Deep Neutral Networks

Homework 0

Yuanteng Chen

1. Gradient Descent Doesn't go nuts with ill-conditions.

Show that for $t>0$, $\|w_t\|_2 \leq \|w_{t-1}\|_2 + \eta\alpha\|y\|_2$:

Sollution;

$$w_t = w_{t-1} - \eta(F^T(Fw_{t-1} - y))$$

according to the assumption: learning rate η is small enough that gradient descent cannot possibly diverse and the Hint $(E - \eta F^T F)$.

first I make an assumption that singular value of $(E - \eta F^T F)$ is less than a specific number M

then I need to convert the original to a formula containing $(E - \eta F^T F)$ so that singular value of $(E - \eta F^T F) < M$ can be used.

$$\|w_t\|_2 = \|w_{t-1} - \eta(F^T(Fw_{t-1} - y))\|_2$$

$= \| W_{t-1} - \eta F^T F W_{t-1} + \eta F^T y \|_2$

$= \| (E - \eta F^T F) W_{t-1} + \eta F^T y \|$

(We know that $\|A+B\|_2 \leq \|A\|_2 + \|B\|_2$)

$\leq \| (E - \eta F^T F) W_{t-1} \|_2 + \| \eta F^T y \|_2$

as the target is $\|W_{t-1}\|_2 + \eta \alpha \|y\|_2$

the latter one is obvious: $\|\eta F^T y\|_2 \leq \eta \alpha \|y\|_2$

But if we want to prove $\|(E - \eta F^T F) W_{t-1}\|_2 \leq \|W_{t-1}\|_2$,

we must prove that specific number M is 1

that is singular value of $(E - \eta F^T F)$ is less that 1

(but I don't know how to prove it)

2. Regularization from the Augmentation Perspective

Show that the ordinary least squares problem

$\underset{W}{\operatorname{argmin}} \|\vec{y} - \hat{X} \vec{w}\|_2^2$ has the same solution as

$$\vec{w} = (X^T X + \Sigma^{-1})^{-1} X^T y$$

Solution: in Tikhonov regularization.

$$\underset{W}{\operatorname{argmin}} \|\vec{y} - \hat{X} \vec{w}\|^2 + W^T \Sigma^{-1} W$$

the MAP (Maximum A Posteriori) of w is:

$$w = (X^TX + \Sigma^{-1})^{-1}X^Ty$$

in the ordinary least squares problem

$$\arg\min_{w} \|\vec{y} - \hat{X}w\|_2^2$$

OLS is a commonly used method for fitting linear models and estimating model parameters:

$$\hat{y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$ ($\hat{\beta}$ is parameter estimate)

when $\hat{X} = \begin{bmatrix} X \\ \Gamma \end{bmatrix} \in R^{(n+d)\times d}$ and $\hat{y} = \begin{bmatrix} y \\ 0_d \end{bmatrix} \in R^{n+d}$

$$
\begin{aligned}
\hat{\beta} &= (X^TX)^{-1}X^Ty \\
&= \left([X^T, \Gamma^T]\cdot\begin{bmatrix} X \\ \Gamma \end{bmatrix}\right)^{-1}[X^T, \Gamma^T]\begin{bmatrix} y \\ 0_d \end{bmatrix} \\
&= (X^TX + \Gamma^T\Gamma)^{-1}(X^T\cdot y + \Gamma^T\cdot 0_d) \\
&= (X^TX + \Gamma^T\Gamma)^{-1}X^Ty + (X^TX + \Gamma^T\Gamma)\cdot\Gamma^T\cdot 0_d \\
&= (X^TX + \Sigma^{-1})^{-1}X^Ty + \underbrace{(X^TX + \Sigma^{-1})\cdot\Gamma^T\cdot 0_d}_{0} \\
&= (X^TX + \Sigma^{-1})^{-1}X^Ty
\end{aligned}
$$

3. Vector Calculus Review

$\vec{x}, \vec{c} \in R^n$, $A \in R^{n\times n}$

(a) show $\frac{\partial}{\partial x}(x^T c) = c^T$

Solution:

$$\frac{\partial}{\partial x}(x^T c) = \frac{\partial}{\partial x}\left(\sum_i x_i \cdot c_i\right)$$

$$= \left[\frac{\partial(\sum_i x_i \cdot c_i)}{\partial x_1}, \frac{\partial(\sum_i x_i \cdot c_i)}{\partial x_2}, \cdots, \frac{\partial(\sum_i x_i c_i)}{\partial x_n}\right]$$

$$= [c_1, c_2, \cdots c_n] = c^T$$

(b). show $\frac{\partial}{\partial x}\|x\|_2^2 = 2x^T$

Solution: $\|x\|_2^2 = x_1^2 + \cdots + x_n^2$

$$\frac{\partial}{\partial x}\|x\|_2^2 = \left[\frac{\partial\|x\|_2^2}{\partial x_1}, \cdots \frac{\partial\|x\|_2^2}{\partial x_n}\right]$$

$$= [2x_1, \cdots 2x_n]$$

$$= 2x^T$$

(c) show $\frac{\partial}{\partial x}(A\vec{x}) = A$

Solution: $\frac{\partial}{\partial x}(A\vec{x})$

$$= \frac{\partial}{\partial x}[A_1 \cdot \vec{x}, A_2 \cdot \vec{x}, \cdots A_n\vec{x}]^T$$

$$\because \frac{\partial([A_1 \cdot \vec{x}, A_2 \cdot \vec{x}, \cdots, A_n\vec{x}]^T)}{\partial x_i}$$

$$= [A_{1i}, A_{2i}, \cdots, A_{ni}]^T = A^i$$

$\therefore \frac{\partial}{\partial x}(A\vec{x}) = [A^1, A^2, \cdots A^i] = A$

(d) show $\frac{\partial}{\partial x}(x^T A x) = x^T(A + A^T)$

Solution:

$$x^T \cdot A \cdot x = [x_1, \cdots x_n] \begin{bmatrix} A_{11}, \cdots A_{1n} \\ \vdots \quad \vdots \\ A_{n1} \cdots A_{nn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \left[\sum_{i=1}^{n} x_i \cdot A_{i1}, \cdots \sum_{i=1}^{n} x_i \cdot A_{in}\right] \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \sum_{j=1}^{n} x_j \cdot \left(\sum_{i=1}^{n} x_i \cdot A_{ij}\right)$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} \cdot x_i \cdot x_j$$

$$\frac{\partial \left(\sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} \cdot x_i \cdot x_j\right)}{\partial x_h} \Leftarrow h \text{ column of } \frac{\partial}{\partial x}(x^T A x)$$

it's obvious that there are only three subsets not equal to zero:

$\sum_{j \neq h} A_{hj} \cdot x_h \cdot x_j$, $\sum_{i \neq h} A_{ih} x_i \cdot x_h$ and $A_{hh} \cdot x_h \cdot x_h$

$$\therefore = \frac{\partial}{\partial x_h}\left(\sum_{j \neq h} A_{hj} x_h \cdot x_j + \sum_{i \neq h} A_{ih} x_i \cdot x_h + A_{hh} \cdot x_h^2\right)$$

$$= \sum_{j \neq h} A_{hj} \cdot x_j + \sum_{i \neq h} A_{ih} \cdot x_i + 2A_{hh} \cdot x_h$$

$$= \sum_{j=1}^{n} A_{hj} \cdot x_j + \sum_{i=1}^{n} A_{ih} \cdot x_i$$

$$= x \cdot A_h + x^T \cdot A^h$$

$$= x^T \cdot A_h^T + x^T \cdot A^h = x^T(A_h^T + A^h)$$

$\therefore \frac{\partial}{\partial X}(X^T \cdot A \cdot X)$

$= [X^T(A_1^T + A^1), \cdots, X^T(A_n^T + A^n)]$

$= X^T[(A_1^T + A^1), \cdots, (A_n^T + A^n)]$

$= X^T(A^T + A)$

(e). Under what condition is the previous derivative equal to $2X^T A$

Solution: in (d) we have proved
$\frac{\partial}{\partial X}(X^T \cdot A \cdot X) = X^T(A + A^T)$
when $A^T = A$ (A is symmetric)
$\frac{\partial}{\partial X}(X^T \cdot A \cdot X) = 2X^T A$

4. ReLu Elbow Update under SGD.

(i) The location of the 'elbow':
Solution: the location of the 'elbow' is where $wx + b = 0 \Leftrightarrow x = -\frac{b}{w}$

(ii) The derivative of the loss w.r.t $\phi(x)$, namely $\frac{dL}{d\phi}$
Solution: $L(x, y, \phi) = \frac{1}{2}\|\phi(x) - y\|_2^2$

$\therefore \frac{\partial L}{d\phi} = \frac{\partial(\frac{1}{2}\|\phi(x) - y\|_2^2)}{d\phi}$

$$= \frac{1}{2}(2\phi(x) - 2y)$$

$$= \phi(x) - y$$

(iii) The partial derivative of the loss w.r.t. w,

namely $\frac{\partial l}{\partial w}$

Solution: $\because$ According to the chain rule

$$\frac{\partial l}{\partial w} = \frac{\partial l}{\partial \phi} \cdot \frac{\partial \phi}{\partial w}$$

we have proved $\frac{\partial l}{\partial \phi} = \phi(x) - y$

$$\frac{\partial \phi}{\partial w} = \begin{cases} x, & wx + b > 0 \\ 0, & \text{else} \end{cases}$$

$$\therefore \frac{\partial l}{\partial w} = \begin{cases} x(\phi(x) - y) & wx + b > 0 \\ 0 & \text{else} \end{cases}$$

(iv) The partial derivative of the loss w.r.t. b.

namely $\frac{\partial l}{\partial b}$

Solution. $\frac{\partial l}{\partial b} = \frac{\partial l}{\partial \phi} \cdot \frac{\partial \phi}{\partial b}$

$$\frac{\partial \phi}{\partial b} = \begin{cases} 1 & wx + b > 0 \\ 0 & \text{else} \end{cases}$$

$$\therefore \frac{\partial l}{\partial b} = \begin{cases} \phi(x) - y & wx + b > 0 \\ 0 & \text{else} \end{cases}$$

(b).

Describe what happens to the slope and elbow of $\phi(x)$ when we perform gradient descent in the following cases:

(i) $\phi(x) = 0$

Solution: after performing gradient descent:

$$b' = b - \Delta b = b - \eta \frac{\partial L}{\partial b} \quad (\eta \text{ is learning rate})$$

$$w' = w - \Delta w = w - \eta \frac{\partial L}{\partial w}$$

when $\phi(x) = 0$, $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial b} = 0$

so both slop and elbow have no changes

(ii) $w > 0$, $x > 0$, and $\phi(x) > 0$.

$$\phi(x) - y = 1$$

$$\begin{cases} \frac{\partial L}{\partial w} = x \\ \frac{\partial L}{\partial b} = 1 \end{cases} \Rightarrow \begin{cases} w' = w - \eta x < w \\ b' = b - \eta \end{cases}$$

$\because$ $w' < w$ $\therefore$ the slope becomes slower.

since I'm not sure if $b > 0$ or $b < 0$

the changes of elbow can't be determined.

(iii) $w > 0$, $x < 0$. and $\phi(x) > 0$

$$\begin{cases} w' = w - \eta x > w \Rightarrow \text{the slope becomes steeper} \\ b' = b - \eta > b. \end{cases}$$

$\because$ $wx + b > 0$ and $x < 0$ $\therefore$ $b > 0$.

$\therefore$ $e' = -\frac{b'}{w'} < \frac{b}{w} < 0$ $\therefore$ elbow moves left

(iv) $w<0$, $x>0$. and $\phi(x)>0$

$$\begin{cases} w' = w - \eta x < w < 0 \quad \therefore |w'| > |w| \\ b' = b - \eta < b \quad (b>0) \end{cases}$$

$|w'| > |w| \Rightarrow$ slope becomes steeper.

$0 < e' = -\frac{b'}{w'} < -\frac{b}{w} \Rightarrow$ elbow moves left

(c) Derive the location e_i of the elbow of the i'th elementwise ReLu activation.

Solution:

assume W_i is the weight of the i'th and b_i is the bais of the i'th

then elbow $= -\frac{b_i}{W_i}$

6. Homework Process and Study Group

(a) stack overflow, CSDN

(b) none

(c) writing : 5 hours

code : 4 hours

$4+5=9$ hours in total