* تمرین جلسه 26 اسفند :

داریم : w ها , b ها $= \varphi$

پس وضعیت وزن ها در پایان epoch اول :

forward : $a_1^1 = g(x_1 w_{11}^1 + x_2 w_{12}^1 + b_1^1) = 0{,}5$

$$a_2^1 = g(x_1 w_{21}^1 + x_2 w_{22}^1 + b_2^1) = 0{,}5$$

$$a_1^2 = g(a_1^1 w_{11}^2 + a_2^1 w_{12}^1 + b_1^2) = 0{,}5$$

backward :

$$\frac{dL}{dw_{11}^1} = \underbrace{\frac{dL}{dz_1^1}}_{\delta_1^1} \cdot \underbrace{\frac{dz_1^1}{dw_{11}^1}}_{x_1} = (a_1^1 - y_1) \cdot x_1 = (0{,}5 - y_1)\, x_1$$

$$\frac{dL}{dw_{12}^1} = \underbrace{\frac{dL}{dz_1^1}}_{\delta_1^1} \cdot \underbrace{\frac{dz_1^1}{dw_{12}^1}}_{x_2} = (a_1^1 - y_1) \cdot x_2 = (0{,}5 - y_1) \cdot x_2$$

$$\frac{dL}{dw_{21}^1} = \underbrace{\frac{dL}{z_2^1}}_{\delta_2^1} \cdot \underbrace{\frac{dz_2^1}{dw_{21}^1}}_{x_1} = (a_2^1 - y_2)\, x_1 = (0{,}5 - y_2) \cdot x_1$$

$$\frac{dL}{dw_{22}^1} = \underbrace{\frac{dL}{dz_2^1}}_{\delta_2^1} \cdot \underbrace{\frac{dz_2^1}{dw_{22}^1}}_{x_2} = (a_2^1 - y_2)\, x_2 = (0{,}5 - y_2)\, x_2$$

$$\frac{dL}{dw^2_{11}} = \underbrace{\frac{dL}{dz^2_1}}_{\delta^2_1} \cdot \underbrace{\frac{dz^2_1}{dw^2_{11}}}_{a^1_1} = (a^2_1 - y_1)\, a^1_1 = \frac{0/5 - y_1}{2}$$

$$\frac{dL}{dw^2_{12}} = \underbrace{\frac{dL}{dz^2_1}}_{\delta^2_1} \cdot \underbrace{\frac{dz^2_1}{dw^2_{12}}}_{a^1_2} = (a^2_1 - y_1)\, a^1_2 = \frac{0/5 - y_1}{2}$$

update :

$$D^1_{11} = \frac{1}{m} \cdot \frac{dL}{dw^1_{11}} + \frac{\lambda}{m} \underbrace{w^1_{11}}_{0} = \frac{1}{m} (0/5 - y_1) x_1$$

$$D^1_{12} = \frac{1}{m} \frac{dL}{dw^1_{12}} + 0 = \frac{1}{m} (0/5 - y_1) \cdot x_2$$

$$D^1_{21} = \frac{1}{m} \cdot \frac{dL}{dw^1_{21}} = \frac{1}{m} (0/5 - y_2) x_1$$

$$D^1_{22} = \frac{1}{m} \cdot \frac{dL}{dw^1_{22}} = \frac{1}{m} (0/5 - y_2) x_2$$

$$D^2_{11} = \frac{1}{m} \cdot \frac{dL}{dw^2_{11}} = \frac{1}{2m} (0/5 - y_1)$$

$$D^2_{12} = \frac{1}{m} \cdot \frac{dL}{d^2_{12}} = \frac{1}{2m} (0/5 - y_1)$$

$$w^1_{11} = \underbrace{w^1_{11}}_{0} - \alpha D^1_{11} = -\frac{\alpha}{m} (0/5 - y_1) x_1$$

$$w^1_{12} = w^1_{12} - \alpha D^1_{12} = -\frac{\alpha}{m} (0/5 - y_1) x_2$$

$$w^1_{21} = w^1_{21} - \alpha D^1_{21} = -\frac{\alpha}{m} (0/5 - y_2) x_1$$

$$w^1_{22} = w^1_{22} - \alpha D^1_{22} = -\frac{\alpha}{m} (0/5 - y_2) x_2$$

$$w^2_{11} = w^2_{11} - \alpha D^2_{11} = -\frac{\alpha}{2m} (0/5 - y_1)$$

$$w^2_{12} = w^2_{12} - \alpha D^2_{12} = -\frac{\alpha}{2m} (0/5 - y_1)$$

$\Longrightarrow$ همانطور که مشخص است، وزن ها update می‌شوند و از صفر تغییر می‌کنند.

ب) وضعیت بایاس ها در پایان epoch اول :

backward :

$$\frac{dL}{db_1^1} = \delta_1^1 = a_1^1 - y_1 = 0{,}5 - y_1$$

$$\frac{dL}{db_2^1} = \delta_2^1 = a_2^1 - y_2 = 0{,}5 - y_2$$

$$\frac{dL}{db_1^2} = \delta_1^2 = a_1^2 - y_1 = 0{,}5 - y_1$$

update :

$$B_1^1 = \frac{1}{m}\frac{dL}{db_1^1} = \frac{1}{m}(0{,}5 - y_1)$$

$$B_2^1 = \frac{1}{m}\frac{dL}{db_2^1} = \frac{1}{m}(0{,}5 - y_2)$$

$$B_1^2 = \frac{1}{m}\frac{dL}{db_1^2} = \frac{1}{m}(0{,}5 - y_1)$$

$$\left.\begin{aligned} b_1^1 &= b_1^1 - \alpha B_1^1 = -\frac{\alpha}{m}(0{,}5 - y_1) \\ b_2^1 &= b_2^1 - \alpha B_2^1 = -\frac{\alpha}{m}(0{,}5 - y_2) \\ b_1^2 &= b_1^2 - \alpha B_1^2 = -\frac{\alpha}{m}(0{,}5 - y_1) \end{aligned}\right\} \Longrightarrow$$

بایاس ها نیز update شدند

که وضعیت وزن‌ها در پایان epoch دوم:

forward:

$$a_1^1 = g\left(-\frac{\alpha}{m}(0{,}5-y_1)\,x_1^2 - \frac{\alpha}{m}(0{,}5-y_1)\,x_2^2 - \frac{\alpha}{m}(0{,}5-y_1)\right)$$

$$= g\left(-\frac{\alpha}{m}(0{,}5-y_1)(x_1^2 + x_2^2 + 1)\right) = m_1 \Rightarrow$$

(نامگذاری جهت کوتاه شدن روابط که در ادامه داریم.)

$$a_2^1 = g\left(-\frac{\alpha}{m}(0{,}5-y_2)\,x_1^2 - \frac{\alpha}{m}(0{,}5-y_2)\,x_2^2 - \frac{\alpha}{m}(0{,}5-y_2)\right)$$

$$= g\left(-\frac{\alpha}{m}(0{,}5-y_2)(x_1^2 + x_2^2 + 1)\right) = m_2$$

$$a_1^2 = g\Big(-\frac{\alpha}{2m}(0{,}5-y_1)\, g\left(-\frac{\alpha}{m}(0{,}5-y_1)(x_1^2+x_2^2+1)\right)$$

$$- \frac{\alpha}{2m}(0{,}5-y_1)\, g\left(-\frac{\alpha}{m}(0{,}5-y_1)(x_1^2+x_2^2+1)\right)$$

$$- \frac{\alpha}{m}(0{,}5-y_1)\Big) = m_3$$

backward:

$$\frac{dL}{dw_{11}^1} = (a_1^1 - y)\cdot x_1 = (m_1 - y_1)\cdot x_1$$

$$\frac{dL}{dw_{12}^1} = (a_1^1 - y_1)\,x_2 = (m_1 - y_1)\cdot x_2$$

$$\frac{dL}{dw_{21}^1} = (a_2^1 - y_2)\,x_1 = (m_2 - y_2)\cdot x_1$$

$$\frac{dL}{dw_{22}^1} = (a_2^1 - y_2)\,x_2 = (m_2 - y_2)\,x_2$$

$$\frac{dL}{dw_{11}^2} = (a_1^2 - y_1)\,a_1^1 = (m_3 - y_1)\cdot m_1$$

$$\frac{dL}{dw_{12}^2} = (a_1^2 - y_1)\,a_2^1 = (m_3 - y_1)\,m_2$$

update :

$$D^1_{11} = \frac{1}{m}\frac{dL}{dw^1_{11}} + \frac{\lambda}{m} w^1_{11} = \frac{1}{m}(m_1 - y_1)x_1 - \frac{\lambda\alpha}{m^2}(0{,}5 - y_1)x_1 = n_1$$

$$D^1_{12} = \frac{1}{m}\frac{dL}{dw^1_{12}} + \frac{\lambda}{m} w^1_{12} = \frac{1}{m}(m_1 - y_1)x_2 - \frac{\lambda\alpha}{m^2}(0{,}5 - y_1)x_2 = n_2$$

$$D^1_{21} = \frac{1}{m}\frac{dL}{dw^1_{21}} + \frac{\lambda}{m} w^1_{21} = \frac{1}{m}(m_2 - y_2)x_1 - \frac{\lambda\alpha}{m^2}(0{,}5 - y_2)x_1 = n_3$$

$$D^1_{22} = \frac{1}{m}\frac{dL}{dw^1_{22}} + \frac{\lambda}{m} w^1_{22} = \frac{1}{m}(m_2 - y_2)x_2 - \frac{\lambda\alpha}{m^2}(0{,}5 - y_2)x_2 = n_4$$

$$D^2_{11} = \frac{1}{m}\frac{dL}{dw^2_{11}} + \frac{\lambda}{m} w^2_{11} = \frac{1}{m}(m_3 - y_1)m_1 - \frac{\lambda\alpha}{2m^2}(0{,}5 - y_1) = n_5$$

$$D^2_{12} = \frac{1}{m}\frac{dL}{dw^2_{12}} + \frac{\lambda}{m} w^2_{12} = \frac{1}{m}(m_3 - y_1)m_2 - \frac{k\alpha}{2m^2}(0{,}5 - y_1) = n6$$

$$w^1_{11} = w^1_{11} - \alpha D^1_{11} = -\frac{\alpha}{m}(0{,}5 - y_1)x_1 - \alpha n_1$$

$$w^1_{12} = w^1_{12} - \alpha D^1_{12} = -\frac{\alpha}{m}(0{,}5 - y_1)x_2 - \alpha n_2$$

$$w^1_{21} = w^1_{21} - \alpha D^1_{21} = -\frac{\alpha}{m}(0{,}5 - y_2)x_1 - \alpha n_3$$

$$w^1_{22} = w^1_{22} - \alpha D^1_{22} = -\frac{\alpha}{m}(0{,}5 - y_2)x_2 - \alpha n_4$$

$$w^2_{11} = w^2_{11} - \alpha D^2_{11} = \frac{-\alpha}{2m}(0{,}5 - y_1) - \alpha n_5$$

$$w^2_{12} = w^2_{12} - \alpha D^2_{12} = \frac{-\alpha}{2m}(0{,}5 - y_1)$$

لذا وضعیت بایاس ها در پایان epoch دوم:

backward: $\frac{dL}{db^1_1} = \delta^1_1 = a^1_1 - y_1 = m_1 - y_1$

$$\frac{dL}{db^1_2} = \delta^1_2 = a^1_2 - y_2 = m_2 - y_2$$

$$\frac{dL}{db^2_1} = \delta^2_1 = a^2_1 - y_1 = m_3 - y_1$$

update: $B^1_1 = \frac{1}{m} \frac{dL}{db^1_1} = \frac{1}{m}(m_1 - y_1)$

$$B^1_2 = \frac{1}{m} \frac{dL}{db^1_2} = \frac{1}{m}(m_2 - y_2)$$

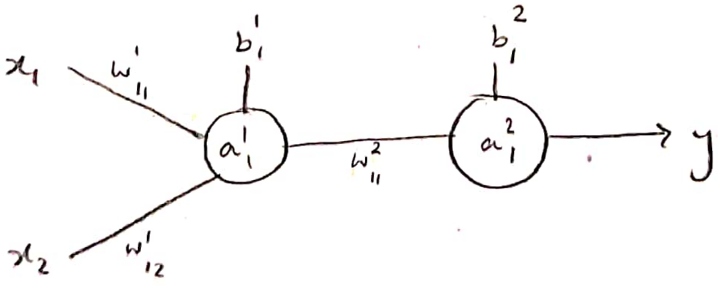
$$B^2_1 = \frac{1}{m} \frac{dL}{db^2_1} = \frac{1}{m}(m_3 - y_1)$$

با توجه به محاسبات :   (هیچ کدام از بایاس ها و وزن ها برابر صفر نمی مانند.)

$$w^1_{11} = w^1_{21} \quad و \quad w^1_{12} = w^1_{22} \quad و \quad b^1_1 = b^1_2 \implies$$

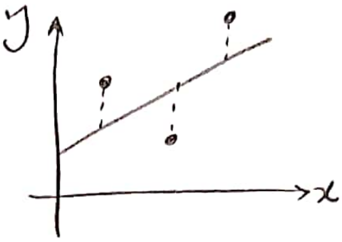
بایاس ها و وزن های نرون های لایه اول، به صورت نظیر به نظیر با هم برابر هستند.

با توجه به همین مسئله، می‌توان همه نرون های لایه ی 1 را فقط با یک نرون جایگزین کرد. شبکه عصبی معادل به این صورت می‌شود:



①

فاصله عمودی برای مشخص کردن این قضیه بهتر است. چرا که متغیر target را روی محور عمودی مختصات رسم کرده‌ایم. و فاصله‌ی مقدار تخمین زده شده و مقدار واقعی را باید نسبت به محور عمودی مختصات بدست آوریم. و نشان دهیم که برای یک x مشخص، فاصله‌ی y تخمین زده شده و y واقعی چقدر است.



فاصله عمودی ✓

---

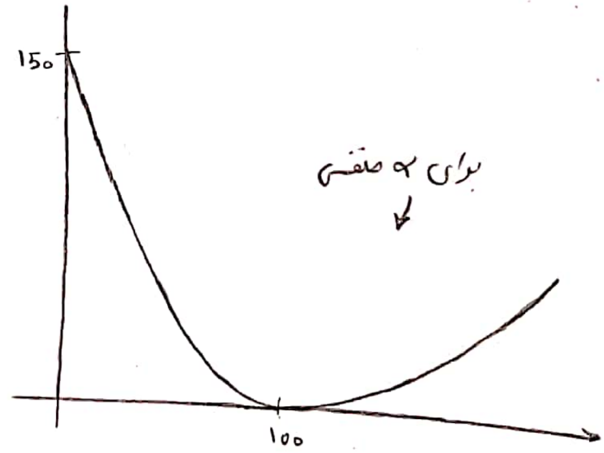
②

Cost_function : $x^2(\mathrm{sgn}(x)+\alpha)^2$

$-1<\alpha<0$

$-1<|\alpha|<1$

که این عدد را منفی انتخاب می‌کنیم تا برای جریمه‌ی زیادتری برای underestimation در نظر بگیرد.



پس این تابع، جریمه‌ی بیشتری برای underestimation قرار می‌دهد نسبت به overestimation. برای همین خیالمان راحت خواهد بود که پیش‌بینی حاصل از مدلمان، underestimate نخواهد بود.

$$f(x,y) = \left(\frac{3}{4}x - \frac{3}{2}\right)^2 + (y-2)^2 + \frac{1}{4}xy$$
$$= \frac{9}{16}x^2 + y^2 + \frac{1}{4}xy - \frac{9}{4}x - 4y + \frac{25}{4}$$

$$\frac{df(x,y)}{dx} = \frac{9}{8}x + \frac{1}{4}y - \frac{9}{4}$$

$$\frac{df(x,y)}{dy} = 2y + \frac{1}{4}x - 4$$

: gradient descent ←

repeat until convergence {

$$x = x - \alpha \frac{df(x,y)}{dx} = x - \alpha\left(\frac{9}{8}x + \frac{1}{4}y - \frac{9}{4}\right)$$

$$y = y - \alpha \frac{df(x,y)}{dy} = y - \alpha\left(2y + \frac{1}{4}x - 4\right)$$

}

* مرحله اول : $\alpha = 0{,}01$ — نقطه شروع : $(5,4)$

$(5,4)$: $f(5,4) = \boxed{14{,}06}$

1

$$x = 5 - (0{,}01)\underbrace{\left(\frac{45}{8} + 1 - \frac{9}{4}\right)}_{\frac{35}{8}} = 4{,}96$$

$$y = 4 - (0{,}01)\underbrace{\left(8 + \frac{5}{4} - 4\right)}_{\frac{11}{4}} = 3{,}97$$

$\Rightarrow (4{,}96\ ,\ 3{,}97)$

$\Rightarrow f(4{,}96, 3{,}97) = \boxed{13{,}73}$

2

⋮

در این حالت، الگوریتم خیلی کند در حال پیشروی است. و مقدار f خیلی آهسته در حال کاهش است، که این نشان دهنده این است که مقدار α خیلی کوچک است. برای اینکه به مقدار بهینه نزدیک شویم، تعداد زیادی مرحله باید طی کنیم...

* مرحله دوم : $\alpha = 0{,}5$ - نقطه شروع : $(5,4)$

$f(5,4) = \boxed{14{,}06}$

1

$x = 5 - (0{,}5)\left(\frac{35}{8}\right) = 2{,}81$

$y = 4 - (0{,}5)\left(\frac{11}{4}\right) = 2{,}63$

$\Rightarrow f(2{,}81 , 2{,}63) = \boxed{2{,}61}$

2

$x = 2{,}81 - (0{,}5)(1{,}57) = 2{,}03$

$y = 2{,}63 - (0{,}5)(1{,}96) = 1{,}65$

$\Rightarrow f(2{,}03 , 1{,}65) = \boxed{0{,}96}$

3

$x = 2{,}03 - (0{,}5)(0{,}45) = 1{,}81$

$y = 1{,}65 - (0{,}5)(-0{,}19) = 1{,}75$

$\Rightarrow f(1{,}81 , 1{,}75) = \boxed{0{,}874}$

4

$x = 1{,}81 - (0{,}5)(0{,}22) = 1{,}7$

$y = 1{,}75 - (0{,}5)(-0{,}05) = 1{,}78$

$\Rightarrow f(1{,}7 , 1{,}78) = \boxed{0{,}855}$

5

$x = 1{,}7 - (0{,}5)(0{,}11) = 1{,}65$

$y = 1{,}78 - (0{,}5)(-0{,}02) = 1{,}79$

$\Rightarrow f(1{,}65 , 1{,}79) = \boxed{0{,}851}$

$\Downarrow$

شرط contingency را این ترتیب تعریف کردیم ، که وقتی اختلاف دو f متوالی ، کمتر از 5×10^{-3} شد ، الگوریتم متوقف شود.

$\Downarrow$

بنابراین ، پس از <u>5 مرحله</u> متوقف می‌شود . و مقدار min را برای تابع f ، برابر <u>0,851</u> برمی‌گرداند . که همان مقدار بهینه است .

* مرحله سوم : $\alpha = 0,75$ و نقطه شروع: $(5,4)$

$f(5,4) = \boxed{14,06}$

1

$x = 5 - (0,75)\left(\frac{35}{8}\right) = 1,72$

$y = 4 - (0,75)\left(\frac{11}{4}\right) = 1,94$

$\Downarrow f(1,72 , 1,94) = \boxed{0,881}$

2

$x = 1,72 - (0,75)(0,17) = 1,59$

$y = 1,94 - (0,75)(0,31) = 1,71$

$\Downarrow f(1,59 , 1,71) = \boxed{0,858}$

3

$x = 1,59 - (0,75)(-0,03) = 1,61$

$y = 1,71 - (0,75)(-0,18) = 1,85$

$\Downarrow$

$f(1,61 , 1,85) = \boxed{0,852}$

4

$x = 1,61 - (0,75)(0,02) = 1,60$

$y = 1,85 - (0,75)(0,10) = 1,78$

$\Downarrow$

$f(1,6 , 1,78) = \boxed{0,850}$

$\Downarrow$

اینجا هم از همان شرط contingency که در قسمت قبل تعریف کرده بودیم، استفاده کردیم برای توقف الگوریتم

$\Downarrow$

در این حالت، پس از 4 مرحله به مقدار بهینه‌ی 0,850 حاصل شد. که نشان می‌دهد که با تعداد مرحله‌ی کمتر، مقدار بهینه‌تری بدست آوردیم نسبت به قسمت قبل. که یعنی نشان می‌دهد که α قبلی که کوچکتر بود، روند کندتری داشت. و این α مناسب‌تر است.

Scanned by CamScanner

(۴)

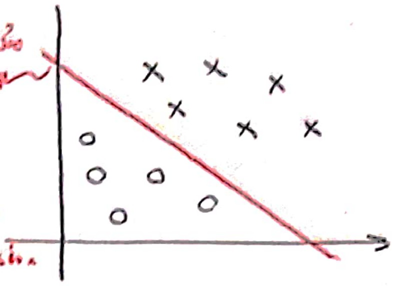
الف) تغییرات در f_1 خیلی آهسته درحال رخ دادن است. و مقدار f_1، خیلی آهسته دارد پایین می‌آید. این نشان دهنده‌ی این است که α خیلی کوچک است. که در این حالت، ما در نهایت به min خواهیم رسید، ولی زمان زیادی طول خواهد کشید.

ب) تغییرات در f_2 نظم یکنواختی ندارد و گاهی مقدار آن کمتر و گاهی بیشتر می‌شود. این نشان می‌دهد که مقدار α خیلی بزرگ است.

ج) تغییرات در f_3 با سرعت خوبی رو به کاهش است. که این نشان می‌دهد مقدار α مناسب است.

د) تغییرات f_4 به صورت کاهشی است و سرعت خوبی هم دارد. که نشان می‌دهد مقدار α مناسب است.

* مثال: یک سری ایمیل داریم. می‌خواهیم آن‌ها را در دو دسته‌ی spam و non-spam دسته‌بندی کنیم. دسته‌ی X نشان دهنده‌ی spam است و O نشان دهنده‌ی non-spam است.



مثلاً خط مدل بدست آمده با logistic regression

برچسبی که در این مثال باید به داده‌ها بدهیم، گسسته است. در linear regression برچسبی که به داده‌ها اختصاص می‌دادیم، پیوسته بود. پس این مسئله، با linear regression قابل حل نیست.

در مسائل linear regression سعی داشتیم تا مدلی روی داده‌ها قرار دهیم و محل قرارگیری آن‌ها را در واقع روی نمودار پیش‌بینی و تخمین کرده باشیم. ولی جنس این سوال متفاوت است. هدف پیدا کردن یک الگو نیست. بلکه هدف جدا سازی دو نوع از داده‌ها و دسته‌بندی آن‌هاست.

تفاوت اصلی linear regression و logistic regression در پیوسته یا گسسته بودن برچسب است. که چون در این مثال، گسسته، باید با logistic آن را حل کنیم.

(6)

الف)

تابع سیگموید : $g(z) = \frac{1}{1+e^{-z}} \longrightarrow \frac{dg(z)}{dz} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}}$

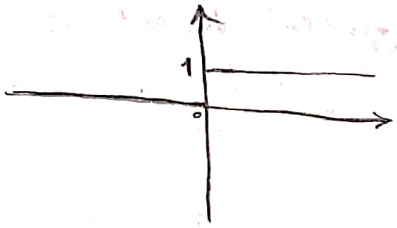
$$\boxed{\frac{dg(z)}{dz} = g(z)\cdot(1-g(z))}$$

---

(6)

ب)

1- Binary step function :



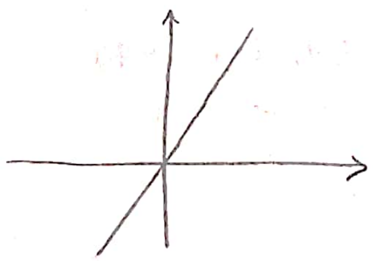
تعریف این تابع به این شکل است :

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

- همانطور که مشخص است، این تابع zero-centered نیست.

- اما ویژگی بدی که دارد، این است که گرادیان آن برابر صفر است؛ چون همانطور که از فرمول آن نیز مشخص است، وابسته به x نیست.

2- linear function :

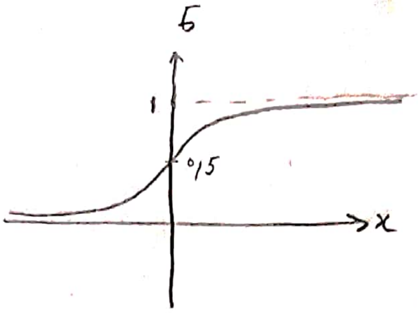


تعریف این تابع : $f(x) = ax$

- این تابع، همانطور که از نمودارش مشخص است، zero-centered نیست.

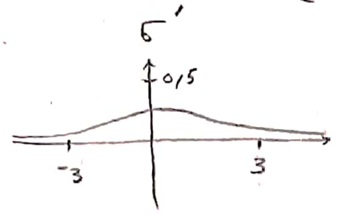
- درباره‌ی گرادیان : درست است که گرادیان آن برابر صفر نیست، اما برابر a است، که عددی کاملاً نامربوط به x است. بنابراین، گرادیان خوبی نیست.

3- Sigmoid :



تعریف تابع: $f(x) = \frac{1}{1+e^{-x}}$

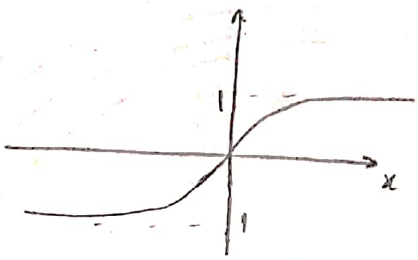
- مشتق این تابع را در بخش قبل بدست آوردیم. این تابع، در همه‌ی نقاط مشتق‌پذیر است. مقادیر گرادیان‌ها، در فاصله‌ی 3- تا 3، تفاوت قابل توجهی باهم دارند. ولی در بقیه‌ی نقاط حالت flat دارد.



این یعنی در مقادیر بیشتر 3 یا کمتر 3-، مشتق به صفر نزدیک می‌شود.

- به علاوه این function، zero-centered نیست.

4- Tanh :

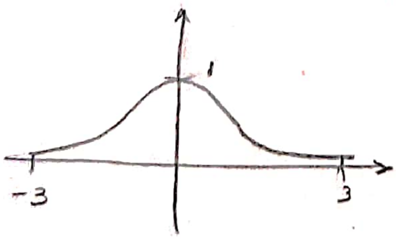


تعریف این تابع:

$\tanh(x) = 2\sigma(2x) - 1$

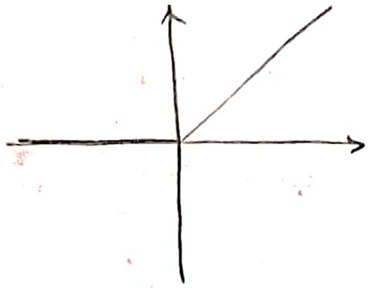
- همانطور که مشخص است، zero-centered است.

- به جز این مورد، بقیه‌ی ویژگی‌های آن مانند σ است. پیوسته است و همه جا مشتق‌پذیر است. ویژگی‌های مشتق آن، مانند σ است.



در نتیجه به خاطر zero-centered بودن، نسبت به σ ترجیح داده می‌شود.

5- ReLU :

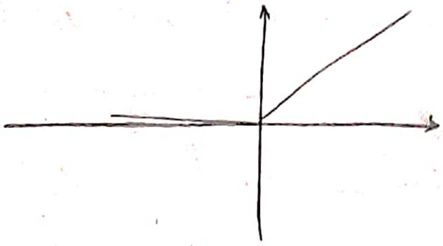


تعریف : $f(x) = \max(0, x)$

- این تابع، zero-centered نیست.

- برای x های منفی، مشتق برابر 0 است

6- Leaky ReLU :



تعریف :

$$f(x) = \begin{cases} 0.01x, & x<0 \\ x, & x \geq 0 \end{cases}$$

- تفاوتی که با ReLU دارد این است که : مشتق برای x های منفی برابر صفر نیست؛ بلکه برابر 0.01 است.

* تعدادی activation function دیگر نیز وجود دارد، که کمتر معروف هستند.

حال بررسی کنیم که کدام یک بهتر است:

- تابع sigmoid، مشتقات آن برای classify کردن بسیار مناسب است.
- تابع sigmoid و tanh گاهاً به دلیل vanishing gradient مورد استفاده قرار نمی‌گیرند.
- ReLu یک تابع عمومی است که برای بیشتر سوال‌ها امروزه کاربرد دارد.
- اگر می‌خواهیم نرون های مرده را نیز به حساب آوریم، Leaky Relu گزینه‌ی خوبی است.

پس واضح است که برای logistic regression، دو تابع Sigmoid و Tanh بهترین گزینه ها هستند. ولی همانطور که گفتیم، Tanh به خاطر zero-centered بودن، به Sigmoid ترجیح داده می‌شود

(7)

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

فعلاً "بدون" ترم regularization در نظر میگیریم.

$$= \frac{1}{m}\sum_{i=1}^{m} T_i$$

$$\Rightarrow \frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum \frac{\partial T_i}{\partial \theta_j} = ?$$

$$T_i = -\left[y^i \log \underbrace{h_\theta(x^i)} + (1-y^i)\log(1-h_\theta(x^i))\right]$$

$$h_\theta(x^i) = \sigma(\underbrace{\theta^T x_j^i}_{z^i}) = \sigma(z^i)$$

$$\Rightarrow \frac{\partial T_i}{\partial \theta_j} = \frac{\partial T_i}{\partial \sigma(z^i)} \times \frac{\partial \sigma(z^i)}{\partial z^i} \times \frac{\partial z^i}{\partial \theta_j}$$

$$\hookrightarrow \frac{\partial T_i}{\partial \sigma(z^i)} = -\left[\frac{y^i}{\sigma(z^i)} - \frac{1-y^i}{1-\sigma(z^i)}\right]$$

$$\hookrightarrow \frac{\partial \sigma(z^i)}{\partial z^i} = \frac{e^{-z^i}}{(1+e^{-z^i})^2} = \sigma(z^i).(1-\sigma(z^i))$$

$$\sigma(z^i) = \frac{1}{1+e^{-z^i}}$$

در جلسه‌های قبل اثبات شد.

$$\hookrightarrow \frac{\partial z^i}{\partial \theta_j} = x_j^i$$

$$\Rightarrow \frac{\partial T_i}{\partial \theta_j} = \left[\sigma(z^i) - y^i\right].x_j^i$$

$$\Rightarrow \frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left[\sigma(z^i) - y^i\right].x_j^i$$

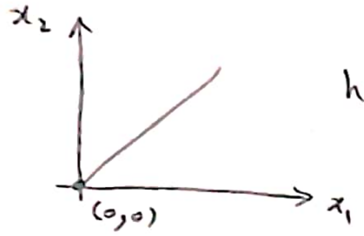
← حال $J(\theta)$ را با وجود ترم regularization در نظر می‌گیریم

$$* \quad \frac{dJ(\theta)}{d\theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left[h_\theta(x^i) - y^i\right].x_j^i \;.2\lambda\theta_j \qquad (j=0,1,2)$$

← داریم : $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

وقتی مقدار λ را خیلی بزرگ در نظر بگیریم، مقدار θ_j ای که در ترم regularization $(\lambda\theta_j^2)$ حضور دارد، برابر صفر می‌شود. حال باید بررسی کنیم که با صفر شدن هرکدام از θ_j ها $(j=0,1,2)$ چه تغییری در خطای آموزش رخ می‌دهد.

+ اگر θ_0 برابر صفر شود : $h_\theta(x) = g(\theta_1 x_1 + \theta_2 x_2)$

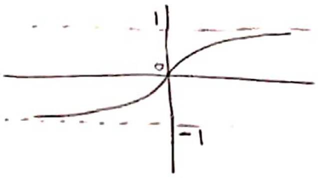


در این صورت، مرز تصمیم حتماً از نقطه‌ی (0,0) عبور می‌کند. که با توجه به این محدودیت، خطای آموزش زیاد خواهد شد.

+ اگر θ_1 یا θ_2 برابر صفر شود، مدل از این هم ساده‌تر و محدودتر می‌شود. (احتمالاً موازی یکی از محور خواهد شد.) بنابراین، خطای آموزش در این دو حالت، از حالت قبلی هم بیشتر می‌شود.

← به طور کلی، ما با داشتن فقط 3 θ، (θ_0 و θ_1، θ_2) مدل خطی داریم و مدل خیلی ساده ای است. وقتی از این θ ها کم کنیم، مدل خیلی ساده می‌شود و خطای آن زیاد می‌شود.

tan-sigmoid :



8

خروجی ها :

$$\begin{cases} a^1 = \text{tanSig}(w^1 p + b^1) = \text{tanSig}(2) \cong 1 \\ a^2 = \text{tanSig}(w^2 a^1 + b^2) = \text{tanSig}(-1) \cong -1 \end{cases}$$

$$\frac{dL}{dw_1} = \frac{dL}{dz_1} \cdot \frac{dz_1}{dw_1} = \delta^1 \cdot p = (a^1 - y_1) p = (1 - y_1) p$$

$$\frac{dL}{dw_2} = \frac{dL}{dz_2} \cdot \frac{dz_2}{dw_2} = \delta^2 \cdot a^1 = (a^2 - y_2) a^1 = -(y_2 + 1)$$

$$\frac{dL}{db_1} = \delta^1 = a^1 - y_1 = 1 - y_1$$

$$\frac{dL}{db_1} = \delta^2 = a^2 - y_2 = -(1 + y_2)$$

$$D^1 = \frac{1}{m} \cdot \frac{dL}{dw^1} + \frac{\lambda}{m} w^1 = \frac{(1-y)p - \lambda}{m}$$

$$D^2 = \frac{1}{m} \frac{dL}{dw^2} + \frac{\lambda}{m} w^2 = \frac{-(y+1) - 2\lambda}{m}$$

$$B^1 = \frac{1}{m} \frac{dL}{db^1} = \frac{1-y}{m}$$

$$B^2 = \frac{1}{m} \frac{dL}{db^2} = \frac{-(1+y)}{m}$$

$$w_1 = w_1 - \alpha D_1 = -1 - \alpha \frac{(1-y)p - \lambda}{m}$$

$$w_2 = w_2 - \alpha D_2 = -2 - \alpha \frac{-(y+1) - 2\lambda}{m}$$

$$b^1 = b^1 - \alpha B^1 = 1 - \alpha \frac{1-y}{m}$$

$$b^2 = b^2 - \alpha B^2 = 1 - \alpha \frac{-(1+y)}{m}$$

(9)

الف)

$$\frac{dL}{dn^2} = a^2 - y = \delta^2 \qquad \frac{dL}{db^2} = \delta^2$$

$$\frac{dL}{dw^2} = \delta^2 \cdot a^1 \qquad \frac{dL}{dw^{2,1}} = \delta^2 \cdot P$$

$$\frac{dL}{dn^1} = a^1 - y = \delta^1 \qquad \frac{dL}{db^1} = \delta^1$$

$$\frac{dL}{dw^1} = \delta^1 \cdot P \qquad \left(\delta^1 = \left[(w^2)^T \delta^2 \right] .* \dot{f}(n^1) \right)$$

الــــ آپدیت کردن پارامترها:

$$D^1 = \frac{1}{m} \cdot \frac{dL}{dw^1} = \frac{1}{m}(a^1 - y) \cdot P = \frac{1}{m}\left[f^1(w^1 p + b^1) - y \right] \cdot P$$

$$D^2 = \frac{1}{m} \cdot \frac{dL}{dw^2} = \frac{1}{m}(a^2 - y) \cdot a^1 = \frac{1}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y \right] \cdot a^1$$

$$D^{2,1} = \frac{1}{m} \cdot \frac{dL}{dw^{2,1}} = \frac{1}{m}(a^2 - y) \cdot P = \frac{1}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y \right] \cdot P$$

$$B^1 = \frac{1}{m} \cdot \frac{dL}{db^1} = \frac{1}{m} \delta^1 = \frac{1}{m}\left[a^1 - y \right] = \frac{1}{m}\left[f^1(w^1 p + b^1) - y \right]$$

$$B^2 = \frac{1}{m} \cdot \frac{dL}{db^2} = \frac{1}{m} \cdot \delta^2 = \frac{1}{m}\left[a^2 - y \right] = \frac{1}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y \right]$$

$$w_1 = w_1 - \alpha D^1 = w^1 - \frac{\alpha}{m}\left[f^1(w^1 p + b^1) - y \right] \cdot P$$

$$w_2 = w_2 - \alpha D^2 = w^2 - \frac{\alpha}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y \right] \cdot a^1$$

$$w^{2,1} = w^{2,1} - \alpha D^{2,1} = w^{2,1} - \frac{\alpha}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y \right] \cdot P$$

$$b^1 = b^1 - \alpha B^1 = b^1 - \frac{\alpha}{m}\left[f^1(w^1 p + b^1) - y \right]$$

$$b^2 = b^2 - \alpha B^2 = b^2 - \frac{\alpha}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y \right]$$

Scanned by CamScanner

ب) اضافه کردن ضریب منظم سازی بر روی D^l اثر خواهد گذاشت :

$$D^1 = \frac{1}{m}\frac{dL}{dw^1} + \frac{\lambda}{m} w^1 = \frac{1}{m}\left[f^1(w^1 p + b^1) - y\right] \cdot p + \frac{\lambda}{m} w^1$$

$$D^2 = \frac{1}{m} \cdot \frac{dL}{dw^2} + \frac{\lambda}{m} w^2 = \frac{1}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y\right] \cdot a^1 + \frac{\lambda}{m} w^2$$

$$D^{2,1} = \frac{1}{m} \cdot \frac{dL}{dw^{2,1}} + \frac{\lambda}{m} w^{2,1} = \frac{1}{m}\left[f^2(w^2 a^1 + w^{2,1} p + b^2) - y\right] \cdot p + \frac{\lambda}{m} w^{2,1}$$

$$w_1 := w_1 - \alpha D^1 = w^1 - \frac{\alpha}{m}\left[\left(f^1(w^1 p + b^1) - y\right) \cdot p + \lambda w^1\right]$$

$$w_2 = w_2 - \alpha D^2 = w_2 - \frac{\alpha}{m}\left[\left(f^2(w^2 a^1 + w^{2,1} p + b^2) - y\right) \cdot a^1 + \lambda w^2\right]$$

$$w^{2,1} = w^{2,1} - \alpha D^{2,1} = w^{2,1} - \frac{\alpha}{m}\left[\left(f^2(w^2 a^1 + w^{2,1} + b2) - y\right) \cdot p + \lambda w^{2,1}\right]$$

$\Leftarrow$ با توجه به اینکه D^l تغییر کند، w^l نیز نسبت به حالت قبل عوض می‌شود

و سپس B^l عوض نمی‌شود $\Longleftarrow$ b^l مانند حالت قبلی خواهد بود.