
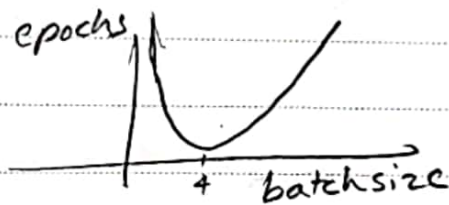


Step

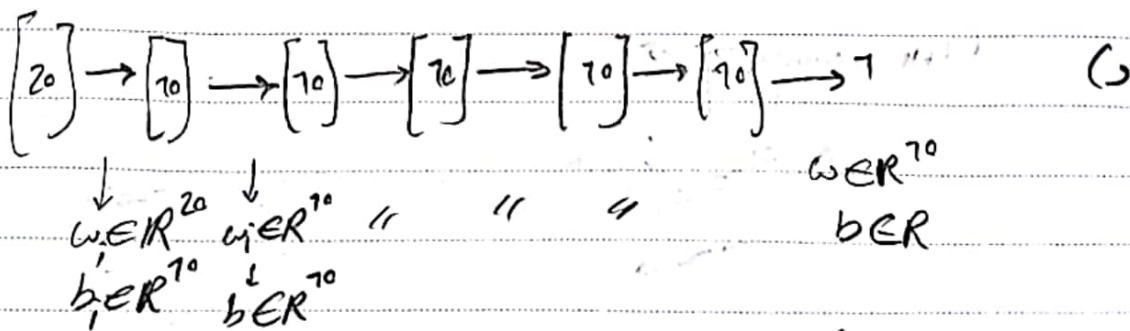


batchsize

تعداد step کا
بہر حال تھوڑی
خصوصی صورت
(قرائے ep rate پر مشتمل)



ج) زیرا تابع σ با مقدار σ همبستگی دارد و این مقدار σ را σ می‌نامند.
 ممکن است بگوییم که این تابع σ را σ می‌نامند و این مقدار σ را σ می‌نامند.
 که σ را σ می‌نامند و این مقدار σ را σ می‌نامند.



$$20 \times 10 + 10 \times 10 + 10 \times 10 + 10 \times 10 + 1 \times 10 + 1 = 661$$

Softmax و Logistic

Softmax و Logistic

$\hat{y}_2 = 1 - \hat{y}_1$ و $\hat{y}_1, \hat{y}_2 \leftarrow \text{softmax}$

$\hat{y} \leftarrow \text{logistic}$

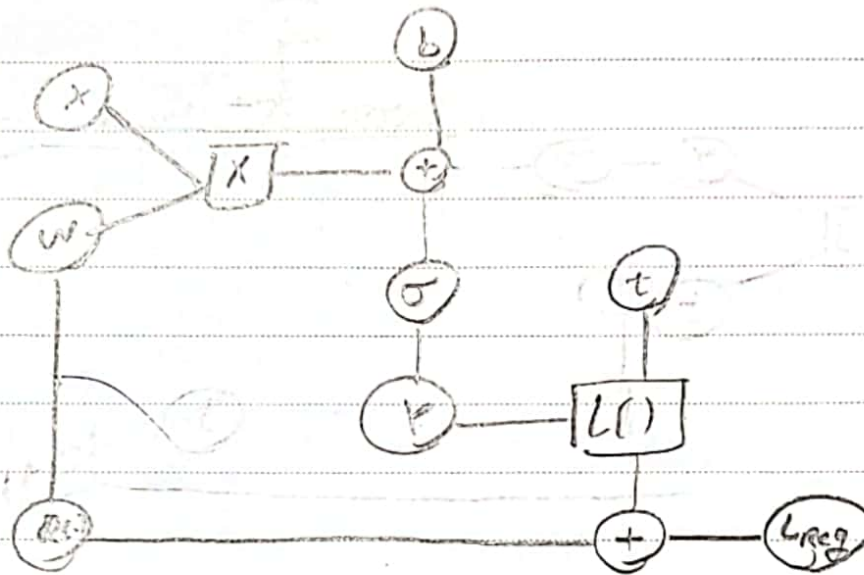
$$\hat{y}_1 = \frac{e^{w_1 x + b_1}}{e^{w_1 x + b_1} + e^{w_2 x + b_2}} = \frac{1}{1 + e^{w_2 x + b_2 - w_1 x - b_1}} = \hat{y}$$

softmax logistic

$$\Rightarrow \begin{cases} w_2 - w_1 = w_c \\ b_2 - b_1 = b_c \end{cases}$$

۲) اگرچه از classifier ها که می‌توانیم با بزرگ متفاوت این کار را انجام دهیم و بر روی داده‌ها
دیتا پوزیشن شده آموخته می‌دهیم. انجام یکی برای توپا آنها با تغییر خروجی‌ها و آنها
به ندرت همکاره دارند اینها انجام می‌دهیم و در صورتی که خطاها کم می‌آید از دیگره منتقل است و آنرا جمع
آنها از بین رفتن این خطاهاست.

ب) این سیکل پوزیشن شامل نرمالیزاسیون و دفعه کردن دیتا با بزرگ متفاوت برای خودیاست
می‌شود، آنها از *elastic distortion*, *scaling*, *rotation* استفاده کرده اند.
این موارد نیز *data augmentation* هم هست که از *overfit* کردن مدل ها جلوگیری می‌کند.
این کار هو مدل نسبت به تغییرات خاص در دیتا می‌شود که از بقیه منتقل است پس هو مدل
در یک سری خصوصیات خاص خطا و ضعف در تشخیص خواهد داشت و اینگونه می‌تواند نسبت
خطاها از عدم منتقل اند و ترکیب جواب مدل ها باعث حذف خطاها شود.



$$\frac{dL_{reg}}{dx} = 1$$

$$\frac{dL_{reg}}{da} = 1$$

$$\frac{dR}{dw} = w$$

$$x, w, b$$

$$y, t$$

$$\frac{dL}{dy} = y - t$$

$$\frac{dL}{dt} = t - y$$

$$\frac{dy}{dz} = y(1-y)$$

$$\frac{dz}{dw} = w$$

$$\frac{dz}{da} = x$$

$$\frac{dz}{db} = 1$$

$$\rightarrow \frac{dL_{reg}}{dy} = y - t, \quad \frac{dL_{reg}}{dz} = (y - t) y (1 - y) \Rightarrow \frac{dL_{reg}}{da} = w(y - t) y (1 - y)$$

$$\frac{dL_{reg}}{db} = (y - t) y (1 - y)$$

$$\Rightarrow \frac{dL_{reg}}{dw} = x(y - t) y (1 - y)$$

ب) فرض کنید در Activation های tanh و sigmoid، اگر شبکه کوچک بودن را رعایت نکنیم
از همان ابتدا در شبکه کوچک بودن را رعایت کنیم. اگر این کار را نکنیم ممکن است در شبکه
تدریس می شود و مدل به جواب خوبی نرسد.

$$x = 1, \quad t = 0.5, \quad \omega = 0.25, \quad b = 10 \quad (2)$$

$$\hookrightarrow y = \sigma(0,25) \approx 0,562 \quad L = \frac{1}{2}(0,562 - 0,5)^2 = 0,01922$$

$$R = \frac{1}{2} (1.25)^2 = 0.3125 \quad L_{reg} = 0.33772$$

$$\frac{d_{\text{reg}}}{dw} = u(y-t)y(1-y) = 0,0752 \quad \frac{d_{\text{reg}}}{d\beta} = (y-t)y(1-y) = 0,0752$$

$$\omega' = \omega - 0,1 \cdot 0,752 \quad b' = b - 0,1 \cdot 0,752 = -0,00152$$

$$= 0,24848 \Rightarrow y' = \sigma(0,24848 - 0,1752) = 0,561$$

(ع) ۵ $g_t = \nabla_{\theta} f_t(\theta_{t-1})$ کے لیے متوالیہ f_t کے لیے θ_t پر

قدم به دست می آوریم

دو صفحه B_2 و B_3 را از زیر 1. ایما و حرکت از آن تعریف می کنیم (هر دو B_2 و حرکت را به ما اثر القدریم که است -

[illegible]

این همان مدل راجه اندروید است و از آنجا که

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

اینکه m_1 باقیمانده بقیه معجزه زمانه می کند. این کار برای حذف bias در قیمت خاص اول است.

$$\hat{m}_1 \leftarrow \frac{m_1 t}{1 - \beta_1 t}$$

مثلاً در قدم اول که $\beta_1 = 0$ می باشد. مخرج کسر نسبتاً برابر با ضریب $(1 - \beta)$ است که در قدم اول برابر با ۱ می باشد. پس اثر آن را در قدم اول اول ضمیمه می کنند

$$\hat{v}_t \leftarrow \frac{v_t}{1 - \beta_t}$$

$$\theta_t = \theta_{t-1} - \frac{\alpha \hat{m}_t}{\sqrt{r_t} + \epsilon}$$

حل با تقسیم \hat{m}_t بر $\sqrt{r_t}$ جهت برآورد
اندازه را نسبت آورده و از ϵ برای
از بین بردن مشکل *division by zero* استفاده می‌کنیم

و ضریب α (معامل λ) را ثابت می‌کنیم

ب) چون مورد در سمت قبل هم نفهم، معادله β ها را نیز بر موهنه تا اثر
ناهمان یک تغییر هم در درجه‌ها دیده شد، θ بیان احتمالی زیاده باشد و *overshoot*

حاکمتر شود. حل در قدم اول (قدم های اول) $m_0 = 0$ را داریم پس در قدم های اول
تا m_t به یک مقدار هم نرسد تا اثر در پس خالی حل می‌شود به اصطلاح m_t به صفر با پس
دارد. برای ضرایب اثر \hat{m}_t را با تقسیم m_t بر $(1 - \beta_t^t)$ می‌سازیم، که این

تعداد در قدم اول دقیقاً معادل تقسیم m_t بر $1 - \beta_t$ می‌باشد که تأثیر بایس صفر را به راحتی
از بین می‌برد. با نزدیک شدن t به $1 - \beta_t^t$ به 1 می‌رود و تأثیر این قدم از بین می‌رود.
این کار باعث حذف بایس صفر m_t می‌شود و \hat{m}_t در قدم های اول کمتر میل می‌کند.

$$\nabla_{\omega} (\omega^T H \omega) = 2H\omega$$

۱.۵ - ۸

$$\Rightarrow \omega_t = \omega_{t-1} - \epsilon (2H\omega_{t-1})$$

$$= \omega_{t-1} (1 - 2\epsilon H) = \omega_{t-1} (1 - 2\epsilon Q \Lambda Q^T)$$

$$= \omega_{t-1} (QQ^T - Q 2\epsilon \Lambda Q^T) = \omega_{t-1} Q (Q^T - 2\epsilon \Lambda Q^T)$$

$$= \omega_{t-1} Q (1 - 2\epsilon \Lambda) Q^T$$

$$\omega_1 = \omega_0 \underbrace{Q^T H Q}_{H'} Q$$

$$\omega_2 = \omega_0 H' H', \omega_3 = \omega_0 H'^3 \Rightarrow \omega_t = \omega_0 Q (1 - 2\epsilon \Lambda)^3 Q^T$$

(۲) در صورتی که دایه‌ها ماتریس $\lim_{t \rightarrow \infty} (1 - 2\epsilon \Lambda)^t$ و ϵ finite باشد

اگر $\forall \lambda_i \quad |1 - 2\epsilon \lambda_i| < 1$: Algorithm converges. (همواره همگرا می‌شود)

از آنجایی که H PD باشد، λ_i ها نیز از صفر بزرگتر هستند.

$$\forall \lambda_i \quad |1 - 2\epsilon \lambda_i| < 1 \Rightarrow 1 - 2\epsilon \lambda_i < 1 \Rightarrow \epsilon \lambda_i < 1$$

$$\lambda_i > 0 \quad \Rightarrow \quad \forall \lambda_i \quad \epsilon < \frac{1}{\lambda_i}$$

به دلیل ϵ بزرگتر از صفر است، $\forall \lambda_i \quad \epsilon < \frac{1}{\lambda_i}$

و از آنجایی که λ_1 بزرگترین مقدار ویژه است: $\epsilon < \frac{1}{\lambda_1}$

$$\omega_t = \omega_{t-1} - \epsilon \cdot H(\omega^{TH}\omega)^{-1} \cdot \nabla_{\omega}(\omega^{TH}\omega) \quad (7)$$

$$= \omega_{t-1} - \epsilon (2H)^{-1} \cdot 2H\omega_{t-1}$$

$$= \omega_{t-1} - \epsilon \omega_{t-1}$$

(8) - نصف بیشتر موارد و محاسبه به سبقت

- ماتریس deep convex نیست و ماتریس Hessian H PD نیست

آن مادر آن نقاط استفاده کرد

SGD

- رویه نیوتون به سبقت Batch محاسبه دارد و محاسبه Hessian نیز بسیار آسان

جواب داد

$$\begin{aligned}
 & E\left[\frac{1}{2}\left(y_d - \sum_{k=1}^n (\omega_k + \delta_k) a_k\right)^2\right] \quad (5) \\
 &= \frac{1}{2} \left(y_d^2 + E\left[\sum_{k=1}^n (\omega_k + \delta_k) a_k\right]^2 - 2y_d E\left[\sum_{k=1}^n (\omega_k + \delta_k) a_k\right] \right) \\
 &= \frac{1}{2} \left(y_d^2 - 2y_d \sum_{k=1}^n \omega_k a_k + E\left[\left(\sum_{k=1}^n \omega_k a_k + \sum_{k=1}^n \delta_k a_k\right)^2\right] \right) \\
 &= \frac{1}{2} \left(y_d^2 - 2y_d \sum_{k=1}^n \omega_k a_k + \left(\sum_{k=1}^n \omega_k a_k\right)^2 + E\left[\left(\sum_{k=1}^n \delta_k a_k\right)^2\right] \right) \\
 &= \frac{1}{2} \left(\left(y_d - \sum_{k=1}^n \omega_k a_k\right)^2 + \sum_{k=1}^n \alpha \omega_k^2 a_k^2 \right) \\
 &\Rightarrow E\left[\frac{\partial J}{\partial \omega_i}\right] = \frac{1}{2} \left(2\alpha \omega_i a_i^2 + \frac{\partial}{\partial \omega_i} \left(y_d - \sum_{k=1}^n \omega_k a_k\right)^2 \right) \\
 &= \alpha \omega_i a_i^2 + \frac{1}{2} \left(-2 a_i y_d + 2 a_i \left(\sum_{k=1}^n \omega_k a_k - \omega_i a_i\right) + 2 \omega_i a_i^2 \right) \\
 &= \alpha \omega_i a_i^2 - a_i y_d + a_i \sum_{k=1}^n \omega_k a_k = \alpha \omega_i a_i^2 + a_i \left(\sum_{k=1}^n \omega_k a_k - y_d\right)
 \end{aligned}$$

ب) هدف دین می شود، تأثیر این نویز را در دسترس اضافه می کنیم

فرم $\alpha \omega_i a_i^2$ به استقامت که در رگرسیون داریم:

که در اینجا می توان نوشت $L_{reg} = L + \lambda \omega_i$

که $\lambda = \alpha a_i^2$ سکه و نویز رگرسیون آنرا اضافه است.