Matrix Calculus Foundation for Machine Learning

LI Xiucheng

SCSE Nanyang Technological University

Outline

Notation

We denote

- scalars with lower-case, x;
- vectors with bold-case, $\mathbf{x}$;
- matrices with upper-case, X;
- the elements of vectors or matrices with x_i or X_{ij};
- trace as $\mathrm{tr}(X) = \sum_{i=1} X_{ii}$ for $X \in \mathbb{R}^{n \times n}$;
- determinant as $|X|$ for $X \in \mathbb{R}^{n \times n}$;
- matrices Hadamard product as $X \odot Y$;
- vector or matrix inner product with $\langle \cdot, \cdot \rangle$.

Background

Vector and Matrix Product

For any $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^{m \times n}$, we define their inner product as

- $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i$.
- $\langle X, Y \rangle = \text{tr}(X^\top Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij}$.

Remark:

- The second one is also known as matrix Frobenius inner product.
- Frobenius inner product is compatible with vector inner product in the sense that when two matrices degrade to vectors Frobenius inner product equals to vector inner product.

Properties of Frobenius inner product

For any $X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^{m \times n}, Z \in \mathbb{R}^{m \times n}, a \in \mathbb{R}$,

- $\langle X, Y \rangle = \langle Y, X \rangle$.
- $\langle aX, Y \rangle = \langle X, aY \rangle = a\langle X, Y \rangle$.
- $\langle X + Z, Y \rangle = \langle X, Y \rangle + \langle Z, Y \rangle$.
- $\langle X, Y \odot Z \rangle = \langle X \odot Y, Z \rangle$.

Properties of Frobenius inner product

Suppose that $A \in \mathbb{R}^{m\times \ell_1}, C \in \mathbb{R}^{\ell_1 \times n}, B \in \mathbb{R}^{m\times \ell_2}, D \in \mathbb{R}^{\ell_2\times n}$, then we have

- $\langle AC, BD\rangle = \langle B^\top AC, D\rangle = \langle C, A^\top BD\rangle$,
- $\langle AC, BD\rangle = \langle ACD^\top, B\rangle = \langle A, BDC^\top\rangle$.

Remark

- The first two equations can be summarized as moving left to left by transposing.
- The last two equations can be summarized as moving right to right by transposing.

Properties of Frobenius inner product

Proof.

The first two equations are pretty obvious by using the definition of inner product; the last two equations use the fact that $\mathrm{tr}(XY) = \mathrm{tr}(YX)$ holds for any two matrices X, Y such that $X^\top$ has the same size with Y. $\square$

Matrix Derivative

Matrix Derivative

Let us denote $f = f(X) \in \mathbb{R}$.

First, consider a scalar x, we have

$$df = f'(x)dx \tag{1}$$

Similarly, for a vector $\mathbf{x}$, we have that

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \langle \nabla_{\mathbf{x}} f, d\mathbf{x} \rangle . \tag{2}$$

The above form is easy to extend to matrix as

$$df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} = \langle \nabla_X f, dX \rangle . \tag{3}$$

Matrix Differentiation Rules

Matrix Differentiation Rules

1. $d(X \pm Y) = dX \pm dY, d(XY) = (dX)Y + XdY, d(X^\top) = (dX)^\top$
2. $d\,\mathrm{tr}(X) = \mathrm{tr}(dX)$
3. $dX^{-1} = -X^{-1}(dX)X^{-1}$
4. $d|X| = \langle \mathrm{adj}(X)^\top, dX \rangle$, where $\mathrm{adj}(X)$ is the adjoint matrix of X
5. $d|X| = |X|\langle (X^{-1})^\top, dX \rangle$ when X is invertible
6. $d(X \odot Y) = (dX) \odot Y + X \odot dY$
7. $d\sigma(X) = \sigma'(X) \odot dX$, where $\sigma(\cdot)$ is an element-wise function such as sigmoid.

Remark: (1) implies that $d\langle X, Y \rangle = \langle dX, Y \rangle + \langle X, dY \rangle$. (4) is known as Jacobi's formula.

Method

The key idea is to use the properties of inner product and the matrix differentiation rules to obtain the inner product form

$$df = \langle \nabla_X f, dX \rangle .$$

Machine Learning Examples

Quadratic Function Optimization

$f(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle.$

$$\begin{aligned} df &= \langle d\mathbf{x}, A\mathbf{x} \rangle + \langle \mathbf{x}, dA\mathbf{x} \rangle \\ &= \langle A\mathbf{x}, d\mathbf{x} \rangle + \langle \mathbf{x}, Ad\mathbf{x} \rangle \\ &= \langle A\mathbf{x}, d\mathbf{x} \rangle + \langle A^\top \mathbf{x}, d\mathbf{x} \rangle \\ &= \langle A\mathbf{x} + A^\top \mathbf{x}, d\mathbf{x} \rangle \end{aligned}$$

Hence,

$$\nabla_{\mathbf{x}} f = A\mathbf{x} + A^\top \mathbf{x}.$$

Linear Regression

$f(\mathbf{w}) = \langle X\mathbf{w} - \mathbf{y}, X\mathbf{w} - \mathbf{y} \rangle.$

$$\begin{aligned} df &= \langle d(X\mathbf{w} - \mathbf{y}), X\mathbf{w} - \mathbf{y} \rangle + \langle X\mathbf{w} - \mathbf{y}, d(X\mathbf{w} - \mathbf{y}) \rangle \\ &= 2\langle X\mathbf{w} - \mathbf{y}, d(X\mathbf{w} - \mathbf{y}) \rangle \\ &= 2\langle X\mathbf{w} - \mathbf{y}, Xd\mathbf{w} \rangle = 2\langle X^\top (X\mathbf{w} - \mathbf{y}), d\mathbf{w} \rangle. \end{aligned}$$

Hence,

$$\nabla_{\mathbf{w}} f = 2X^\top (X\mathbf{w} - \mathbf{y}).$$

And

$$\nabla_{\mathbf{w}} f = 0 \quad \Longrightarrow \quad \mathbf{w}^* = (X^\top X)^{-1} X^\top \mathbf{y}.$$

Softmax Regression

In Softmax regression, $\mathbf{y}$ is a one-hot vector defining the class target distribution, $\hat{\mathbf{y}} = \text{softmax}(X\mathbf{w})$ is the model predicted distribution. The loss function is defined as the cross-entropy between $\mathbf{y}$ and $\hat{\mathbf{y}}$, i.e.,

$$\begin{aligned}
f(\mathbf{w}) &= -\langle \mathbf{y}, \log \text{softmax}(X\mathbf{w}) \rangle \\
&= -\left\langle \mathbf{y}, \log \frac{\exp(X\mathbf{w})}{\langle \mathbf{1}, \exp(X\mathbf{w}) \rangle} \right\rangle \\
&= -\langle \mathbf{y}, X\mathbf{w} - \mathbf{1} \log \langle \mathbf{1}, \exp(X\mathbf{w}) \rangle \rangle \\
&= -\langle \mathbf{y}, X\mathbf{w} \rangle + \log \langle \mathbf{1}, \exp(X\mathbf{w}) \rangle \langle \mathbf{y}, \mathbf{1} \rangle \\
&= -\langle \mathbf{y}, X\mathbf{w} \rangle + \log \langle \mathbf{1}, \exp(X\mathbf{w}) \rangle,
\end{aligned}$$

note that $\langle \mathbf{y}, \mathbf{1} \rangle = 1$.

Softmax Regression

$$f(\mathbf{w}) = -\langle \mathbf{y}, X\mathbf{w}\rangle + \log\langle \mathbf{1}, \exp(X\mathbf{w})\rangle.$$

$$\begin{aligned}
df &= -\langle \mathbf{y}, Xd\mathbf{w}\rangle + \frac{d\langle \mathbf{1}, \exp(X\mathbf{w})\rangle}{\langle \mathbf{1}, \exp(X\mathbf{w})\rangle} \\
&= -\langle \mathbf{y}, Xd\mathbf{w}\rangle + \left\langle \frac{\mathbf{1}}{\langle \mathbf{1}, \exp(X\mathbf{w})\rangle}, \exp(X\mathbf{w}) \odot d(X\mathbf{w}) \right\rangle \\
&= -\langle \mathbf{y}, Xd\mathbf{w}\rangle + \left\langle \frac{\mathbf{1} \odot \exp(X\mathbf{w})}{\langle \mathbf{1}, \exp(X\mathbf{w})\rangle}, d(X\mathbf{w}) \right\rangle \\
&= -\langle \mathbf{y}, Xd\mathbf{w}\rangle + \left\langle \frac{\exp(X\mathbf{w})}{\langle \mathbf{1}, \exp(X\mathbf{w})\rangle}, Xd\mathbf{w} \right\rangle \\
&= -\left\langle X^\top \left(\mathbf{y} - \frac{\exp(X\mathbf{w})}{\langle \mathbf{1}, \exp(X\mathbf{w})\rangle} \right), d\mathbf{w} \right\rangle
\end{aligned}$$

Hence,

$$\nabla_{\mathbf{w}} f = X^\top \frac{\exp(X\mathbf{w})}{\langle \mathbf{1}, \exp(X\mathbf{w})\rangle} - -X^\top \mathbf{y}.$$

Estimating the Covariance of Gaussian Distribution

$f(\Sigma) = \log|\Sigma| + \frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{x}_i - \boldsymbol{\mu}, \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\rangle$. The first term

$$d\log|\Sigma| = |\Sigma|^{-1} d|\Sigma| = \langle \Sigma^{-1}, d\Sigma\rangle.$$

The second term

$$\begin{aligned} d\frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{x}_i - \boldsymbol{\mu}, \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\rangle &= \frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{x}_i - \boldsymbol{\mu}, d\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\langle \mathbf{x}_i - \boldsymbol{\mu}, \Sigma^{-1}(d\Sigma)\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\langle \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top\Sigma^{-1}, d\Sigma\right\rangle. \end{aligned}$$

Estimating the Covariance of Gaussian Distribution

Let $S = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$, then

$$df = \left\langle \Sigma^{-1} - \Sigma^{-1} S \Sigma^{-1}, d\Sigma \right\rangle .$$

Hence,

$$\nabla_\Sigma f = (\Sigma^{-1} - \Sigma^{-1} S \Sigma^{-1})^\top .$$