补充习题答案

在作业要求掌握的内容之外，还必须掌握的：

习题 1　能运用LU分解、Cholesky分解求解适定方程组

利用作业中习题4.1的结果，解线性方程组

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

其中$\boldsymbol{b} = \begin{pmatrix} 24 \\ 24 \\ 12 \end{pmatrix}$

解　先解$\boldsymbol{L}\boldsymbol{y} = \boldsymbol{b}$，得

$$\boldsymbol{y} = \begin{pmatrix} 24 \\ 12 \\ -4 \end{pmatrix}$$

再解$\boldsymbol{U}\boldsymbol{x} = \boldsymbol{y}$，得

$$\boldsymbol{x} = \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}$$

习题 2　能运用法方程组、Cholesky分解、QR分解求解方程组的最小二乘解

利用作业中习题4.3的结果，解线性方程组的最小二乘解

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

其中$\boldsymbol{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

解 $\boldsymbol{Q}=\begin{pmatrix}\frac{3}{\sqrt{11}} & \frac{-1}{\sqrt{66}} & \frac{-1}{\sqrt{6}}\\ \frac{1}{\sqrt{11}} & \frac{7}{\sqrt{66}} & \frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{11}} & -\frac{4}{\sqrt{66}} & \frac{2}{\sqrt{6}}\end{pmatrix}, \boldsymbol{R}=\begin{pmatrix}\sqrt{11} & \frac{8}{\sqrt{11}}\\ 0 & \frac{12}{\sqrt{66}}\\ 0 & 0\end{pmatrix}, \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{b}=\begin{pmatrix}\frac{3}{\sqrt{11}}\\ \frac{-1}{\sqrt{66}}\\ \frac{-1}{\sqrt{6}}\end{pmatrix}$ 解

$$\begin{pmatrix}\sqrt{11} & \frac{8}{\sqrt{11}}\\ 0 & \frac{12}{\sqrt{66}}\end{pmatrix}\boldsymbol{x}=\begin{pmatrix}\frac{3}{\sqrt{11}}\\ \frac{-1}{\sqrt{66}}\end{pmatrix}$$

得最小二乘解：

$$\boldsymbol{x}=\begin{pmatrix}1/3\\ -1/12\end{pmatrix}$$

习题 3 熟练运用含有迹、行列式、矩阵的逆及其复合微分的计算方法

- $f(\boldsymbol{X})=\ln|\boldsymbol{X}|$，$\boldsymbol{X}$可逆，求$\frac{\partial f}{\partial \boldsymbol{X}}$
- $f(\boldsymbol{X})=\|\boldsymbol{A}\boldsymbol{X}^{-1}\|_F^2$，$\boldsymbol{X}$可逆，求$\frac{\partial f}{\partial \boldsymbol{X}}$

解 (1) $f(\boldsymbol{X})=\ln|\boldsymbol{X}|$

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{X}}&=\frac{\partial f}{\partial|\boldsymbol{X}|}\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}}\\ &=\frac{1}{|\boldsymbol{X}|}|\boldsymbol{X}|\boldsymbol{X}^{-T}\\ &=\boldsymbol{X}^{-T}\end{aligned}$$

因此

$$\frac{\partial f}{\partial \boldsymbol{X}}=\boldsymbol{X}^{-T}$$

(2)

$$f(\boldsymbol{X})=\|\boldsymbol{X}^{-1}\|_F^2=\mathrm{Tr}(\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1})$$

$$\begin{aligned}\mathrm{d}f(\boldsymbol{X})&=\mathrm{Tr}[\mathrm{d}(\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1})]\\ &=\mathrm{Tr}[(\mathrm{d}\boldsymbol{X}^{-T})\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}+\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\mathrm{d}\boldsymbol{X}^{-1}]\\ &=\mathrm{Tr}[(\boldsymbol{X}^{-1}\mathrm{d}\boldsymbol{X}\boldsymbol{X}^{-1})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}+\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}\mathrm{d}\boldsymbol{X}\boldsymbol{X}^{-1}]\\ &=-\mathrm{Tr}[\boldsymbol{X}^{-T}\mathrm{d}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}+\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}\mathrm{d}\boldsymbol{X}\boldsymbol{X}^{-1}]\\ &=-\mathrm{Tr}[(\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{X}^{-T})^{\mathrm{T}}\mathrm{d}\boldsymbol{X}+(\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{X}^{-T})^{\mathrm{T}}\mathrm{d}\boldsymbol{X}]\end{aligned}$$

所以

$$\frac{\partial f}{\partial \boldsymbol{X}} = -2\boldsymbol{X}^{-T}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{X}^{-T}$$

习题 4 理解MLE、MAP的思想

给定N个独立同分布样本，$\boldsymbol{x}_t, t = 1, 2, \cdots, N$，服从多元正态分布

$$G(\boldsymbol{x}_t|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu})\}$$

其中$\boldsymbol{x}_t, \boldsymbol{\mu} \in \mathbb{R}^d$，$\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$是可逆对称矩阵。利用极大似然估计(MLE)估计参数$\boldsymbol{\mu}, \boldsymbol{\Sigma}$。

解

$$p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{t=1}^{N} G(\boldsymbol{x}_t|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

对上式取对数得到对数似然

$$L = \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{t=1}^{N} \ln G(\boldsymbol{x}_t|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

得

$$L = -\frac{Nd}{2}ln(2\pi) - \frac{N}{2}ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_t (\boldsymbol{x}_t - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu})$$

对数似然在$\frac{\partial L}{\partial \boldsymbol{\mu}} = \mathbf{0}$，$\frac{\partial L}{\partial \boldsymbol{\Sigma}} = \boldsymbol{O}$时取得极大值，因此可以利用MLE估计参数。

1.估计$\boldsymbol{\mu}$

$\frac{\partial L}{\partial \boldsymbol{\mu}} = \sum_t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu})$ 令$\frac{\partial L}{\partial \boldsymbol{\mu}} = \mathbf{0}$

$$\mu = \frac{1}{N}\sum_t \boldsymbol{x}_t$$

2.估计$\boldsymbol{\Sigma}$

$$d\boldsymbol{L} = d[\frac{N}{2}ln|\boldsymbol{\Sigma}|] - d[\frac{1}{2}\sum_t (\boldsymbol{x}_t - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu})]$$

第一项为

$$d[\frac{N}{2}ln|\boldsymbol{\Sigma}|] = -\frac{N}{2}d[ln|\boldsymbol{\Sigma}|] = -\frac{N}{2}\mathrm{Tr}[\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}]$$

第二项为

$$
\begin{aligned}
d[\frac{1}{2}\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu})] &= -\frac{1}{2}d\mathrm{Tr}[\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu})] \\
&= -\frac{1}{2}d\mathrm{Tr}[\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}] \\
&= -\frac{1}{2}\mathrm{Tr}[\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_t-\boldsymbol{\mu})^T(-\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1})] \\
&= \frac{1}{2}\mathrm{Tr}[\boldsymbol{\Sigma}^{-1}\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}]
\end{aligned}
$$

得到

$$
\frac{\partial L}{\partial \boldsymbol{\Sigma}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}
$$

令$\frac{\partial L}{\partial \boldsymbol{\Sigma}}=0$,易得$\boldsymbol{\Sigma}=\frac{1}{N}\sum_t(\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_t-\boldsymbol{\mu})^T$

习题 5 理解熵、互信息、交叉熵、KL散度的含义并掌握计算方法

证明，在多分类问题中，利用交叉熵函数作为损失函数和用KL散度作为损失函数是等价的。

解 设第i个样本$\boldsymbol{x}_i$属于y_i类，则其真实标签分布为$\boldsymbol{p}_i$，$\boldsymbol{p}_i$的第y_i个分量为1，其余分量为0的向量。设预测器f预测样本$\boldsymbol{x}_i$标签分布为$\boldsymbol{q}_i=f(\boldsymbol{x}_i;\boldsymbol{\theta})$，$\boldsymbol{q}_i$的每个分量为预测器预测的$\boldsymbol{x}_i$属于每个分类的概率，$\boldsymbol{\theta}$是预测器中要学习的参数。

$$
\text{KL散度} = (\boldsymbol{p}_i^{\mathrm{T}}\log\boldsymbol{p}_i - \boldsymbol{p}_i^{\mathrm{T}}\log\boldsymbol{q}_i)
$$

$$
\text{交叉熵} = (-\boldsymbol{p}_i^{\mathrm{T}}\log\boldsymbol{q}_i)
$$

由于真实标签分布是真实存在、固定不变的，因此

$$
\arg\min_{\boldsymbol{\theta}}\text{KL散度} = \arg\min_{\boldsymbol{\theta}}\text{交叉熵}
$$

习题 6 运用拉格朗日对偶函数解最优化问题

已知矩阵$\boldsymbol{A}\in\mathbb{R}^{p\times q}$，$\boldsymbol{B}\in\mathbb{R}^{p\times r}$。$\mathrm{rank}(\boldsymbol{A})=\min(p,q)$。未知矩阵$\boldsymbol{X}\in\mathbb{R}^{q\times r}$，列出以下优化问题并求解。

- 若$p<q$，求Frobenius范数最小的矩阵$\boldsymbol{X}$，使得$\boldsymbol{A}\boldsymbol{X}=\boldsymbol{B}$。

优化问题为

$$\begin{aligned} \texttt{min} \quad & f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X}\|_F^2 \\ \texttt{s.t.} \quad & \boldsymbol{AX} = \boldsymbol{B} \end{aligned}$$

- 若$p > q$，求矩阵$\boldsymbol{X}$，使得矩阵$\boldsymbol{AX} - \boldsymbol{B}$的Frobenius范数最小。
  优化问题为

$$\min f(\boldsymbol{X}) = \|\boldsymbol{AX} - \boldsymbol{B}\|_F^2$$

解 (1) 优化问题为

$$\begin{aligned} \texttt{min} \quad & f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X}\|_F^2 \\ \texttt{s.t.} \quad & \boldsymbol{AX} = \boldsymbol{B} \end{aligned}$$

Lagrange函数为：

$$L(\boldsymbol{X}, \boldsymbol{\Lambda}) = \frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \mathrm{Tr}(\boldsymbol{\Lambda}^{\mathrm{T}}(\boldsymbol{AX} - \boldsymbol{B}))$$

$$\frac{\partial L}{\partial \boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}$$

令$\frac{\partial L}{\partial \boldsymbol{X}} = 0$，有：

$$\boldsymbol{X} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Lambda}$$

$$g(\boldsymbol{\Lambda}) = -\frac{1}{2}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{AA}^{\mathrm{T}}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{B}$$

令$\frac{\partial g}{\partial \boldsymbol{\Lambda}} = 0$：

$$-\boldsymbol{AA}^{\mathrm{T}}\boldsymbol{\Lambda} + \boldsymbol{B} = 0$$

由$\boldsymbol{A} \in \mathbb{R}^{p\times q}, \mathrm{rank}(\boldsymbol{A}) = p$得$\boldsymbol{AA}^{\mathrm{T}}$可逆，因此

$$\boldsymbol{\Lambda} = (\boldsymbol{AA}^{\mathrm{T}})^{-1}\boldsymbol{B}$$

因此，$\boldsymbol{X}$满足$\boldsymbol{AX} = \boldsymbol{B}$的最小二范数解：

$$\boldsymbol{X} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{AA}^{\mathrm{T}})^{-1}\boldsymbol{B}$$

因此，

$$\boldsymbol{X} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{AA}^{\mathrm{T}})^{-1}\boldsymbol{B}$$

是满足$\boldsymbol{AX} = \boldsymbol{B}$的Frobinius范数最小的矩阵。

(2)若$p > q$，优化问题

$$\min f(\boldsymbol{X}) = \|\boldsymbol{AX} - \boldsymbol{B}\|_F^2$$

$f(\boldsymbol{X}) = \mathrm{Tr}((\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B})^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B})$，当$f$关于$\boldsymbol{X}$梯度为零时$f$最小。解方程

$$\frac{\partial f}{\partial \boldsymbol{X}} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}) = 0$$

12

由$p > q, \mathrm{rank}\ \boldsymbol{A} = q$可得$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$可逆。得

$$\boldsymbol{X} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}$$

因此使得$\boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}$的Frobenius范数最小的矩阵$\boldsymbol{X} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}$ 14

习题 7 运用梯度下降法和牛顿法迭代求解单变量及多变量的最优化问题

第二次作业习题15，优化问题改为$\min_x x^3 - ax$，其中$a > 0$，定义域$\{x|x > 0\}$。

解

$$f(x) = x^3 - ax$$

$f'(x) = 3x^2 - a$，$f''(x) = 6x$ 所以迭代格式

$$x_{n+1} = x_n - \frac{3x_n^2 - a}{6x_n}$$