حمیدرضا امیرزاده
۴۰۱۲۰۶۹۹۹

موضوع: تمارین سری اول درس Deep Learning   تاریخ:

① الف) در رابطه $w = X^{-1}y$، ممکن است X یک ماتریس مربعی نباشد و در نتیجه معکوس پذیر نیست.

اگر از تابع هزینه $J = \frac{1}{m}\sum_{i}^{m}(y_i-\hat{y}_i)^2$ نسبت به w مشتق بگیریم و برابر صفر قرار دهیم پاسخ فرم بسته به صورت زیر به دست خواهد آمد:
$$w = (X^TX)^{-1}X^Ty$$

ب) عدم وجود auto correlation که یکی از فرضیات مسائل رگرسیون خطی است به معنای فرض عدم ~~وجود~~ وجود ارتباط (correlation) بین داده‌ها و خطاهاست.

برای بررسی وجود یا عدم وجود این ارتباط یک تست آماری به نام durbin watson وجود دارد.

← خطی که با کمک رگرسیون خطی بر داده‌ها فیت می‌شود برابر است با $y = 0.976x + 12.702$

← در این تست آماری مطابق جداول موجود برای $\alpha = 0.05$ مقادیر بحرانی زیر را داریم: $d_U = 1.32$ ، $d_L = 0.879$

← در این تست آماری d از رابطه زیر به دست می‌آید:

$$d = \frac{\sum_{t=2}(e_t - e_{t-1})^2}{\sum_{t=1} e_t^2}$$

جدول زیر را تشکیل می‌دهیم:

n	x	y	$\hat{y}$	error	$error^2$
۱	۴۸٫۹	۶۰٫۸	۶۰٫۴۳۸	۰٫۳۰۲	۰٫۱۳۸۳
۲	۵۰٫۶	۶۲٫۵	۶۲٫۰۸	۰٫۴۲	۰٫۱۷۶۴
۳	۵۲٫۹	۶۴٫۶	۶۴٫۳۳	۰٫۲۷	۰٫۰۷۲
۴	۵۵	۶۶٫۱	۶۶٫۳۸	-۰٫۳۸	۰٫۰۷۸
۵	۵۶٫۸	۶۷٫۷	۶۸٫۱۳	-۰٫۴۳	۰٫۱۸۴
۶	۵۸٫۸	۶۹٫۱	۷۰٫۰۹	-۰٫۹۹	۰٫۹۸
۷	۶۱٫۲	۷۱٫۷	۷۲٫۴۳	-۰٫۷۳	۰٫۵۳۲
۸	۶۳٫۵	۷۳٫۵	۷۳٫۷	-۰٫۲	۰٫۰۴
۹	۶۴٫۷	۷۶٫۲	۷۵٫۸۴	۰٫۳۶	۰٫۱۲۹
۱۰	۶۵	۷۷٫۳	۷۶٫۴	۱٫۱۶	۱٫۳۴۵

مجموع مجذورات خطا برابر است با ۳٫۶۷۴

$\varepsilon_2 - \varepsilon_1 =$ ۰٫۰۴۸ $\xrightarrow{\text{توان ۲}}$ ۰٫۰۰۲

$\varepsilon_3 - \varepsilon_2 =$ −۰٫۱۵ $\rightarrow$ ۰٫۰۲۲

$\varepsilon_4 - \varepsilon_3 =$ −۰٫۵۵ $\rightarrow$ ۰٫۳۰۲

$\varepsilon_5 - \varepsilon_4 =$ ۰٫۱۵ $\rightarrow$ ۰٫۰۲۲

$\varepsilon_6 - \varepsilon_5 =$ −۰٫۵۶ $\rightarrow$ ۰٫۳۱۳

$\varepsilon_7 - \varepsilon_6 =$ ۰٫۲۶ $\rightarrow$ ۰٫۰۶۷

$\varepsilon_8 - \varepsilon_7 =$ ۰٫۵۳ $\rightarrow$ ۰٫۲۸

$\varepsilon_9 - \varepsilon_8 =$ ۰٫۵۶ $\rightarrow$ ۰٫۳۱۳

$\varepsilon_{10} - \varepsilon_9 =$ ۰٫۸ $\rightarrow$ ۰٫۶۴

مجموع مجذورات اختلافات خطاها برابر است با : ۱٫۹۶

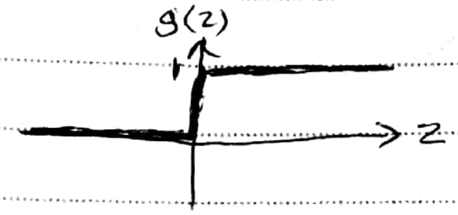
$$d = \frac{\sum_{t=2} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1} \varepsilon_t^2} = \frac{1.96}{3.674} = 0.53$$

$d = 0.53 < d_L = 0.879 \Rightarrow$ داده‌ها Postively autocorrelated هستند

$4 - d = 3.47 \not< d_u = 1.32 \Rightarrow$ داده‌ها Negatively autocorrelated نیستند

② توابع فعال سازی الف)

• Binary step



← برای $z \neq 0$ مشکل محو شدن گرادیان را دارد زیرا در این نقاط گرادیان برابر صفر است.

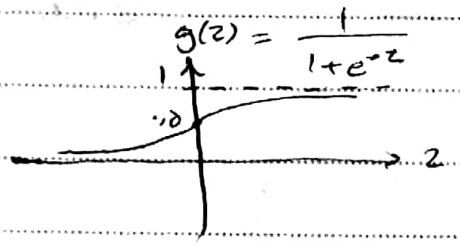
$$g(z)=U(z)=\begin{cases}1 & ; z>0\\ 0 & ; z<0\end{cases} \qquad g'(z)=U'(z)=\begin{cases}0 & ; z>0\\ \delta(0) & ; z=0\\ 0 & ; z<0\end{cases}$$

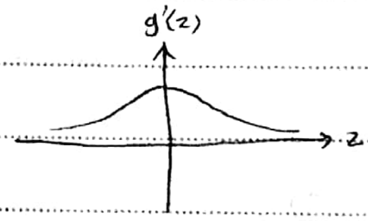
← برای $z \neq 0$ مشتق پذیر است ولی در $z=0$ مشتق ناپذیر است که می‌توان مشتق آن را افزایی گرفت و در آن نقطه مشتق را برابر میانگین مشتقات راست و چپ قرار داد.

← از نظر هزینه محاسباتی عملاً کمترین هزینه ممکن را دارد و مشکلی از این لحاظ وجود ندارد.

← zero-centered نیست.

• Sigmoid





$$g(z)=\sigma(z)=\frac{1}{1+e^{-z}} \qquad g'(z)=\sigma'(z)=\frac{1}{1+e^{-z}}\left(\frac{e^{-z}}{1+e^{-z}}\right)=\sigma(z)(1-\sigma(z))$$

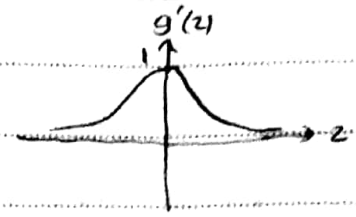
← برای $z \to \pm\infty$ مشکل محو شدن گرادیان وجود دارد زیرا گرادیان به صفر میل می‌کند.

← به ازای تمامی نقاط مشتق پذیر است.

← از نظر هزینه محاسباتی دارای پیچیدگی است و استفاده از آن ممکن است باعث کند شدن فرآیند یادگیری شود.

← zero-centered نیست.

- tanh



$g(z) = \tanh(z) = \frac{2}{1+e^{-2z}} - 1 = 2\sigma(2z) - 1$ $g'(z) = 1 - (\tanh(z))^2$

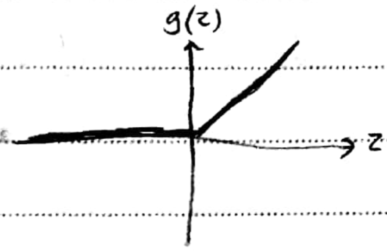
← برای $z \to \pm\infty$ مشکل محو شدن گرادیان وجود دارد زیرا گرادیان به صفر میل می‌کند

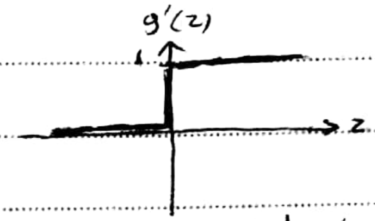
← به ازای تمامی نقاط مشتق‌پذیر است.

← از نظر محاسباتی دارای پیچیدگی است و ممکن است باعث کند شدن روند یادگیری شود.

← zero-centered است

- Relu





$g(z) = Relu(z) = \max(0, z)$ $g'(z) = \begin{cases} 1 & ; z>0 \\ ? & ; z=0 \\ 0 & ; z<0 \end{cases}$

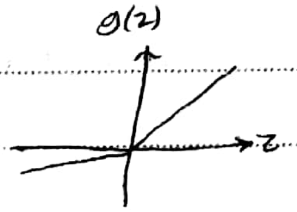
← برای $z<0$ مشکل محو شدن گرادیان وجود دارد زیرا گرادیان برابر صفر است.

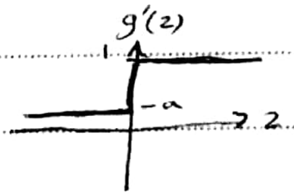
← به ازای $z \neq 0$ مشتق‌پذیر است. در $z=0$ مشتق‌ناپذیر است که می‌توان مشتق نرم‌افزاری گرفت و برابر ۰٫۵ قرار داد.

← از نظر محاسباتی بسیار ساده است.

← zero-centered نیست

- Leaky Relu





$g(z) = LRelu(z) = \begin{cases} z & ; z>0 \\ az & ; z<0 \end{cases}$ $g'(z) = \begin{cases} 1 & ; z>0 \\ ? & ; z=0 \\ a & ; z<0 \end{cases}$

← مشکل محو شدن گرادیان وجود ندارد.

← به ازای $z \neq 0$ مشتق‌پذیر است. در $z=0$ مشتق‌ناپذیر است که می‌توان مشتق نرم‌افزاری گرفت و برابر $\frac{1-a}{2}$ قرار داد.

← از نظر محاسباتی بسیار ساده است

← zero-centered نیست

موضوع : .... تاریخ : ....

sigmoid

ب) تابع فعال ساز swish به صورت روبرو تعریف شده است: $g(x) = x\sigma(\beta x)$

$\beta = 0 \Rightarrow g(x) = x/2$ ، $\beta = 1 \Rightarrow g(x) = x\sigma(x)$ ، $\beta \to \infty \Rightarrow g(x) \to Relu(x)$

مشتق این تابع به صورت زیر است: $g'(x) = \sigma(\beta x) + \beta x \sigma(\beta x)(1 - \sigma(\beta x))$

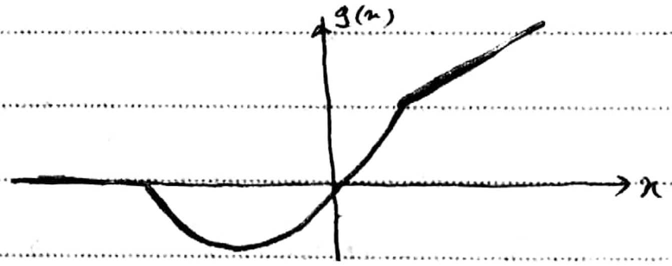
* از آنجایی که وقتی $\beta = 0$ آنگاه $g(-x) = -g(x)$ آنگاه $g(x)$ به ازای $\beta = 0$ ، zero-centered است.

* از روی نمودارها می‌توان دریافت که وقتی $x \to -\infty$ مشکل محو شدن گرادیان وجود دارد زیرا گرادیان به صفر میل می‌کند. لازم به ذکر است که با افزایش β مشتق زودتر به صفر میل می‌کند.

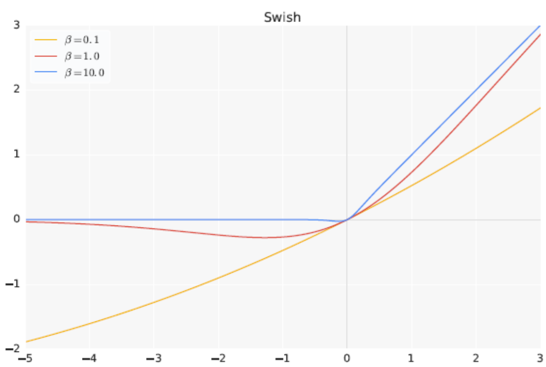
* از نظر عملکرد swish بر روی دیتاست‌های CIFAR-10 ، CIFAR-100 ، IMAGENET در تمامی موارد عملکردی بهتر یا مشابه Relu داشته است. اما در کاربرد ترجمه ماشینی، بر روی دیتاست WMT2014 در برخی موارد Relu عملکرد بهتری داشته است.

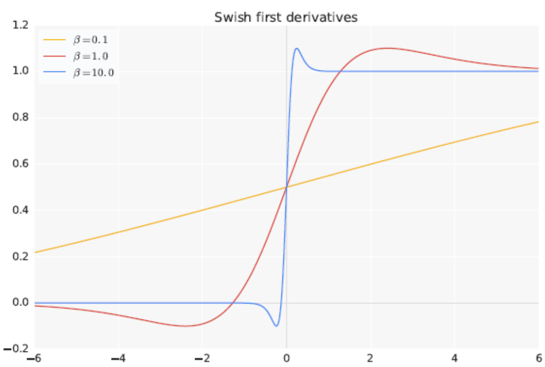
- تابع فعال ساز hard swish به صورت روبرو است: $g(x) = \frac{x}{6} Relu6(x+3)$

$$Relu6(x) = \begin{cases} 0 & ; \quad x < 0 \\ x & ; \quad 0 \leq x \leq 6 \\ 6 & ; \quad x > 6 \end{cases}$$



* از آنجایی که در رابطه تابع فعال ساز swish از تابع sigmoid استفاده شده، هزینه محاسباتی بیشتری نسبت به استفاده از تابع فعال ساز hard swish دارد. بنابراین می‌توان گفت که در دستگاه‌هایی که منابع محاسباتی کمتری دارند نمی‌توان از swish استفاده کرد اما می‌توان از hard swish استفاده کرد و با این حال که هزینه محاسباتی کمتری دارد، دقتی مشابه swish دارد.





موضوع:
تاریخ:

۳) الف)

RMSprop → مشکی

(۱) زیرا در سطوح آشنا زودتر همگرا می‌شود. نوسانات آن در یک جهت کمتر از جهت دیگر است. بنابراین می‌توان با برداشتن گام‌های بلندتر زودتر به جواب رسید.

Momentum → آبی

از آنجایی که در این روش گرادیان‌های قبلی نیز لحاظ می‌شوند، می‌توان گفت که از جمع‌شان گرادیان‌ها استفاده می‌کند. بنابراین نسبت به gd نوسانات کمتری دارد و زودتر همگرا می‌شود.

Nesterov Momentum → قرمز

زیرا نقطه‌ی بعدی که گرادیان در آن حساب می‌شود را پیش‌بینی می‌کند و برای همین نوسانات کمتری نسبت به Momentum دارد. بنابراین زودتر از Momentum همگرا می‌شود.

Gradient descent → سبز

از آنجایی که gd فقط براساس گرادیان در نقطه‌ی فعلی تصمیم می‌گیرد و در جهت آن حرکت می‌کند، ممکن است دچار نوسانات زیادی شود.

(۲) مزایای این سه روش در بالا ذکر شده. از معایب روش‌های ذکرشده می‌توان به نیاز داشتن به چندین هایپرپارامتر که نیاز به تنظیم شدن دارند ذکر کرد. در واقع تعداد هایپرپارامترها بیشتر می‌شود.

(۳) روش Adam از آنجایی که ترکیب روش‌های Momentum و RMSprop است و نسبت به روش Momentum از مزایای روش RMSprop نیز بهره می‌برد. بنابراین می‌توان گفت که در روش Momentum ممکن است دچار مینیمم‌های محلی شویم. اما با روش Adam شانس بیشتری برای فرار از این مینیمم‌های محلی وجود دارد و به سمت flat Minimum ها حرکت می‌کنیم.

در مورد روش bias correction ها، از آنجایی که مقادیر اولیه برای m_0 و v_0 داریم، به کمک تقسیم کردن بر $1-\beta^t$ اثر این مقادیر اولیه را کم می‌کنیم و حساسیت محاسبات اولیه به β_1 و β_2 با این تقسیم کاهش می‌یابد.

۳ ب) Multi-task learning زمانی اطلاق می‌شود که یک مدل شبکه عصبی سعی می‌کند که بتواند چندین تسک را به صورت همزمان یاد بگیرد. این تکنیک زمانی استفاده می‌شود که تسک‌هایی که می‌خواهیم یاد گرفته شوند، بتوانند از low-level feature های مشترک بهره ببرند. در این صورت است که کارایی یک شبکه عصبی multi task بهتر از چندین شبکه عصبی single task می‌باشد. از آنجایی که تمامی تسک‌ها از این ویژگی‌های سطح پایین مشترک استفاده می‌کنند، می‌گوییم که این ویژگی shared parameter هستند.

دلیل اینکه این روش می‌تواند باعث افزایش generalization مدل شود این است که می‌توان از داده‌های بیشتری برای آموزش شبکه استفاده کرد. زیرا می‌توانیم از داده‌هایی استفاده کنیم که شاید برای تنها یک تسک باشند اما درنهایت از آنجایی که از ویژگی shared parameter استفاده می‌کنیم، باعث می‌شود که عملکرد مدل روی دیگر تسک‌ها نیز افزایش یابد. در واقع می‌توان از ویژگی‌های سطح پایین برای تسک‌ها که مشترک هستند بهره برد.

۳ ج)

(۱) از آنجایی که در روش drop out به صورت تصادفی در هر iteration واحدهایی را حذف می‌کنیم می‌توان گفت که Complexity شبکه کاهش می‌یابد و درنتیجه مدل نیز کاهش می‌یابد و این یعنی کاهش overfit یا افزایش عمومیت پذیری و قدرت تعمیم پذیری مدل.

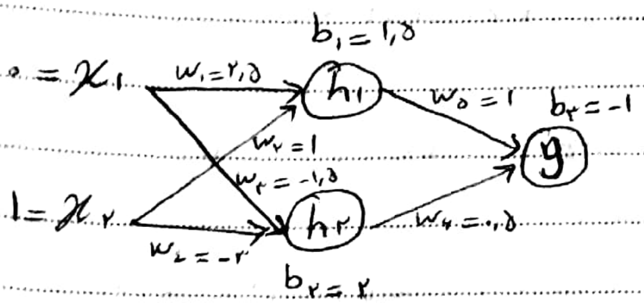
(۲) در ensemble learning چندین مدل با قدرت نسبی کمتر آموزش می‌بینند و نتیجه نهایی براساس نظر اکثریت مدل‌ها است. در روش dropout نیز به طور مشابه در هر iteration یک مدل کم قدرت‌تر را آموزش می‌دهیم (با حذف کردن تصادفی تعدادی از واحدهای پنهان). درنهایت می‌توان گفت میانگینی از این مدل‌های کم قدرت‌تر را استفاده می‌کنیم و درنتیجه نهایی نظر اکثریت مدل‌ها تأثیرگذار بوده است.

(۳) استفاده از dropout حین train به این صورت است که برای هر لایه یک احتمال نگهداری واحدهای آن لایه را درنظر می‌گیریم و به این صورت به یک شبکه خلاصه‌تر می‌رسیم. اما در زمان test [redacted] توصیه می‌شود که از dropout استفاده نکنیم. تمامی واحدهای هر لایه را نگه داریم. زیرا در زمان test می‌خواهیم از تمام قابلیت مدل و پارامترهایی که شبکه یاد گرفته است استفاده کنیم.

موضوع: .... تاریخ: ....

Backpropagation ۴

الف) $\alpha = 0.1$, $y = 1$



forward: $net_{h_1} = w_1 x_1 + w_2 x_2 + b_1 = 2.5 \times 0 + 1 \times 1 + 1.5 = 2.5$

$out_{h_1} = \sigma(net_{h_1}) = 1/1+e^{-2.5} = 0.924$

$net_{h_2} = w_3 x_1 + w_4 x_2 + b_2 = -1.5 \times 0 - 3 \times 1 + 2 = -1$

$out_{h_2} = \sigma(net_{h_2}) = 1/1+e^{+1} = 0.269$

$net_y = w_5\, out_{h_1} + w_6\, out_{h_2} + b_3 = 1 \times 0.924 + 0.5 \times 0.269 + (-1) = 0.059$

$out_y = \sigma(net_y) = 1/1+e^{-0.0589} = 0.515$

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

$$\xrightarrow{m=1} J = -\left(y \log out_y + (1-y) \log(1-out_y) \right) \overset{\text{جایگذاری}}{=} 0.5604$$

Backward: $$\frac{\partial J}{\partial w_5} = \frac{\partial J}{\partial out_y} \frac{\partial out_y}{\partial net_y} \frac{\partial net_y}{\partial w_5} = -1.941 \times 0.249 \times 0.924 = -0.446$$

$$\partial J / \partial out_y = -\left(\frac{y}{out_y} - \frac{1-y}{1-out_y} \right) = -\left(\frac{1}{0.515} - \frac{1-1}{1-0.515} \right) = -1.941$$

$\partial out_y / \partial net_y = \sigma_{net_y} (1 - \sigma_{net_y}) = 0.515 (1 - 0.515) = 0.249$

$\partial net_y / \partial w_5 = out_{h_1} = 0.924$

$$\frac{\partial J}{\partial w_6} = \frac{\partial J}{\partial out_y} \frac{\partial out_y}{\partial net_y} \frac{\partial net_y}{\partial w_6} = -1.941 \times 0.249 \times 0.269 = -0.130$$

$\partial net_y / \partial w_6 = out_{h_2} = 0.269$

$$\frac{\partial J}{\partial b_3} = \frac{\partial J}{\partial out_y} \frac{\partial out_y}{\partial net_y} \frac{\partial net_y}{\partial b_3} = -1.941 \times 0.249 \times 1 = -0.483$$

$\partial net_y / \partial b_3 = 1$

Scanned with CamScanner

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial out_{h_1}} \frac{\partial out_{h_1}}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_1} = -۰٫۴۸۸ \times ۰٫۰۷۰۲ \times ۰ = ۰$$

$$\frac{\partial J}{\partial out_{h_1}} = \frac{\partial J}{\partial out_y} \frac{\partial out_y}{\partial net_y} \frac{\partial net_y}{\partial out_{h_1}} = -۱٫۹۶۱ \times ۰٫۲۴۹ \times ۱ = -۰٫۴۸۸$$

$\partial net_y / \partial out_{h_1} = w_5 = 1$ $\qquad \frac{\partial out_{h_1}}{\partial net_{h_1}} = \sigma_{net_{h_1}}(1-\sigma_{net_{h_1}}) = ۰٫۹۲۴(۱-۰٫۹۲۴) = ۰٫۰۷۰۲$

$\partial net_{h_1} / \partial w_1 = x_1 = ۰$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial out_{h_1}} \frac{\partial out_{h_1}}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_2} = -۰٫۴۸۸ \times ۰٫۰۷۰۲ \times ۱ = -۰٫۰۳۴۳$$

$\partial net_{h_1} / \partial w_2 = x_2 = 1$

$$\frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial out_{h_1}} \frac{\partial out_{h_1}}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial b_1} = -۰٫۴۸۸ \times ۰٫۰۷۰۲ \times ۱ = -۰٫۰۳۴۳$$

$\partial net_{h_1} / \partial b_1 = 1$

$$\frac{\partial J}{\partial w_3} = \frac{\partial J}{\partial out_{h_2}} \frac{\partial out_{h_2}}{\partial net_{h_2}} \frac{\partial net_{h_2}}{\partial w_3} = -۰٫۲۴۴ \times ۰٫۱۹۴ \times ۰ = ۰$$

$$\frac{\partial J}{\partial out_{h_2}} = \frac{\partial J}{\partial out_y} \frac{\partial out_y}{\partial net_y} \frac{\partial net_y}{\partial out_{h_2}} = -۱٫۹۶۱ \times ۰٫۲۴۹ \times ۰٫۵ = -۰٫۲۴۴$$

$\partial net_y / \partial out_{h_2} = w_6 = ۰٫۵$ $\qquad \frac{\partial out_{h_2}}{\partial net_{h_2}} = \sigma_{net_{h_2}}(1-\sigma_{net_{h_2}}) = ۰٫۷۳۹(۱-۰٫۷۳۹) = ۰٫۱۹۴$

$\partial net_{h_2} / \partial w_3 = x_1 = ۰$

$$\frac{\partial J}{\partial w_4} = \frac{\partial J}{\partial out_{h_2}} \frac{\partial out_{h_2}}{\partial net_{h_2}} \frac{\partial net_{h_2}}{\partial w_4} = -۰٫۲۴۴ \times ۰٫۱۹۴ \times ۱ = -۰٫۰۴۷۲$$

$\partial net_{h_2} / \partial w_4 = x_2 = 1$ $\qquad \frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial out_{h_2}} \frac{\partial out_{h_2}}{\partial net_{h_2}} \frac{\partial net_{h_2}}{\partial b_2} = -۰٫۲۴۴ \times ۰٫۱۹۴ \times ۱ = -۰٫۰۴۷۲$

حال می‌توان وزن‌ها و بایاس‌ها را مطابق الگوریتم gradient descent به روز رسانی کرد:

$$\frac{\partial J}{\partial w_1} = 0 \qquad \frac{\partial J}{\partial w_2} = -0.0339 \qquad \frac{\partial J}{\partial w_3} = 0 \qquad \frac{\partial J}{\partial w_4} = -0.0472$$

$$\frac{\partial J}{\partial w_5} = -0.344 \qquad \frac{\partial J}{\partial w_6} = -0.129 \qquad \frac{\partial J}{\partial b_1} = -0.0339 \qquad \frac{\partial J}{\partial b_2} = -0.0472 \qquad \frac{\partial J}{\partial b_3} = -0.483$$

$$w_1^+ = w_1 - \alpha \frac{\partial J}{\partial w_1} = 2.5 - 0.1 \times 0 = 2.5$$

$$w_2^+ = w_2 - \alpha \frac{\partial J}{\partial w_2} = 1 - 0.1 \times (-0.0339) = 1.00339$$

$$w_3^+ = w_3 - \alpha \frac{\partial J}{\partial w_3} = -1.5 - 0.1 \times 0 = -1.5$$

$$w_4^+ = w_4 - \alpha \frac{\partial J}{\partial w_4} = -3 - 0.1 \times (-0.0472) = -2.9953$$

$$w_5^+ = w_5 - \alpha \frac{\partial J}{\partial w_5} = 1 - 0.1 \times (-0.344) = 1.0344$$

$$w_6^+ = w_6 - \alpha \frac{\partial J}{\partial w_6} = 0.5 - 0.1 \times (-0.129) = 0.5129$$

$$b_1^+ = b_1 - \alpha \frac{\partial J}{\partial b_1} = 1.5 - 0.1 \times (-0.0339) = 1.50339$$

$$b_2^+ = b_2 - \alpha \frac{\partial J}{\partial b_2} = 2 - 0.1 \times (-0.0472) = 2.00472$$

$$b_3^+ = b_3 - \alpha \frac{\partial J}{\partial b_3} = -1 - 0.1 \times (-0.483) = -0.95171$$

موضوع : .... تاریخ : ....

ب)

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial h_2}\,\frac{\partial h_2}{\partial z_2}\,\frac{\partial z_2}{\partial w_2}$$

$$\frac{\partial J}{\partial h_{2j}} = \sum_{i=1}^{D_y} \frac{\partial J}{\partial \hat{y}_i}\,\frac{\partial \hat{y}_i}{\partial h_{2j}}$$

$$J = -\sum_{i=1}^{D_y} y_i \log \hat{y}_i \rightsquigarrow \frac{\partial J}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

$$\begin{cases} \hat{y}_1 = \dfrac{e^{h_{21}}}{\sum_{i=1}^{D_y} e^{h_{2i}}} \rightsquigarrow \dfrac{\partial \hat{y}_1}{\partial h_{21}} = \dfrac{e^{h_{21}}\sum e^{h_{2i}} - (e^{h_{21}})^2}{(\sum e^{h_{2i}})^2} = \hat{y}_1(1-\hat{y}_1) \\ \hat{y}_2 = \dfrac{e^{h_{22}}}{\sum e^{h_{2i}}} \rightsquigarrow \dfrac{\partial \hat{y}_2}{\partial h_{21}} = \dfrac{-e^{h_{21}} e^{h_{22}}}{(\sum e^{h_{2i}})^2} = -\hat{y}_1 \hat{y}_2 \\ \hat{y}_3 = \dfrac{e^{h_{23}}}{\sum e^{h_{2i}}} \rightsquigarrow \dfrac{\partial \hat{y}_3}{\partial h_{21}} = \dfrac{-e^{h_{21}} e^{h_{23}}}{(\sum e^{h_{2i}})^2} = -\hat{y}_1 \hat{y}_3 \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \hat{y}_{D_y} = \dfrac{e^{h_{2D_y}}}{\sum e^{h_{2i}}} \rightsquigarrow \dfrac{\partial \hat{y}_{D_y}}{\partial h_{21}} = \dfrac{-e^{h_{21}} e^{h_{2D_y}}}{(\sum e^{h_{2i}})^2} = -\hat{y}_1 \hat{y}_{D_y} \end{cases}$$

$$\Rightarrow \frac{\partial J}{\partial h_{21}} = -\frac{y_1}{\hat{y}_1}\hat{y}_1(1-\hat{y}_1) + \frac{y_2}{\hat{y}_2}\hat{y}_1\hat{y}_2 + \frac{y_3}{\hat{y}_3}\hat{y}_1\hat{y}_3 + \dots + \frac{y_{D_y}}{\hat{y}_{D_y}}\hat{y}_1\hat{y}_{D_y}$$

$$= \hat{y}_1 y_1 + \hat{y}_1 y_2 + \hat{y}_1 y_3 + \dots + \hat{y}_1 y_{D_y} - y_1 = \Big(\sum_{i=1}^{D_y} y_i\Big)\hat{y}_1 - y_1$$

به طریق مشابه می‌توان دریافت که :

$$\frac{\partial J}{\partial h_{2j}} = \Big(\sum_{i=1}^{D_y} y_i\Big)\hat{y}_j - y_j$$

$$\Rightarrow \frac{\partial J}{\partial h_2} = \begin{bmatrix} (\sum y_i)\hat{y}_1 - y_1 \\ (\sum y_i)\hat{y}_2 - y_2 \\ \vdots \\ (\sum y_i)\hat{y}_{D_y} - y_{D_y} \end{bmatrix} = \hat{y}\,[1\ 1 \cdots 1]_{1\times D_y}\, y - y$$

$$h_2 = Relu(z_2) \Rightarrow \frac{\partial h_2}{\partial z_2} = \begin{cases} 1 & ; \; z_2 > 0 \\ 0 & ; \; z_2 < 0 \end{cases}$$

$$z_2 = w_2 h_1 + b_2 \Rightarrow \frac{\partial z_2}{\partial w_2} = h_1 \quad , \quad \frac{\partial z_2}{\partial b_2} = 1$$

$$\Rightarrow \frac{\partial J}{\partial w_2} = \begin{cases} (\hat{y}\,[1\;1 \ldots 1]_{1\times D_y}\, y - y)\, h_1^T & ; \; z_2 > 0 \\ 0 & ; \; z_2 < 0 \end{cases}$$

$$\Rightarrow \frac{\partial J}{\partial b_2} = \begin{cases} (\hat{y}\,[1\;1 \ldots 1]_{1\times D_y}\, y - y) & ; \; z_2 > 0 \\ 0 & ; \; z_2 < 0 \end{cases}$$

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial h_2}\,\frac{\partial h_2}{\partial z_2}\,\frac{\partial z_2}{\partial h_1}\,\frac{\partial h_1}{\partial z_1}\,\frac{\partial z_1}{\partial w_1}$$

$$\Rightarrow \frac{\partial J}{\partial w_1} = \begin{cases} w_2^T\,(\hat{y}\,[1\;1 \ldots 1]_{1\times D_y}\, y - y)\, x^T & ; \; z_1, z_2 > 0 \\ 0 & ; \; else \end{cases}$$

$$\frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial h_2}\,\frac{\partial h_2}{\partial z_2}\,\frac{\partial z_2}{\partial h_1}\,\frac{\partial h_1}{\partial z_1}\,\frac{\partial z_1}{\partial b_1}$$

$$\Rightarrow \frac{\partial J}{\partial b_1} = \begin{cases} w_2^T\,(\hat{y}\,[1\;1 \ldots 1]_{1\times D_y}\, y - y) & ; \; z_1, z_2 > 0 \\ 0 & ; \; else \end{cases}$$