{۲}} + \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \frac{\partial h^{(۲)}_{۲}}{\partial h^{(۱)}_{۲}} =& - \frac{\partial E}{\partial h^{(۲)}_{۱}} \cdot \mathrm{ReLU}'(ah^{(۱)}_{۱} - bh^{(۱)}_{۲})b \\ &- \frac{\partial E}{\partial h^{(۲)}_{۲}} \cdot \mathrm{ReLU}'(bh^{(۱)}_{۱} - ah^{(۱)}_{۲})a\end{aligned}$$

$$\frac{\partial E}{W^{(۱)}_{۱۱}} = \frac{\partial E}{\partial h^{(۱)}_{۱}} \cdot \frac{\partial h^{(۱)}_{۱}}{\partial W^{(۱)}_{۱۱}} = \frac{\partial E}{\partial h^{(۱)}_{۱}} \cdot h^{(۱)}_{۱}(۱ - h^{(۱)}_{۱})$$

$$\frac{\partial E}{b^{(۱)}_{۱}} = \frac{\partial E}{\partial h^{(۱)}_{۱}} \cdot \frac{\partial h^{(۱)}_{۱}}{\partial b^{(۱)}_{۱}} = \frac{\partial E}{\partial h^{(۱)}_{۱}} \cdot h^{(۱)}_{۱}(۱ - h^{(۱)}_{۱})$$

$$\frac{\partial E}{W^{(۱)}_{۲۱}} = \frac{\partial E}{\partial h^{(۱)}_{۱}} \cdot \frac{\partial h^{(۱)}_{۱}}{\partial W^{(۱)}_{۲۱}} = ۰$$

$$\frac{\partial E}{W^{(۱)}_{۱۲}} = \frac{\partial E}{\partial h^{(۱)}_{۲}} \cdot \frac{\partial h^{(۱)}_{۲}}{\partial W^{(۱)}_{۱۲}} = \frac{\partial E}{\partial h^{(۱)}_{۲}} \cdot h^{(۱)}_{۲}(۱ - h^{(۱)}_{۲})$$

$$\frac{\partial E}{W^{(۱)}_{۲۲}} = \frac{\partial E}{\partial h^{(۱)}_{۲}} \cdot \frac{\partial h^{(۱)}_{۲}}{\partial W^{(۱)}_{۲۲}} = ۰$$

$$\frac{\partial E}{b^{(۱)}_{۲}} = \frac{\partial E}{\partial h^{(۱)}_{۲}} \cdot \frac{\partial h^{(۱)}_{۲}}{\partial b^{(۱)}_{۲}} = \frac{\partial E}{\partial h^{(۱)}_{۲}} \cdot h^{(۱)}_{۲}(۱ - h^{(۱)}_{۲})$$

به طوری که در عبارات بالا داریم:

$$\mathrm{ReLU}'(x) = \begin{cases} ۱ & x \geqslant ۰ \\ ۰ & x < ۰ \end{cases}$$

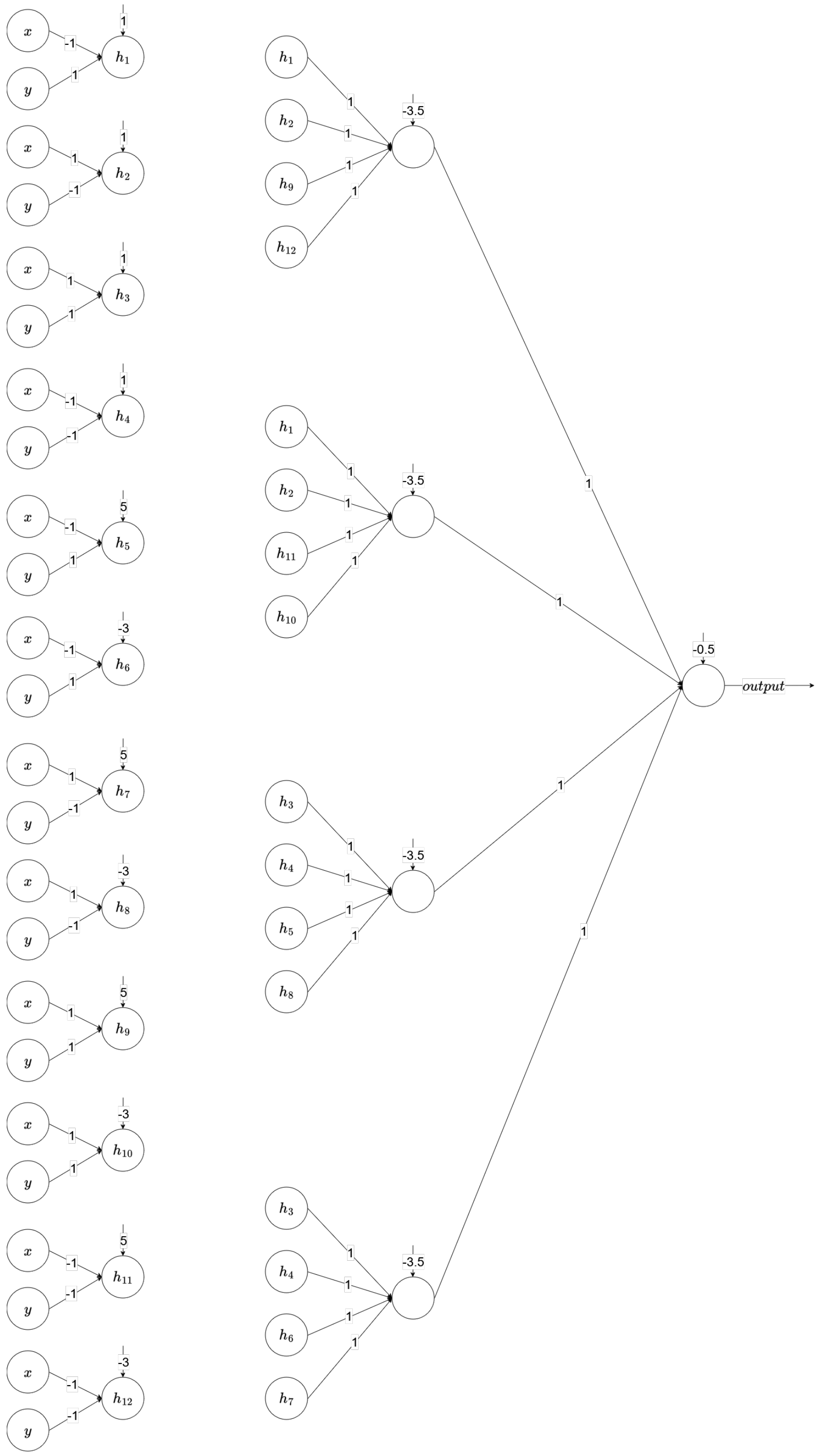
د) برای هر کدام از $W_{ij}^{(l)}$ ها و نیز هر کدام از $b_i^{(l)}$ ها ($i, j \in \{۱, ۲\} \ \wedge \ l \in \{۱, ۲, ۳, ۴\}$)، مقدار پارامتر را به شکل زیر بروزرسانی می کنیم:

$$param_{\text{new}} = param - \eta \frac{\partial E}{\partial param}$$

با استفاده از گرادیان های به دست آمده در قسمت قبل و همچنین مقادیر اولیه پارامتر ها (داده شده در صورت سوال)، می توان پارامتر های موجود در شبکه عصبی را بروزرسانی نمود.

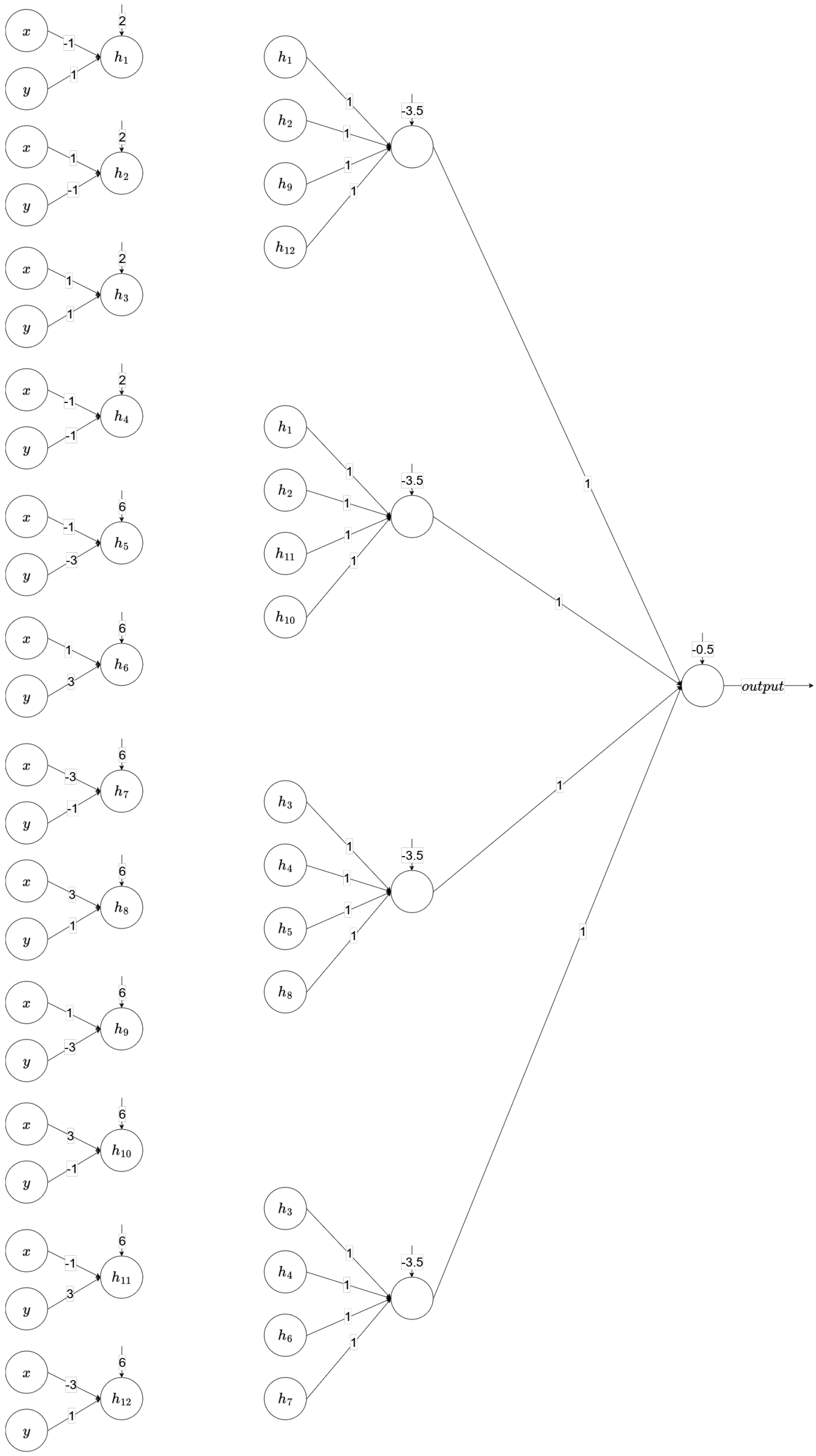
پاسخ مسئله‌ی ۵.

الف) شکل زیر یک شبکه مطلوب با ۱۷ عدد TLU را نمایش می دهد. برای جلوگیری از پیچیدگی زیاد شکل، ورودی ها و نیز برخی از نورون ها چند بار رسم شده اند.



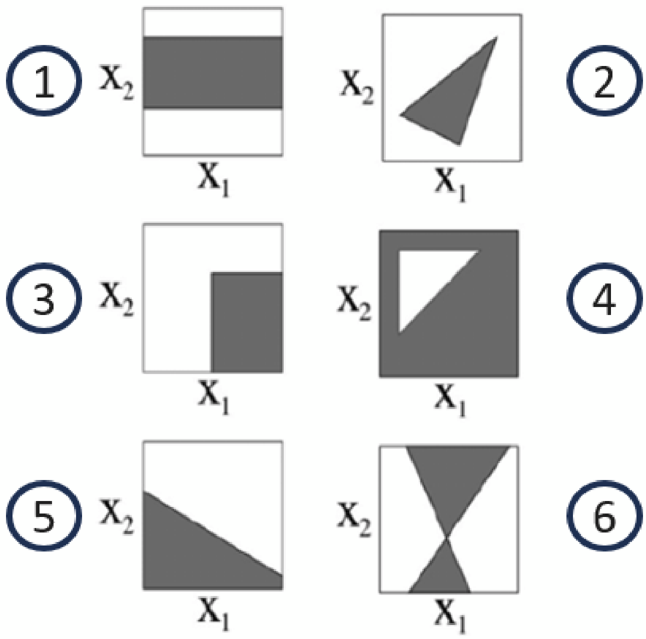
شکل ۵: معماری شبکه مطلوب

ب) بله. شبکه عصبی مطلوب را در شکل زیر مشاهده می کنید.



شکل ۶: معماری شبکه مطلوب

ج) شکل ها را به صورت زیر شماره گذاری می کنیم:



شکل ۷: شماره گذاری اشکال

شبکه A با شکل شمارهٔ ۵ مطابقت دارد، زیرا در این شبکه تنها می‌توان یک خط رسم کرد که صفحه را به وسیلهٔ آن به دو بخش مجزا تقسیم می‌کنیم.

شبکه B منطبق بر شکل شمارهٔ ۱ است؛ دلیل آن این است که در این شکل دو خط وجود دارد. در این شبکه نیز دو خط ایجاد می‌کنیم و سپس در لایهٔ دوم، خروجی‌ها را با استفاده از عملگر منطقی AND ترکیب می‌کنیم تا ناحیهٔ مشترک تعیین شود.

شبکه C با شکل شمارهٔ ۶ مرتبط است. در این حالت، ابتدا در لایهٔ اول دو خط ترسیم می‌کنیم، اما برای ترکیب خروجی‌ها نیاز به عملگر XOR داریم، که این امر مستلزم افزودن دو لایهٔ دیگر به شبکه است.

شبکه D به شکل شمارهٔ ۳ اختصاص دارد؛ زیرا می‌توانیم در لایهٔ اول یک خط افقی و یک خط عمودی رسم کنیم. سپس، در لایهٔ دوم، با بهره‌گیری از عملگر منطقی AND، ناحیهٔ بین این دو خط را مشخص می‌کنیم.

شبکه E متناظر با شکل شمارهٔ ۲ است. در این شبکه، نیاز است که در لایهٔ اول سه خط ترسیم کنیم و سپس در لایهٔ بعدی، خروجی‌ها را با هم توسط عملگر AND ترکیب کنیم تا ناحیهٔ مورد نظر حاصل شود.

شبکه F مرتبط با شکل شمارهٔ ۴ است؛ زیرا این شکل از یک خط افقی، یک خط عمودی و یک خط مورب تشکیل شده است. در این شبکه، یکی از نورون‌ها فقط ورودی x_1 را دریافت می‌کند، دیگری تنها ورودی x_2، و دیگری هر دو ورودی را دریافت می‌کند. سپس در لایهٔ بعدی، خروجی‌ها را با عملگر منطقی OR ترکیب می‌کنیم تا ناحیهٔ مطلوب تعیین شود.

پاسخ مسئله‌ی ۶.

الف) در ابتدا دقت کنید که لایه l ام شبکه عصبی را با $\mathcal{L}_l$ نمایش می دهیم، به طوری که $\mathcal{L}_۰$ لایه ورودی و $\mathcal{L}_{L+۱}$ لایه خروجی بوده و برای هر $l \in \{۱, \cdots, L\}$، $\mathcal{L}_l$ همان l امین لایه مخفی شبکه عصبی می باشد. برای i امین نورون مخفی در $\mathcal{L}_l$، $\sigma\left(z_i^{(l)}\right)$ به شکل زیر است:

$$\sigma\left(z_i^{(l)}\right) = \begin{cases} z_i^{(l)} & z_i^{(l)} \geqslant ۰ \\ ۰ & z_i^{(l)} < ۰ \end{cases}$$

و درنتیجه $\sigma\left(z_i^{(l)}\right)$ از دو بخش خطی تشکیل شده است. حال توجه نمایید که برای یک شبکه عصبی داده شده، یک نمونه $x \in \mathcal{X}$ مقدار $z_i^{(l)}$ و در نتیجه مقدار $\sigma\left(z_i^{(l)}\right)$ را به طور یکتا تعیین می کند. به عبارت دیگر همانطور که در صورت سوال هم اشاره شده است، می توان وضعیت نورون مخفی i ام در $\mathcal{L}_l$ را با یک متغیر دودویی مانند $c_i^{(l)} \in \{۰, ۱\}$ تعیین کرد، به طوری که $c_i^{(l)} = ۱$ اگر و تنها اگر $z_i^{(l)} \geqslant ۰$ و $c_i^{(l)} = ۰$ اگر و تنها اگر $z_i^{(l)} < ۰$.

فرض کنید $c^{(l)} = \left[c_۱^{(l)}, \cdots, c_{n_l}^{(l)}\right]$ ماتریس وضعیت تمامی نورون های مخفی در $\mathcal{L}_l$ بوده و $C = \left[c^{(۱)}, \cdots, c^{(L)}\right]$ نشان دهنده وضعیت تمامی نورون های مخفی در شبکه عصبی باشد. برای یک شبکه عصبی داده شده، ماتریس C با داشتن نمونه x به طور یکتا تعیین می شود.

در ادامه، تابعی که نمونه x را به ماتریس $C \in \{۰, ۱\}^N$ (N تعداد کل نورون های مخفی است) نگاشت می کند، با $\mathcal{C} : \mathcal{X} \rightarrow \{۰, ۱\}^N$ نمایش می دهیم. بنابراین خواهیم داشت:

$$a^{(l)} = \sigma\left(z^{(l)}\right) = c^{(l)} \odot z^{(l)}$$

به طوری که در عبارت بالا $\odot$ نشان دهنده ضرب elementwise می باشد. در نتیجه می توان نوشت:

$$z^{(l+۱)} = W^{(l)} a^{(l)} + b^{(l)} = W^{(l)}\left(c^{(l)} \odot z^{(l)}\right) + b^{(l)} = \tilde{W}^{(l)} z^{(l)} + b^{(l)}$$

به طوری که در عبارت بالا، $\tilde{W}^{(l)} = W^{(l)} \odot c^{(l)}$ می باشد. با جایگذاری مکرر تساوی اخیر برای لایه های قبل تر، می توان $z^{(l+۱)}$ را برای $l \in \{۱, \cdots, L\}$ به شکل زیر بازنویسی کرد:

$$z^{(l+۱)} = \left(\prod_{k=۱}^{l} \tilde{W}^{(k)}\right) z^{(۱)} + \sum_{h=۱}^{l}\left(\prod_{k=h+۱}^{l} \tilde{W}^{(k)}\right) b^{(h)}$$

با جایگذاری $z^{(۱)} = W^{(۰)} x + b^{(۱)}$ در تساوی اخیر به دست می آید:

$$\begin{aligned} z^{(l+۱)} &= \left(\prod_{k=۱}^{l} \tilde{W}^{(k)}\right)\left(W^{(۰)} x + b^{(۱)}\right) + \sum_{h=۱}^{l}\left(\prod_{k=h+۱}^{l} \tilde{W}^{(k)}\right) b^{(h)} \\ &= \left(\prod_{k=۰}^{l} \tilde{W}^{(k)}\right) x + b^{(۱)}\left(\prod_{k=۱}^{l} \tilde{W}^{(k)}\right) + \sum_{h=۱}^{l}\left(\prod_{k=h+۱}^{l} \tilde{W}^{(k)}\right) b^{(h)} \\ &= \tilde{W}^{(۰:l)} x + b^{(۰:l)} \end{aligned}$$

به طوری که در تساوی صفحه قبل، $\tilde{W}^{(\cdot:l)}$ ماتریس ضریب x و $b^{(\cdot:l)}$ جمع بایاس های باقی مانده می باشد. با این تعریف، به وضوح داریم:

$$F_A(x, W) = \operatorname{softmax}\left(z^{(L+۱)}\right) = \operatorname{softmax}\left(\tilde{W}^{(\cdot:L)}x + b^{(\cdot:L)}\right)$$

برای یک شبکه عصبی و یک نمونه x مشخص و ثابت، $\tilde{W}^{(\cdot:l)}$ و $b^{(\cdot:l)}$ پارامتر های ثابتی هستند که با توجه به ماتریس وضعیت نورون های مخفی یا همان $C = \mathcal{C}(x)$ به صورت یکتا و منحصر به فرد تعیین می شوند. به علاوه، ماتریس C منجر به ایجاد یک خانواده از نابرابری ها به شکل زیر می شود: (با توجه به اولین تساوی این بخش)

$$\left(۲c^{(l+۱)} - ۱\right) \odot \left(\tilde{W}^{(\cdot:l)}x + b^{(\cdot:l)}\right) \geqslant ۰ \quad , \quad l \in \{۰, \cdots, L-۱\}$$

که این نمایش دهنده یک convex polytope می باشد و در نتیجه حکم مسئله برقرار است.

برای مشاهده منبع حل این سوال، روی این لینک کلیک کنید.

ب) درابتدا نشان می دهیم که $r(k,m) \leqslant \mathcal{O}(k^m)$ می باشد. برای این منظور در ابتدا دقت کنید که می توان نوشت:

$$\binom{k}{i} = \frac{k!}{(k-i)!i!} = \frac{k(k-۱)\cdots(k-i+۱)}{i!} \leqslant \frac{k^i}{i!} \tag{۱}$$

از طرف دیگر می توان نوشت:

$$e^i = \sum_{j=۰}^{\infty} \frac{i^j}{j!} \geqslant \frac{i^i}{i!} \quad \Longrightarrow \quad \frac{۱}{i!} \leqslant \left(\frac{e}{i}\right)^i \tag{۲}$$

از نابرابری های (۱) و (۲) نتیجه می شود:

$$\binom{k}{i} \leqslant \left(\frac{ek}{i}\right)^i$$

در نتیجه خواهیم داشت:

$$\begin{aligned} r(k,m) \leqslant \sum_{i=۰}^{m} \binom{k}{i} &\leqslant \sum_{i=۰}^{m} \left(\frac{ek}{i}\right)^i \\ &\leqslant ۱ + m + \frac{m(m-۱)}{۲} + \sum_{i=۳}^{m} k^i \\ &= \mathcal{O}(m^۲) + \frac{k^{m+۱} - k^۳}{k-۱} = \mathcal{O}(k^m) \end{aligned}$$

دقت کنید که نابرابری های موجود در مراحل بالا برای $k > ۱$ نوشته شده اند، چرا که برای $k = ۱$ به وضوح $r(k,m) = \mathcal{O}(۱)$ می باشد.

حال توجه نمایید که هر یک از نورون های موجود در شبکه عصبی نمایش دهنده یک خط می باشد؛ چرا که خروجی آن به شکل $\sigma\left(W_i^{(l)} h_i^{(l)} + b_i^{(l)}\right)$ است که مشخص می کند $h_i^{(l)}$ در بالا یا پایین خطی که با بردار نرمال $W_i^{(l)}$ و بایاس $b_i^{(l)}$ تعیین می شود، قرار می گیرد. بنابراین با عبور از هر لایه در یک شبکه عصبی، k خط جدید در فضای نمونه ها ایجاد می شوند و تعداد نواحی ایجاد شده در این فضا را $r(k, m) \leqslant \mathcal{O}(k^m)$ برابر می کنند. از آنجا که در ابتدا کل فضا یک ناحیه واحد است، پس از عبور از تمامی n لایه موجود در شبکه عصبی تعداد کل نواحی موجود در فضای نمونه ها حداکثر $\mathcal{O}\left((k^m)^n\right) = \mathcal{O}(k^{mn})$ خواهد بود. از طرف دیگر، تناظری یک به یک میان activation pattern ها و تعداد نواحی ایجاد شده در فضای نمونه ها وجود دارد و می توان نتیجه گرفت:

$$\mathcal{A}\left(F_{A_{(n,k)}}\left(\mathbb{R}^m; W\right)\right) \leqslant \mathcal{O}(k^{mn})$$

که این همان حکم مسئله است.

در حل این تمرین با **عسل مسکین** همفکری صورت گرفته است.