DDA5001 Machine Learning

Notes: Matrix Differentiation

Teaching Team

The Chinese University of Hong Kong, Shenzhen

Matrix Differentiation

We have learned some commonly used results in lecture:

- $\frac{\partial \boldsymbol{c}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{c}$,
- $\frac{\partial \|\boldsymbol{x}\|^2}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$,
- $\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top$,
- $\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A}^\top + \boldsymbol{A})\boldsymbol{x}$,
- $\frac{\partial \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2}{\partial \boldsymbol{x}} = 2\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$.

How about more complex functions? What if we want to calculate the gradient with respect to a matrix?

- We need a general rule to do matrix differentiation.

Matrix Differentiation

Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ and gradient $\nabla f(\boldsymbol{x}) := \left[\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right] \in \mathbb{R}^d$.

Basic rule:

$$df = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} dx_i = \nabla f(\boldsymbol{x})^\top d\boldsymbol{x}.$$

More general case where $\boldsymbol{X}$ is a $\mathbb{R}^{m\times n}$ matrix:

$$df = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial \boldsymbol{X}_{ij}} d\boldsymbol{X}_{ij} = \mathrm{Tr}(\frac{\partial f}{\partial \boldsymbol{X}}^\top d\boldsymbol{X}), \tag{1}$$

where $\frac{\partial f}{\partial \boldsymbol{X}} := \left[\frac{\partial f}{\partial \boldsymbol{X}_{ij}}\right] \in \mathbb{R}^{m\times n}$ is the gradient.

- The equation (1) can be used to calculate the gradient $\frac{\partial f}{\partial \boldsymbol{X}}$.

Matrix Differentiation

$$df = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial \boldsymbol{X}_{ij}} d\boldsymbol{X}_{ij} = \mathrm{Tr}(\frac{\partial f}{\partial \boldsymbol{x}}^\top d\boldsymbol{X}).$$

Example: Calculate the gradient of function $f(\boldsymbol{X}) := \boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}$.

$$\begin{aligned} df &= \boldsymbol{a}^\top d\boldsymbol{X}\boldsymbol{b} \quad (d(\boldsymbol{X}\boldsymbol{Y}) = (d\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}(d\boldsymbol{Y})) \\ \implies \mathrm{Tr}(df) &= \mathrm{Tr}(\boldsymbol{a}^\top d\boldsymbol{X}\boldsymbol{b}) \\ &= \mathrm{Tr}(\boldsymbol{b}\boldsymbol{a}^\top d\boldsymbol{X}) \quad (\mathrm{Tr}(\boldsymbol{X}\boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{X})). \end{aligned}$$

Hence, $\frac{\partial f}{\partial \boldsymbol{x}}^\top = \boldsymbol{b}\boldsymbol{a}^\top \implies \frac{\partial f}{\partial \boldsymbol{x}} = \boldsymbol{a}\boldsymbol{b}^\top$.

Matrix Differential Step: (scalar to matrix/vector)

- Calculate df and apply trace operator to df.
- Establish the relation in (1) based on basic rules.

Basic Rules

Differential Rules:

(1) $d(\boldsymbol{X} \pm \boldsymbol{Y}) = d\boldsymbol{X} \pm d\boldsymbol{Y}$.

(2) $d(\boldsymbol{X}\boldsymbol{Y}) = (d\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}(d\boldsymbol{Y})$.

(3) $d\boldsymbol{X}^{-1} = -\boldsymbol{X}^{-1}d\boldsymbol{X}\boldsymbol{X}^{-1}$.

(4) $d(\boldsymbol{X} \odot \boldsymbol{Y}) = d\boldsymbol{X} \odot \boldsymbol{Y} + \boldsymbol{X} \odot d\boldsymbol{Y}$, where $\odot$ is element-wise multiplication.

(5) $d\sigma(\boldsymbol{X}) = \sigma'(\boldsymbol{X}) \odot d\boldsymbol{X}$, where $\sigma(\boldsymbol{X}) = [\sigma(\boldsymbol{X}_{ij})]$ is element-wise function mapping.

Trace Rules:

(1) $a = \mathrm{Tr}(a)$, where a is scalar.

(2) $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{X})$, given $\boldsymbol{X}$ and $\boldsymbol{Y}^\top$ have the same size.

(3) $\mathrm{Tr}(\boldsymbol{X}^\top) = \mathrm{Tr}(\boldsymbol{X})$.

(4) $\mathrm{Tr}(\boldsymbol{X} \pm \boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{X}) \pm \mathrm{Tr}(\boldsymbol{Y})$.

(5) $\mathrm{Tr}(\boldsymbol{X}^\top(\boldsymbol{Y} \odot \boldsymbol{C})) = \mathrm{Tr}((\boldsymbol{X} \odot \boldsymbol{Y})^\top \boldsymbol{C})$, given $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{C}$ have the same size.

Exercise 1: Least Square

Problem 1: calculate the gradient of the least square problem

$$f(\boldsymbol{\theta}) = \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2.$$

Solution.

$$\begin{aligned} df &= d\left[(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^\top (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})\right] \\ &= [d(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})]^\top (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) + (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^\top [d(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})] \\ &= (\boldsymbol{X} d\boldsymbol{\theta})^\top (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) + (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^\top \boldsymbol{X} d\boldsymbol{\theta} \\ &= 2(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^\top \boldsymbol{X} d\boldsymbol{\theta}. \end{aligned}$$

Thus,

$$\begin{aligned} df &= \mathrm{Tr}(df) = \mathrm{Tr}(2(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^\top \boldsymbol{X} d\boldsymbol{\theta}) \\ &\implies \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = 2\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}). \end{aligned}$$

Exercise 2: Multi-Class Logistic Regression

Problem 2: Let $\mathbf{\Theta} \in \mathbb{R}^{K \times d}$ be the parameter. The objective of one-sample logistic regression is

$$\min_{\mathbf{\Theta}} \; f(\mathbf{\Theta}) := -\boldsymbol{y}^\top \log \sigma(\mathbf{\Theta}\boldsymbol{x}),$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input sample, $\boldsymbol{y} \in \mathbb{R}^K$ is one-hot vector whose value is 1 for the corresponding class and 0 otherwise. $\sigma(\boldsymbol{z}) := \frac{\exp(\boldsymbol{z})}{1^\top \exp(\boldsymbol{z})} \in \mathbb{R}^K$ is the softmax function, where $\exp(\boldsymbol{z})$ is element-wise exponential function.

Derive the gradient $\frac{\partial f(\mathbf{\Theta})}{\partial \mathbf{\Theta}}$.

Exercise 2: Solution

Let $\mathbf{1}$ denotes all-one vector, we have

$$\begin{aligned} f &= -\boldsymbol{y}^\top(\log\exp(\boldsymbol{\Theta}\boldsymbol{x}) - \mathbf{1}\log(\mathbf{1}^\top\exp(\boldsymbol{\Theta}\boldsymbol{x})) \\ &= -\boldsymbol{y}^\top\boldsymbol{\Theta}\boldsymbol{x} + \log(\mathbf{1}^\top\exp(\boldsymbol{\Theta}\boldsymbol{x})). \end{aligned}$$

Differentiating f:

$$\begin{aligned} df &= -\boldsymbol{y}^\top(d\boldsymbol{\Theta})\boldsymbol{x} + \frac{\mathbf{1}^\top\exp(\boldsymbol{\Theta}\boldsymbol{x}) \odot (d\boldsymbol{\Theta})\boldsymbol{x}}{\mathbf{1}^\top\exp(\boldsymbol{\Theta}\boldsymbol{x})} \\ &= -\boldsymbol{y}^\top(d\boldsymbol{\Theta})\boldsymbol{x} + \frac{\exp(\boldsymbol{\Theta}\boldsymbol{x})^\top(d\boldsymbol{\Theta})\boldsymbol{x}}{\mathbf{1}^\top\exp(\boldsymbol{\Theta}\boldsymbol{x})}. \qquad (\mathbf{1}^\top\boldsymbol{u}\odot\boldsymbol{v} = \boldsymbol{u}^\top\boldsymbol{v}) \end{aligned}$$

Take trace operator and rearrange using exchange property:

$$\operatorname{Tr}(df) = \operatorname{Tr}\left(\boldsymbol{x}\left(-\boldsymbol{y}^\top + \frac{\exp(\boldsymbol{\Theta}\boldsymbol{x})^\top}{\mathbf{1}^\top\exp(\boldsymbol{\Theta}\boldsymbol{x})}\right)d\boldsymbol{\Theta}\right).$$

Hence, the gradient is $\frac{\partial f}{\partial \boldsymbol{\Theta}} = (-\boldsymbol{y} + \sigma(\boldsymbol{\Theta}\boldsymbol{x}))\boldsymbol{x}^\top$.

(The loss is only for one-sample. How about multi-sample case?)