Deep learning تمرین نظری سری اول

402212269 امیرمحمد عذری

سوال ①

$$\begin{cases} x: & n\times 1 \\ y: & m\times 1 \end{cases} \qquad \begin{cases} y = Ax \\ A: \ m\times n \end{cases}$$

1) $$y_i = \sum_{K=1}^{n} a_{iK}\, x_K \overset{j\in[1,n]}{\Longrightarrow} \frac{\partial y_i}{\partial x_j} = a_{ij} \Rightarrow \frac{\partial y}{\partial x} = A_{m\times n}$$

$$\frac{\partial y}{\partial x_j} = \begin{bmatrix} \frac{\partial y_1}{\partial x_j} \\ \frac{\partial y_2}{\partial x_j} \\ \vdots \\ \frac{\partial y_m}{\partial x_j} \end{bmatrix} \Longrightarrow \frac{\partial y}{\partial x} = A = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & & \\ \frac{\partial y_m}{\partial x_1} & & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

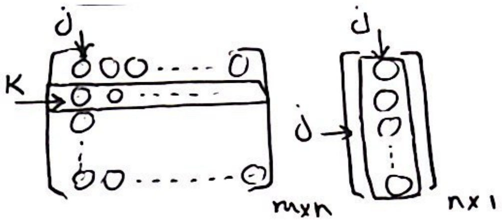
مستقل

2) $$\frac{\partial y}{\partial z} = \frac{\partial (Ax)}{\partial z} = \cancel{\frac{\partial A}{\partial z}}^{\,0}\, x + A\frac{\partial x}{\partial z} = A\frac{\partial x}{\partial z}$$

3) $$\alpha = y^T A x \Rightarrow \alpha = \sum_{i=1}^{m} y_i \sum_{j=1}^{n} a_{ij}\, x_j = \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j$$

$$\frac{\partial \alpha}{\partial x_K} = \sum_{i=1}^{m} y_i\, a_{iK} = \sum_{i=1}^{m} \begin{bmatrix} \boxed{\circ} \\ \circ \\ \vdots \\ \circ \end{bmatrix}_{m\times 1} \begin{bmatrix} \boxed{\circ\ \circ\ \circ \cdots \circ} \\ \circ\ \circ\ \circ \\ \vdots \\ \circ \cdots \circ \end{bmatrix}_{m\times n} \Rightarrow$$

$$\frac{\partial \alpha}{\partial x} = y^T_{1\times m} A_{m\times n} \qquad [1\times n]$$

3) $$\frac{\partial \alpha}{\partial y_K} = \sum_{j=1}^{n} a_{Kj}\, x_j = \sum_{j=1}^{n}$$ 

$$\frac{\partial \alpha}{\partial y} = \underset{1\times n}{x^T}\ \underset{n\times m}{A^T} \qquad [1\times m]$$

4) $$\alpha = \underset{1\times n}{y^T}\ \underset{n\times 1}{x} \qquad [1\times 1] \quad \text{scalar}$$

$$\frac{\partial \alpha}{\partial z} = \frac{\partial \alpha}{\partial x}\frac{\partial x}{\partial z} + \frac{\partial \alpha}{\partial y}\frac{\partial y}{\partial z} \xrightarrow{\alpha = y^T x}$$

$$\circledast\ \frac{\partial \alpha}{\partial x} = y^T, \quad \frac{\partial \alpha}{\partial y} = x^T \qquad \text{"taken from previous part"}$$

$$\Rightarrow \quad \frac{\partial \alpha}{\partial z} = y^T \frac{\partial x}{\partial z} + x^T \frac{\partial y}{\partial z}$$

5) if A is non-Singular then $\det(A) \neq 0$ and A^{-1} exists.

So: $$AA^{-1} = I \Rightarrow \frac{\partial (AA^{-1})}{\partial \alpha} = \frac{\partial I}{\partial \alpha} = 0$$

$$\frac{\partial (AA^{-1})}{\partial \alpha} = \frac{\partial A}{\partial \alpha} A^{-1} + A \frac{\partial A^{-1}}{\partial \alpha} = 0 \quad \Rightarrow$$

$$A \frac{\partial A^{-1}}{\partial \alpha} = -\frac{\partial A}{\partial \alpha} A^{-1} \xrightarrow{A^{-1}\times} I \frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha} A^{-1}$$

$$\Rightarrow \frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

سوال ②

$$y = \psi(u, v, z)$$

$$\nabla\psi = \left(\frac{\partial\psi}{\partial u}, \frac{\partial\psi}{\partial v}, \frac{\partial\psi}{\partial z}\right)_{3\times1}$$

$$J(\nabla\psi) = \begin{bmatrix} \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial u}\right) & \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial v}\right) & \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial z}\right) \\ \frac{\partial}{\partial v}\left(\frac{\partial\psi}{\partial u}\right) & \frac{\partial}{\partial v}\left(\frac{\partial\psi}{\partial v}\right) & \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial z}\right) \\ \frac{\partial}{\partial z}\left(\frac{\partial\psi}{\partial u}\right) & \frac{\partial}{\partial z}\left(\frac{\partial\psi}{\partial v}\right) & \frac{\partial}{\partial z}\left(\frac{\partial\psi}{\partial z}\right) \end{bmatrix}$$

$$\Rightarrow J(\nabla\psi) = \begin{bmatrix} \frac{\partial^2\psi}{\partial u^2} & \frac{\partial^2\psi}{\partial u\partial v} & \frac{\partial^2\psi}{\partial u\partial z} \\ \frac{\partial^2\psi}{\partial v\partial u} & \frac{\partial^2\psi}{\partial v^2} & \frac{\partial^2\psi}{\partial v\partial z} \\ \frac{\partial^2\psi}{\partial z\partial u} & \frac{\partial^2\psi}{\partial z\partial v} & \frac{\partial^2\psi}{\partial z^2} \end{bmatrix}$$

$$H = \nabla^2\psi = \begin{bmatrix} \frac{\partial^2\psi}{\partial u^2} & \frac{\partial^2\psi}{\partial u\partial v} & \frac{\partial^2\psi}{\partial u\partial z} \\ \frac{\partial^2\psi}{\partial v\partial u} & \frac{\partial^2\psi}{\partial v^2} & \frac{\partial^2\psi}{\partial v\partial z} \\ \frac{\partial^2\psi}{\partial z\partial u} & \frac{\partial^2\psi}{\partial z\partial v} & \frac{\partial^2\psi}{\partial z^2} \end{bmatrix}$$

So Hessian of ψ function equals to the Jacobian of its gradient.

سوال ③

برخی از انواع داده ها دارای تقارن ذاتی هستند و این تقارن به اشکال مختلف می تواند وجود داشته باشد.

برای مثال تقارن فضایی در تصاویر صورت وجود دارد و یک چهره ممکن است در امتداد محور عمودی تقارن داشته باشد. یعنی یک طرف صورت، طرف مقابل را می تواند منعکس کند.

همچنین داده های سری زمانی می توانند دارای تقارن زمانی باشند؛ یعنی الگوهایی داشته باشند که در طول زمان تکرار شوند. (بصورت ماهیانه، فصلی و ...)

این تقارن به صورت های مختلف مثل افزایش داده ها، مهندسی feature، انتخاب بهتر معماری شبکه و انتخاب regularization بهتر باعث بهبود عملکرد تسک classification شود.

افزایش داده: وجود تقارن به ما امکان تولید داده های جدید با اعمال تبدیلات بازتاب و چرخش می دهد.

مهندسی feature: وجود تقارن کمک می کند که در مهندسی ویژگی، ویژگی های مبتنی بر تقارن را برای ثبت ویژگی های داده در نظر گرفت.

معماری مدل: می توان از شبکه های عصبی ای که دانش مربوط به تقارن را می تواند بدست بیاورد مثل شبکه های CNN استفاده کرد. چون این شبکه ها، امکان بهره گیری از تقارن های translational را دارند.

regularization: از تکنیک های منظم سازی که وزن های شبکه را تشویق به یادگیری فیچرهای متقارن در داده می کنند، می توان استفاده کرد. ویژگی weight-sharing در CNN باعث یادگیری تقارن هایی در پارامترهایش می شود.

ادامه سوال ③

چون ترم regularization به تابع هزینه اضافه می شود و شبکه به دنبال کاهش این ترم خواهد بود. سپس باید این ترم regularization را طوری انتخاب کنیم تا وزن هایمان اختلاف زیادی نداشته باشند.

$$X = [x_1 \quad x_2] \quad (1,2)$$

$$W = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad (2,1)$$

$$R(w) = w^T S w$$

اگر بخواهیم w_1 و w_2 نزدیک هم باشند و نامتقارن نشوند، باید S را طوری انتخاب کنیم تا $(w_1 - w_2)^2$ را در ترم رگولاریزیشن ایجاد کنیم.

$$R(w) = [w_1 \quad w_2] \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = a w_1^2 + b w_1 w_2 + c w_1 w_2 + d w_2^2$$

if $\quad R(w) = (w_1 - w_2)^2 = w_1^2 - 2 w_1 w_2 + w_2^2$

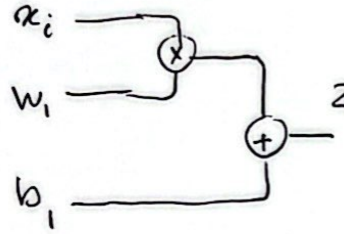
then $\quad a = 1, \; b = -1, \; c = -1, \; d = 1$

$$S = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

سوال 4

$x_i : [D_x , 1]$

$Z_1 : [D_{a_1} , 1]$



$$Z_1 = W_1 x^{(i)} + b_1$$

$$[D_{a_1}, 1] = [D_{a_1}, D_x] \cdot [D_x, 1] + [D_{a_1}, 1]$$

1) $W_1 : [D_{a_1} , D_x]$, $b_1 : [D_{a_1} , 1]$

$a_1 , z_1 : [D_{a_1} , 1]$ $\quad Z_2 = W_2 a_1 + b_2$

$$[D_{a_2}, 1] = [D_{a_2}, D_{a_1}] \cdot [D_{a_1}, 1] + [D_{a_2}, 1]$$

$W_2 : [D_{a_2} , D_{a_1}]$, $b_2 : [D_{a_2} , 1]$

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

3) $$\frac{\partial \hat{y}^{(i)}}{\partial Z_2} = \sigma(Z_2)\left(1 - \sigma(Z_2)\right) = \delta_2$$

4) $$\frac{\partial Z_2}{\partial a_1} = W_2 = \delta_3$$

5) $$\frac{\partial a_1}{\partial z_1} = \begin{cases} 1 & z_1 \geqslant 0 \\ 0 & z_1 < 0 \end{cases} = \delta_4$$

6) $$\frac{\partial z_1}{\partial W_1} = x^T = \delta_5$$

سوال ④

7) $$\frac{\partial J}{\partial W_1} = \frac{\partial J}{\partial y^{(i)}} \times \frac{\partial y^{(i)}}{\partial z_2} \times \frac{\partial z_2}{\partial a_1} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z_1}{\partial W_1}$$

$$\Rightarrow \frac{\partial J}{\partial W_1} = \delta_1 \times \delta_2 \times \delta_3 \times \delta_4 \times \delta_5 \Rightarrow$$

$$\frac{\partial J}{\partial W_1} = \frac{-1}{m} \left(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1-y^{(i)}}{1-\hat{y}^{(i)}} \right) \sigma(z_2)\left(1-\sigma(z_2)\right) W_2 \, x^T$$

for $z_1 \geqslant 0$

$$\frac{\partial J}{\partial W_1} = 0$$ for $z_1 < 0$

سوال (5)

1) تابع beale

$$f(x) = (1.5 - x_1 + x_1 x_2)^2 + (2.25 - x_1 + x_1 x_2^2)^2 + (2.625 - x_1 + x_1 x_2^3)^2$$

تابع beale غیرمحدب است زیرا ماتریس Hessian آن در همه جا positive semi-definite نیست. همچنین بطور خاص‌تر در نقطه (0,1) دارای saddle point است و در این نقطه دارای مقادیر ویژه مثبت و منفی در ماتریس Hessian است.

این موضوع نشان می‌دهد تابع در برخی جهات به سمت بالا منحنی می‌شود و در برخی جهات به سمت پایین. پس بصورت بدیهی Convex نمی‌باشد.

به دلایل مختلفی Convexity در optimization اهمیت دارد:

- توابع محدب مشکل local minima را ندارند و هر local minima در توابع محدب یک global minima نیز هست.
- بهینه‌سازی توابع محدب ساده‌تر از توابع غیرمحدب است و می‌توان راحت‌تر global minima را یافت.
- پایداری و robustness توابع محدب نسبت به انحراف کوچک بیشتر از توابع غیرمحدب است.

$$\nabla f = \begin{bmatrix} 2xy^6 + 2xy^4 + 5.25y^3 - 4xy^3 + 4.5y^2 - 2xy^2 + 3y - 4xy + 6x - 12.75 \\ 6x^2y^5 + 4x^2y^3 - 6x^2y^2 - 2x^2y - 2x^2 + 15.75xy^2 + 9xy + 3x \end{bmatrix}$$

$$\nabla f(0,1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

$$(x^{(K+1)}, y^{(K+1)}) = (x^{(K)}, y^{(K)}) - \alpha \nabla f(x^{(K)}, y^{(K)})$$

$$= (0,1) - \alpha(0,0) = (0,1)$$

(2

Q is positive semi definite

$$h(x) = \frac{1}{2} x^T Q x + x^T c + b$$

فرض کنیم تابع $h(x)$ بصورت زیر باشد :

$$h(x) = \frac{1}{2} x^T Q x \qquad Q \text{ is PSD}$$

اعمال گرادیان کاهشی :

$$x_K = x_{K-1} - \alpha \nabla h(x_{K-1})$$

$$\Rightarrow x_k = x_{K-1} - \alpha Q x_{K-1} = (I - \alpha Q) x_{K-1}$$

اگر بصورت بازگشتی ادامه دهیم :

$$x_K = (I - \alpha Q)^K x_0$$

اگر ماتریس Q را بصورت تجزیه eigen value هایش بنویسیم :

$$Q = V \Lambda V^T \qquad \begin{cases} V: & \text{eigen vectors} \\ \Lambda: & \text{eigen values} \end{cases}$$

$$(I - \alpha Q)^K x_0 = \left(I - \alpha \left(V \Lambda V^T\right)\right)^K x_0$$

$$= \left(I - \alpha V \Lambda V^T\right)^K x_0 = \left(V (I - \alpha \Lambda) V^T\right)^K x_0$$

$$x_K = V (I - \alpha \Lambda)^K V^T x_0$$

پس الگوریتم بهینه‌سازی گرادیان کاهشی برای $h(x)$ فرضی در جهت بردارهای ویژه ماتریس Q است .

چون الگوریتم بهینه سازی گرادیان کاهشی نسبت به شیفت تغییرناپذیر است.

پس می‌توانیم تابع $h(x)$ را به تابع درجه 2 با مرکز غیر صفر نیز تعمیم دهیم.

$$h(x) = \frac{1}{2} x^T Q x + x^T c + b$$

---

3) الگوریتم Adam

$$m^{(K)} = \beta_1 m^{(K-1)} + (1-\beta_1)(\partial_w E)^{(K)}$$

$$v^{(K)} = \beta_2 v^{(K-1)} + (1-\beta_2)(\partial_w E)^{2(K)}$$

$$\hat{m}^{(K)} = \frac{\hat{m}^{(K)}}{1-\beta_1^K} , \quad \hat{v}^{(K)} = \frac{\hat{v}^{(K)}}{1-\beta_2^K}$$

$$w^{(K+1)} = w^{(K)} - \frac{\eta}{\sqrt{\hat{v}^{(K)}+\varepsilon}} \hat{m}^{(K)}$$

$$\circledast \quad \partial_w E^{(K)} = g_K$$

$$v_K = \beta_2 v_{K-1} + (1-\beta_2) g_K^2 \implies v_K = (1-\beta_2) \sum_{j=1}^{K} \beta_2^{K-j} g_j^2$$

از 2-norm به ∞-norm تغییر می‌دهیم:

$$v_K = (1-\beta_2^P) \sum_{j=1}^{K} \beta_2^{P(K-j)} |g_j|^P$$

$$\lim_{P\to\infty} (v_K)^{1/P} = \lim_{P\to\infty} \left((1-\beta_2^P) \sum_{j=1}^{K} \beta_2^{P(K-j)} |g_j|^P \right)^{1/P} =$$

$$\lim_{P\to\infty} \left(\sum_{j=1}^{K} \left(\beta_2^{(K-j)} \cdot |g_j| \right)^P \right)^{1/P} = \max_j \left(\beta_2^{K-j} |g_j| \right)$$

اگر ∞-norm گرادیان را در Step K برابر u_K بنامیم:

$$u^{(K)} = \max\left(\beta_2 \cdot u^{(K-1)}, |g_K|\right)$$

پس الگوریتم جدید بشکل زیر خواهد بود.

$$m^{(K)} = \beta_1 \, m^{(K-1)} + (1-\beta_1)\left(\partial_w E\right)^{(K)}$$

$$u^{(K)} = \max\left(\beta_2 \, u^{(K-1)}, \left|\left(\partial_w E\right)^{K}\right|\right)$$

$$\theta^{(K)} = \theta^{(K-1)} - \frac{\eta}{(1-\beta_1^{K})\, u^{(K)}} \, m^{(K)}$$

برگرفته از مقاله Adam: a method for stochastic optimization

- الگوریتم حاصل شده (Adamax) در زمان هایی که گرادیان ها بزرگ هستند stable تر است یا زمانی که مشخصه های داده در طول زمان تغییر پذیرند و نیاز است با توجه به آن learning rate را تغییر دهیم، مناسب است.

الگوریتم Adam بصورت عمومی optimizer مناسبی است. علی الخصوص زمانی که دیتا ست بزرگ و پارامترهای زیادی داریم. این الگوریتم تغییر ناپذیر نسبت به rescale شدن گرادیان ها است و این ویژگی زمانی که گرادیان ها بین مسئله های متفاوت بصورت قابل توجه تغییر می کنند یا حتی در بخش های مختلف یک مسئله متغیر هستند، کمک می کند.