

Mathematics Notes
for
Computer Science
Information Technology

Hazer-BJTU

2024 / 2 / 16

目录

1	深度学习中的线性代数/概率论	3
1.1	多元函数微分	3
1.2	线性回归模型的解析解	3

1 深度学习中的线性代数/概率论

1.1 多元函数微分

考虑定义在 \mathbb{R}^n 上的函数 f ，其输出为一个向量 $\mathbf{y} \in \mathbb{R}^m$ ，如果存在线性函数 L ，使得：

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + L(\mathbf{h}) + O(\|\mathbf{h}\|_2)$$

其中线性函数 L 满足：

$$L(\mathbf{x} + \mathbf{y}) = L(\mathbf{x}) + L(\mathbf{y})$$

$$L(\lambda \cdot \mathbf{x}) = \lambda \cdot L(\mathbf{x}), \lambda \in \mathbb{R}$$

那么我们就认为该函数 f 是可微的，一般来说，我们可以将线性函数 L 简单理解为线性变换，如果我们限制函数 f 的输出为一个实数 $y \in \mathbb{R}$ ，则微分也可以被表示为如下形式：

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \mathbf{w}^\top \mathbf{h} + O(\|\mathbf{h}\|_2), \mathbf{w} \in \mathbb{R}^n$$

一个基本的事实是可微 \Rightarrow 偏导数存在，因为：

$$\begin{aligned} \frac{f(\mathbf{x} + \Delta \mathbf{x}_i) - f(\mathbf{x})}{\Delta \mathbf{x}_i} &= \frac{\mathbf{w}_i \cdot \Delta \mathbf{x}_i}{\Delta \mathbf{x}_i} + \frac{O(\Delta \mathbf{x}_i)}{\Delta \mathbf{x}_i} = \mathbf{w}_i + \frac{O(\Delta \mathbf{x}_i)}{\Delta \mathbf{x}_i} \\ \Rightarrow \lim_{\Delta \mathbf{x}_i \rightarrow 0} \frac{f(\mathbf{x} + \Delta \mathbf{x}_i) - f(\mathbf{x})}{\Delta \mathbf{x}_i} &= \mathbf{w}_i + \lim_{\Delta \mathbf{x}_i \rightarrow 0} \frac{O(\Delta \mathbf{x}_i)}{\Delta \mathbf{x}_i} = \mathbf{w}_i \\ \Rightarrow \frac{\partial f}{\partial \mathbf{x}_i} &= \mathbf{w}_i \end{aligned}$$

由此可见，实际上向量 \mathbf{w} 就是由函数 f 关于各分量的偏导数构成的：

$$\mathbf{w} = \left(\frac{\partial f}{\partial \mathbf{x}_0}, \frac{\partial f}{\partial \mathbf{x}_1}, \frac{\partial f}{\partial \mathbf{x}_2}, \dots, \frac{\partial f}{\partial \mathbf{x}_n} \right)$$

定义对于向量 $\mathbf{x} \in \mathbb{R}^n$ ： $d\mathbf{x} = (d\mathbf{x}_0, d\mathbf{x}_1, d\mathbf{x}_2, \dots, d\mathbf{x}_n)$ ，则根据全微分公式可以得出如下关系：

$$d\mathbf{x}^\top \mathbf{x} = 2\mathbf{x}^\top d\mathbf{x}$$

$$d(\mathbf{x} + \mathbf{y}) = d\mathbf{x} + d\mathbf{y}$$

$$dA\mathbf{x} = A d\mathbf{x}$$

留给读者自证。

1.2 线性回归模型的解析解

一般的线性模型可以被描述为以下形式，其中 $\hat{y} \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^d, \mathbf{w} \in \mathbb{R}^d$ ：

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + \mathbf{b}$$

而对于批量的样本数据，使用 $\mathbf{X} \in \mathbb{R}^{n \times d}$ 表示 n 组样本， $\hat{\mathbf{Y}} \in \mathbb{R}^n$ 表示对于数据集上所有样本的预测结果向量，则可以进行如下矩阵表示：

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{w} + \mathbf{B}$$

对于真实的数据 \mathbf{Y} ，线性回归要求我们最小化损失 $\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2$ ，这是一个十分简单的优化问题，存在解析解，证明如下：

$$\begin{aligned}\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2 &= \sqrt{(\hat{\mathbf{Y}} - \mathbf{Y})^\top (\hat{\mathbf{Y}} - \mathbf{Y})} \\ &= \sqrt{(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})^\top (\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})}\end{aligned}$$

故问题转化为最小化 $(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})^\top (\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})$ ，这是一个二次型，我们对于 \mathbf{w} 求导：

$$\begin{aligned}d(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})^\top (\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y}) &= 2(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})^\top d(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y}) \\ &= 2(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})^\top \mathbf{X}d\mathbf{w} \\ &= 0\end{aligned}$$

故可以得到：

$$(\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y})^\top \mathbf{X} = \mathbf{O}$$

等式两边同时取转置可知：

$$\begin{aligned}\mathbf{X}^\top (\mathbf{X}\mathbf{w} + \mathbf{B} - \mathbf{Y}) &= \mathbf{O} \\ \mathbf{X}^\top \mathbf{X}\mathbf{w} &= \mathbf{X}^\top (\mathbf{Y} - \mathbf{B}) \\ \mathbf{w} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{Y} - \mathbf{B})\end{aligned}$$

即可得到参数的最优解，前提是矩阵 $\mathbf{X}^\top \mathbf{X}$ 可逆。