Mathematics Notes
for
Computer Science
Information Technology

Hazer-BJTU

2024 / 2 / 16

目录

0.1 深度学习中的线性代数/概率论

0.1.1 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：

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

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：

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

一个基本的事实是可微⇒偏导数存在，因为：

$$\frac{f(\mathbf{x}+\mathbf{h}_i) - f(\mathbf{x})}{\Delta \mathbf{x}_i} = \frac{\mathbf{w}_i \cdot \Delta \mathbf{x}_i}{\Delta \mathbf{x}_i} + \frac{O(\Delta \mathbf{x}_i)}{\Delta \mathbf{x}_i} = \mathbf{w}_i + \frac{O(\Delta \mathbf{x}_i)}{\Delta \mathbf{x}_i}$$
$$\Rightarrow \lim_{\Delta \mathbf{x}_i \to 0} \frac{f(\mathbf{x}+\mathbf{h}_i) - f(\mathbf{x})}{\Delta \mathbf{x}_i} = \mathbf{w}_i + \lim_{\Delta \mathbf{x}_i \to 0} \frac{O(\Delta \mathbf{x}_i)}{\Delta \mathbf{x}_i} = \mathbf{w}_i$$
$$\Rightarrow \frac{\partial f}{\partial \mathbf{x}_i} = \mathbf{w}_i$$

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：

$$\mathbf{w} = \left(\frac{\partial f}{\partial \mathbf{x}_1}, \frac{\partial f}{\partial \mathbf{x}_2}, \frac{\partial f}{\partial \mathbf{x}_3}, \ldots, \frac{\partial f}{\partial \mathbf{x}_n}\right)^\top$$

定义对于向量$\mathbf{x} \in \mathbb{R}^n$: $\mathrm{d}\mathbf{x} = (\mathrm{d}\mathbf{x}_1, \mathrm{d}\mathbf{x}_2, \mathrm{d}\mathbf{x}_3, \ldots, \mathrm{d}\mathbf{x}_n)$，则根据全微分公式可以得出如下关系：

$$\mathrm{d}\mathbf{x}^\top \mathbf{x} = 2\mathbf{x}^\top \mathrm{d}\mathbf{x}$$
$$\mathrm{d}(\mathbf{x}+\mathbf{y}) = \mathrm{d}\mathbf{x} + \mathrm{d}\mathbf{y}$$
$$\mathrm{d}\mathbf{A}\mathbf{x} = \mathbf{A}\mathrm{d}\mathbf{x}$$
$$\mathrm{d}\mathbf{x}^\top \mathbf{A}\mathbf{x} = 2\mathbf{x}^\top \mathbf{A}\mathrm{d}\mathbf{x}$$

在此只证明最后一条，注意到：

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{A}_{i,j}\mathbf{x}_i\mathbf{x}_j$$
$$\frac{\partial}{\partial \mathbf{x}_i}\mathbf{x}^\top \mathbf{A}\mathbf{x} = 2\mathbf{A}_{i,i}\mathbf{x}_i + 2\sum_{1 \le j \le n, j \ne i} \mathbf{A}_{i,j}\mathbf{x}_j = 2\sum_{j=1}^{n} \mathbf{A}_{i,j}\mathbf{x}_j$$
$$\Rightarrow \mathrm{d}\mathbf{x}^\top \mathbf{A}\mathbf{x} = 2\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{A}_{i,j}\mathbf{x}_j\mathrm{d}\mathbf{x}_i = 2\mathbf{x}^\top \mathbf{A}\mathrm{d}\mathbf{x}$$

与一元函数同理，如果上述函数f满足二阶偏导数连续的条件，则我们也可以利用Hessian矩阵做出更高阶的估计：

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{w}^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \mathbf{H}\mathbf{h} + O(\|\mathbf{h}\|_2^2)$$

其中Hessian矩阵的形式为：

$$\mathbf{H}_{i,j} = \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}$$

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

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

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：

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

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。

0.1.3 SVD奇异值分解