数据科学与工程数学基础 作业4 _

2021年6月2日 上午

13k 字 111 分钟

一

> 构建模型使得预测值与真实值的误差最小常用向量2-范数度量，求解模型过程中需要计算梯度，求梯度：
>
> - $f(A)=\frac{1}{2}||Ax+b-y||_2^2$，求 $\frac{\partial f}{\partial A}$
> - $f(x)=\frac{1}{2}||Ax+b-y||_2^2$，求 $\frac{\partial f}{\partial x}$
>
> ，其中 $A\in\mathbb{R}^{m\times n}, x\in\mathbb{R}^n, b, y\in\mathbb{R}^m$

由

$$\begin{aligned} f(\mathbf{A},x) &= \frac{1}{2}||\mathbf{Ax}+\mathbf{b}-\mathbf{y}||_2^2 \\ &= \frac{1}{2}(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T(\mathbf{Ax}+\mathbf{b}-\mathbf{y}) \end{aligned}$$

可知

$$
\begin{aligned}
df &= d\left[Tr\left(\frac{1}{2}(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\right)\right] \\
&= \frac{1}{2}Tr\left[d\left((\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\right)\right] \\
&= \frac{1}{2}Tr\left[d(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})+(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\right] \\
&= \frac{1}{2}Tr\left[\mathbf{x}^T\cdot d\mathbf{A}^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})+(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\cdot\mathbf{x}\right] \\
&= \frac{1}{2}\left\{Tr\left[\mathbf{x}^T\cdot d\mathbf{A}^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\right]+Tr\left[(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\cdot\mathbf{x}\right]\right\} \\
&= \frac{1}{2}\left\{Tr\left[d\mathbf{A}^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\cdot\mathbf{x}^T\right]+Tr\left[\mathbf{x}\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\right]\right\} \\
&= \frac{1}{2}\left\{Tr\left[\mathbf{x}\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\right]+Tr\left[\mathbf{x}\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\right]\right\} \\
&= \frac{1}{2}Tr\left[2\cdot\mathbf{x}\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\right] \\
&= Tr\left[\mathbf{x}\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d\mathbf{A}\right]
\end{aligned}
$$

故

$$
\frac{\partial f}{\partial \mathbf{A}} = \left(\mathbf{x}\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\right)^T = (\mathbf{Ax}+\mathbf{b}-\mathbf{y})\cdot\mathbf{x}^T
$$

又

$$
\begin{aligned}
df &= \frac{1}{2}Tr\left[d(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})+(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot d(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\right] \\
&= \frac{1}{2}Tr\left[d\mathbf{x}^T\cdot\mathbf{A}^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})+(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\cdot d\mathbf{x}\right] \\
&= \frac{1}{2}\left\{Tr\left[d\mathbf{x}^T\cdot\mathbf{A}^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})\right]+Tr\left[(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\cdot d\mathbf{x}\right]\right\} \\
&= \frac{1}{2}\left\{Tr\left[(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\cdot d\mathbf{x}\right]+Tr\left[(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\cdot d\mathbf{x}\right]\right\} \\
&= \frac{1}{2}Tr\left[2\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\cdot d\mathbf{x}\right] \\
&= Tr\left[(\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\cdot d\mathbf{x}\right]
\end{aligned}
$$

故

$$
\frac{\partial f}{\partial \mathbf{x}} = \left((\mathbf{Ax}+\mathbf{b}-\mathbf{y})^T\cdot\mathbf{A}\right)^T = \mathbf{A}^T\cdot(\mathbf{Ax}+\mathbf{b}-\mathbf{y})
$$

二

> 利用迹微分法求解
>
> $$\frac{\partial \backslash\mathbf{tr}(W^{-1})}{\partial W}$$
>
> ，其中 $W \in \mathbb{R}^{m\times m}$

由

$$
\begin{aligned}
d\,Tr(\mathbf{W}^{-1}) &= Tr\left[d\left(\mathbf{W}^{-1}\right)\right] \\
&= Tr\left[-\mathbf{W}^{-1}\cdot d\mathbf{W}\cdot\mathbf{W}^{-1}\right] \\
&= Tr\left[-\left(\mathbf{W}^{-1}\right)^2\cdot d\mathbf{W}\right]
\end{aligned}
$$

可知

$$
\frac{\partial Tr(\mathbf{W}^{-1})}{\partial \mathbf{W}} = -\left(\mathbf{W}^{-2}\right)^T
$$

三

二次型是数据分析中常用函数，求

$$\frac{\partial x^T Ax}{\partial x}, \frac{\partial x^T Ax}{\partial A}$$

，其中 $A \in \mathbb{R}^{m\times m}, x \in \mathbb{R}^m$

由

$$\begin{aligned} d\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) &= d\,Tr\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) \\ &= Tr\left[d\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)\right] \\ &= Tr\left[d\mathbf{x}^T\cdot\mathbf{A}\mathbf{x}+\mathbf{x}^T\mathbf{A}\cdot d\mathbf{x}\right] \\ &= Tr\left[\mathbf{x}^T\mathbf{A}^T d\mathbf{x}\right]+Tr\left[\mathbf{x}^T\mathbf{A}d\mathbf{x}\right] \\ &= Tr\left[\mathbf{x}^T(\mathbf{A}^T+\mathbf{A})d\mathbf{x}\right] \end{aligned}$$

故

$$\frac{\partial\mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} = \left(\mathbf{x}^T(\mathbf{A}^T+\mathbf{A})\right)^T = (\mathbf{A}+\mathbf{A}^T)\mathbf{x}$$

又

$$\begin{aligned} d\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) &= Tr\left[d\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)\right] \\ &= Tr\left[\mathbf{x}^T\cdot d\mathbf{A}\cdot\mathbf{x}\right] \\ &= Tr\left[\mathbf{x}\mathbf{x}^T d\mathbf{A}\right] \end{aligned}$$

故

$$\frac{\partial\mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial\mathbf{A}} = \left(\mathbf{x}\mathbf{x}^T\right)^T = \mathbf{x}\mathbf{x}^T$$

四

定义 $(\exp(z))_i = \exp(z_i), (\ln(z))_i = \ln(z_i)$，则

$$f(z) = \frac{\exp(z)}{\mathbf{1}^T\exp(z)}$$

成为Softmax函数，如果 $q = f(z), J = -p^T\ln(q)$，其中 $p, q, z \in \mathbb{R}^n$，并且 $\mathbf{1}^T p = 1$，则

- 证明：$\frac{\partial J}{\partial z} = q - p$
- 若 $z = Wx$，其中 $W \in \mathbb{R}^{n\times m}, x \in \mathbb{R}^m$, $\frac{\partial J}{\partial W} = (q-p)x^T$ 是否成立。

$\forall\mathbf{x},\mathbf{y},\mathbf{z}\in\mathbb{R}^2, \lambda\in\mathbb{R}$

(1) 任取 $i, j \in \{1, 2, \ldots, n\}$

易得

$$\frac{\partial J}{\partial q_j} = -\frac{p_j}{q_j}$$

当 $i \neq j$ 时，

$$\frac{\partial q_j}{\partial z_i} = -\frac{e^{z_i+z_j}}{\left(\sum_{k=1}^n e^{z_k}\right)^2} = -q_i\cdot q_j$$

当 $i = j$ 时，

$$\frac{\partial q_j}{\partial z_i} = \frac{e^{z_i}\cdot\left(\sum_{k=1}^n e^{z_k}\right) - e^{2z_i}}{\left(\sum_{k=1}^n e^{z_k}\right)^2} = q_i - q_i^2$$

故

$$
\begin{aligned}
\frac{\partial J}{\partial z_i} &= \sum_{j=1}^{n} \frac{\partial J}{\partial q_j} \cdot \frac{\partial q_j}{\partial z_i} \\
&= \sum_{j \neq i} \left(-\frac{p_j}{q_j}\right)(-q_i q_j) + \left(-\frac{p_i}{q_i}\right)(q_i - q_i^2) \\
&= q_i \cdot \sum_{j \neq i} p_j - p_i (1 - q_i)
\end{aligned}
$$

于是由 $1^T p = \sum_{i=1}^{n} p_i = 1$ 可知

$$
\frac{\partial J}{\partial z_i} = q_i(1-p_i) - p_i(1-q_i) = q_i - p_i
$$

即

$$
\frac{\partial J}{\partial \mathbf{z}} = \mathbf{q} - \mathbf{p}
$$

(2) 由 $d\,Tr(\mathbf{W}\mathbf{x}) = Tr\left(d\mathbf{W} \cdot \mathbf{x}\right) = Tr(\mathbf{x} \cdot d\mathbf{W})$ 可知

$$
\frac{\partial J}{\partial \mathbf{W}} = \mathbf{x}^T
$$

故

$$
\frac{\partial J}{\partial \mathbf{W}} = \frac{\partial J}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial \mathbf{W}} = (\mathbf{q} - \mathbf{p})\mathbf{x}^T
$$

成立

五

> 以下内容是利用极大似然估计求解多元正态分布模型的关键步骤：
>
> $$
> L = -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_t (x_t - \mu)^T \Sigma^{-1} (x_t - \mu)
> $$
>
> ，L 是对数似然，N 为样本数，d 为样本维数，$\Sigma \in \mathbb{R}^{d \times d}$ 为协方差矩阵（对称矩阵），$\mu \in \mathbb{R}^d$ 为期望向量。
>
> - 求 $\frac{\partial L}{\partial \mu}$
> - 当 $\mu = \frac{1}{N}\sum_t x_t$ 使，求 $\frac{\partial L}{\partial \Sigma}$，并求使 $\frac{\partial L}{\partial \Sigma} = 0$ 成立的 Σ。

(1)

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}}\left[(\mathbf{x}_t - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_t - \boldsymbol{\mu})\right] \\
&= -\frac{1}{2}\sum_{t=1}^{N} \frac{\partial\left[(\mathbf{x}_t - \boldsymbol{\mu})^T\right]}{\partial \boldsymbol{\mu}} \cdot \frac{\partial\left[(\mathbf{x}_t - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_t - \boldsymbol{\mu})\right]}{\partial [\mathbf{x}_t - \boldsymbol{\mu}]} \\
&= -\frac{1}{2}\sum_{t=1}^{N} \left[-2\Sigma^{-1}(\mathbf{x}_t - \boldsymbol{\mu})\right] \\
&= \Sigma^{-1} \cdot \sum_{t=1}^{N} (\mathbf{x}_t - \boldsymbol{\mu})
\end{aligned}
$$

(2)

由

$$
\begin{aligned}
dL &= Tr\left[d\left(-\frac{Nd}{2}\ln(2\pi)-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\sum_{t=1}^{N}(\mathbf{x}_t-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_t-\boldsymbol{\mu})\right)\right] \\
&= Tr\left[-\frac{N}{2}d\left(\ln|\Sigma|\right)-\frac{1}{2}\sum_{t=1}^{N}(\mathbf{x}_t-\boldsymbol{\mu})^T\cdot d\left(\Sigma^{-1}\right)\cdot(\mathbf{x}_t-\boldsymbol{\mu})\right] \\
&= Tr\left[-\frac{N}{2|\Sigma|}\cdot|\Sigma|\Sigma^{-1}d\Sigma+\frac{1}{2}\sum_{t=1}^{N}(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})^T\cdot\Sigma^{-1}d\Sigma\cdot\Sigma^{-1}\right] \\
&= Tr\left[\left(-\frac{N}{2}\cdot\Sigma^{-1}+\frac{1}{2}\sum_{t=1}^{N}\Sigma^{-1}(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})^T\cdot\Sigma^{-1}\right)d\Sigma\right]
\end{aligned}
$$

及 Σ 为对称矩阵可知

$$
\frac{\partial L}{\partial\Sigma}=\frac{1}{2}\sum_{t=1}^{N}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}-\frac{N}{2}\Sigma^{-1}
$$

故当 $\Sigma=\frac{1}{N}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$ 时，$\frac{\partial L}{\partial\Sigma}=0$

六

> 求
>
> $$\frac{\partial|X_k|}{\partial X}$$
>
> ，其中 $X\in\mathbb{R}^{m\times m}$ 为可逆矩阵。

由 $\mathbf{X}\in\mathbb{R}^{m\times m}$ 可逆可知

$$
\begin{aligned}
\frac{\partial\left|\mathbf{X}^k\right|}{\partial\mathbf{X}} &= \frac{\partial\left|\mathbf{X}^k\right|}{\partial|\mathbf{X}|}\cdot\frac{\partial|\mathbf{X}|}{\partial\mathbf{X}} \\
&= k|\mathbf{X}|^{k-1}\cdot|\mathbf{X}|\cdot(\mathbf{X}^{-1})^T \\
&= k|\mathbf{X}|^k\left(\mathbf{X}^{-1}\right)^T
\end{aligned}
$$

七

> 求
>
> $$\frac{\partial\backslash\mathrm{tr}(AXBX^TC)}{\partial X}$$
>
> ，其中 $A\in\mathbb{R}^{m\times n},X\in\mathbb{R}^{n\times k},B\in\mathbb{R}^{k\times k},C\in\mathbb{R}^{n\times m}$

由

$$
\begin{aligned}
d\left(\mathbf{AxBx}^T\mathbf{C}\right) &= Tr\left[d\left(\mathbf{AxBx}^T\mathbf{C}\right)\right] \\
&= Tr\left[\mathbf{A}\cdot d\mathbf{x}\cdot\mathbf{Bx}^T\mathbf{C}+\mathbf{AxB}\cdot d\mathbf{x}^T\cdot\mathbf{C}\right] \\
&= Tr\left[\mathbf{Bx}^T\mathbf{CA}d\mathbf{x}\right]+Tr\left[d\mathbf{x}^T\cdot\mathbf{CAxB}\right] \\
&= Tr\left[\mathbf{Bx}^T\mathbf{CA}d\mathbf{x}\right]+Tr\left[\mathbf{B}^T\mathbf{x}^T\mathbf{A}^T\mathbf{C}^T d\mathbf{x}\right] \\
&= Tr\left[\left(\mathbf{Bx}^T\mathbf{CA}+\mathbf{B}^T\mathbf{x}^T\mathbf{A}^T\mathbf{C}^T\right)d\mathbf{x}\right]
\end{aligned}
$$

可知

$$
\frac{\partial Tr\left(\mathbf{AxBx}^T\mathbf{C}\right)}{\partial\mathbf{X}}=\left(\mathbf{Bx}^T\mathbf{CA}+\mathbf{B}^T\mathbf{x}^T\mathbf{A}^T\mathbf{C}^T\right)^T=\mathbf{A}^T\mathbf{C}^T\mathbf{xB}^T+\mathbf{CAxB}
$$

八

> 求激活函数
>
> $$\sigma(x) = \frac{1}{1+e^{-x}}$$
>
> 的导数

$$\frac{d\sigma}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}\left(\frac{1}{1+e^{-\mathbf{x}}}\right) = \frac{e^{-\mathbf{x}}}{(1+e^{-\mathbf{x}})^2} = \sigma(\mathbf{x})\,(1-\sigma(\mathbf{x}))$$

九

> 求
>
> $$\frac{\partial}{\partial x}\exp\left\{-\frac{1}{2||\sigma||_2^2}||x-\mu||_2^2\right\}$$
>
> ，其中 $x, \mu, \sigma \in \mathbb{R}^n$

由

$$\begin{aligned} d\left(e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}\right) &= Tr\left[d\left(e^{-\frac{2}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}\right)\right] \\ &= Tr\left[-\frac{1}{2||\boldsymbol{\sigma}||^2}e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}\cdot d\left(||\mathbf{x}-\boldsymbol{\mu}||_2^2\right)\right] \\ &= Tr\left[-\frac{1}{2||\boldsymbol{\sigma}||^2}e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}\cdot d\left((\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\right)\right] \\ &= Tr\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^T}{||\boldsymbol{\sigma}||^2}e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}\cdot d\mathbf{x}\right] \end{aligned}$$

可知

$$\frac{\partial}{\partial\mathbf{x}}e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2} = \left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T}{||\boldsymbol{\sigma}||^2}e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}\right)^T = -\frac{(\mathbf{x}-\boldsymbol{\mu})}{||\boldsymbol{\sigma}||^2}e^{-\frac{1}{2||\boldsymbol{\sigma}||^2}||\mathbf{x}-\boldsymbol{\mu}||_2^2}$$

十

> 阅读以下代码，填写更新梯度部分的代码。（提交时，需要提交补全的代码，以及最后10次输出的截图）

实现代码：

Copy

```python
import numpy as np

N, D_in, H, D_out = 64, 1000, 100, 10
# 随机创建一些训练数据
x = np.random.randn(N, D_in)
y = np.random.randn(N, D_out)
w1 = np.random.randn(D_in, H)
w2 = np.random.randn(H, D_out)
learning_rate = 1e-6
for it in range(500):
    # Forward pass
    h = x.dot(w1)  # N * H
    h_relu = np.maximum(h, 0)  # N * H
    y_pred = h_relu.dot(w2)  # N * D_out
    # compute loss
    loss = np.square(y_pred - y).sum()
    print(it, loss)
    # Backward pass
    # compute the gradient
    grad_y_pred = y_pred - y
    grad_w2 = h_relu.T.dot(grad_y_pred)
    grad_h_relu = grad_y_pred.dot(w2.T)
    grad_h = grad_h_relu.copy()
    grad_h[h < 0] = 0
    grad_w1 = x.T.dot(grad_h)
    w1 -= learning_rate * grad_w1
    w2 -= learning_rate * grad_w2
```

输出结果（最后10次循环）：

数据科学数学基础 Mathematics DataScience

操作系统实验 内存管理

操作系统实验 I/O子系统