# 2024 秋《机器学习概论》作业 2 解答

张文锦

满分 100 分，每道题 25 分（尝试解答给 1 分，答案正确第一/二/四大题每小题给6分，第四大题每小题给4分），中文或英文作答均可。

## 1 正则化的贝叶斯解释

### (a)

我们有

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} p(\theta|x,y) \\ &= \arg\max_{\theta} \frac{p(\theta)p(y|x,\theta)}{p(y|x)} \\ &= \arg\max_{\theta} p(\theta)p(y|x,\theta)\end{aligned}$$

这里，我们可以直接去掉分母中的 $p(y|x)$，因为它不依赖于 $\theta$。根据假设，有 $p(\theta|x) = p(\theta)$。

### (b)

由(a)可得

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} \log(p(\theta)p(y|x,\theta)) \\ &= \arg\min_{\theta} -\log p(y|x,\theta) - \log\left(\frac{1}{(2\pi)^{d/2}|\eta^2 I|^{1/2}}\exp\left(-\frac{\|\theta\|_2^2}{2\eta^2}\right)\right) \\ &= \arg\min_{\theta} -\log p(y|x,\theta) + \frac{\|\theta\|_2^2}{2\eta^2}\end{aligned}$$

从而$\lambda = 1/2\eta^2$.

### (c)

我们的模型可以表示为 $\vec{\mathbf{y}} = X\theta + \vec{\varepsilon}$，其中 $\vec{\varepsilon} \sim \mathcal{N}(0,\sigma^2 I)$，$\theta \sim \mathcal{N}(0,\eta^2 I)$。因此，$\vec{\mathbf{y}}|X,\theta \sim \mathcal{N}(X\theta,\sigma^2 I)$。根据(b)中的结果，我们有：

$$
\begin{aligned}
\theta_{MAP} =& \arg\min_\theta -\log p(\vec{\mathbf{y}}|x,\theta) + \frac{\|\theta\|_2^2}{2\eta^2} \\
&= \arg\min_\theta -\log\left[\frac{1}{(2\pi)^{d/2}|\sigma^2 I|^{1/2}}\exp\left(-\frac{\|\vec{\mathbf{y}}-X\theta\|_2^2}{2\sigma^2}\right)\right] + \frac{\|\theta\|_2^2}{2\eta^2} \\
&= \arg\min_\theta \frac{\|\vec{\mathbf{y}}-X\theta\|_2^2}{2\sigma^2} + \frac{\|\theta\|_2^2}{2\eta^2} \\
&= \arg\min_\theta \|\vec{\mathbf{y}}-X\theta\|_2^2 + \frac{\sigma^2}{\eta^2}\|\theta\|_2^2
\end{aligned}
$$

令 $J(\theta)$ 为上述目标函数。为了最小化 $J(\theta)$，我们对 $\theta$ 计算其梯度并令其等于 0。我们得到：

$$
\begin{aligned}
\nabla_\theta J(\hat\theta) =& 2X^T(X\hat\theta - \vec{\mathbf{y}}) + \frac{2\sigma^2}{\eta^2}\hat\theta \\
=& \left(X^TX + \frac{\sigma^2}{\eta^2}I\right)\hat\theta - X^T\vec{\mathbf{y}} \\
=& 0
\end{aligned}
$$

从而$\hat\theta_{MAP} = (X^TX + \frac{\sigma^2}{\eta^2}I)^{-1}X^T\vec{y}$.

## (d)

此时，我们的模型为$\vec{\mathbf{y}} = X\theta + \vec{\varepsilon}$, 其中$\vec{\varepsilon} \sim \mathcal{N}(0,\sigma^2 I)$且$\theta_i \sim \text{Laplace}(0,b)$， $i = 1,\ldots,n$. 类似的, $\vec{\mathbf{y}}|X,\theta \sim \mathcal{N}(X\theta,\sigma^2 I)$, $p(\theta_i) = \frac{1}{2b}\exp\left(-\frac{|\theta_i|}{b}\right)$. 进而我们有:

$$
\begin{aligned}
\theta_{MAP} =& \arg\min_\theta -\log(p(\theta)p(\vec{\mathbf{y}}|x,\theta)) \\
=& \arg\min_\theta -\log p(\vec{\mathbf{y}}|x,\theta) - \log\prod_{i=1}^d p(\theta_i) \\
=& \arg\min_\theta -\log\left[\frac{1}{(2\pi)^{d/2}|\sigma^2 I|^{1/2}}\exp\left(-\frac{\|\vec{\mathbf{y}}-X\theta\|_2^2}{2\sigma^2}\right)\right] - \sum_{i=1}^d \log\left(\frac{1}{2b}\exp\left(-\frac{|\theta_i|}{b}\right)\right) \\
=& \arg\min_\theta \frac{\|\vec{\mathbf{y}}-X\theta\|_2^2}{2\sigma^2} + \sum_{i=1}^d \frac{|\theta_i|}{b} \\
=& \arg\min_\theta \|\vec{\mathbf{y}}-X\theta\|_2^2 + \frac{2\sigma^2}{b}\|\theta\|_1
\end{aligned}
$$

从而, $\gamma = \frac{2\sigma^2}{b}$.

## 2 逻辑回归：训练稳定性

### (a)

对于正确分类的样本$x^{(i)}$，正例即$\theta^T x^{(i)} > 0$，负例有$\theta^T x^{(i)} < 0$. 用$c\theta$代替$\theta$，则对正例有$c\theta \to \infty$,对负例有$c\theta \to -\infty$，进而

$$P(y=1|x^{(i)}) = \frac{1}{1+e^{-c\theta^T x^{(i)}}} \begin{cases} \to 1, & \text{正例}, \\ \to 0, & \text{负例}. \end{cases}$$

### (b)

此时存在$\theta$使得对所有正例$x^{(i)}$有$\theta^T x^{(i)} > 0$，对负例$x^{(i)}$有$\theta^T x^{(i)} < 0$。

回溯$L(c\theta) = \sum_{i=1}^n (y^{(i)} \log(1+e^{-c\theta^T x^{(i)}}) + (1-y^{(i)}) \log(1+e^{c\theta^T x^{(i)}}))$，对所有正例，有$y^{(i)} = 1$，进而

$$y^{(i)} \log(1+e^{-c\theta^T x^{(i)}}) + (1-y^{(i)}) \log(1+e^{c\theta^T x^{(i)}}) = \log(1+e^{-c\theta^T x^{(i)}}) \tag{1}$$

结合$\theta^T x^{(i)} > 0$,故随$c \to \infty$而不断减小；类似的，对所有负例，有$y^{(i)} = 0$，进而

$$y^{(i)} \log(1+e^{-c\theta^T x^{(i)}}) + (1-y^{(i)}) \log(1+e^{c\theta^T x^{(i)}}) = \log(1+e^{c\theta^T x^{(i)}}),$$

结合$\theta^T x^{(i)} < 0$,亦随$c \to \infty$而不断减小。从而随着$c$的增大，$c\theta$每例对应的损失不断减小。进而不可能在有限点处取得最小值，故无法收敛。

### (c)

此时线性不可分，结合定义不妨设存在正例$x^{(i)}$使得$\theta^T x^{(i)} < 0$，结合(1)可知对给定$\theta$，$c \to \infty$时(1)亦趋于$\infty$，从而损失函数在$\{c\theta \mid c \in \mathbb{R}\}$上最小值点有限，进而知损失函数在整个定义域上最小值点有限，故梯度下降最终收敛。

### (d)

- 修改学习率的选择方案不会改变线性可分性，因此不会收敛。
- 同上（）
- 线性变换不会去除数据集的可分性，因此仍然不会收敛。
- 在这种情况下，正则化可以有所帮助。代价函数被修改，现在可能最小值有限。
- 加入噪声项可能会打破可分性，从而使模型收敛。

## 3 多分类问题

### (a)

由定义知$\frac{1}{\frac{1}{e}+1+e}[\frac{1}{e},1,e] \approx [0.09, 0.245, 0.665]$.

### (b)

带入NLL式子即得$-\log 0.2 = \log 5 \approx 1.6094$.

### (c)

求导得$\nabla_{W^L} NLL(a^L, \mathbf{y}) = \vec{x} \cdot (a^L - \vec{y})^T$，代入数据即得

$$\nabla_{W^L} NLL(a^L, \mathbf{y}) = \frac{1}{\frac{1}{e}+1+e} \begin{pmatrix} 1 & -1-\frac{1}{e} & \frac{1}{e} \\ 1 & -1-\frac{1}{e} & \frac{1}{e} \end{pmatrix} \approx \begin{pmatrix} 0.245 & -0.335 & 0.090 \\ 0.245 & -0.335 & 0.090 \end{pmatrix}.$$

### (d)

应为$\frac{e}{\frac{1}{e}+1+e} = 0.665$.

### (e)

应为$W^L - 0.5\nabla_{W^L} NLL = \frac{1}{\frac{1}{e}+1+e} \begin{pmatrix} \frac{1}{e}+\frac{1}{2}+e & -\frac{1}{2e}-\frac{1}{2}-e & -\frac{5}{2e}-2-2e \\ -\frac{1}{e}-\frac{3}{2}-e & \frac{5}{2}+\frac{5}{2e}+2e & 1+e+\frac{1}{2e} \end{pmatrix} \approx \begin{pmatrix} 0.878 & -0.833 & -2.045 \\ -1.122 & 2.167 & 0.955 \end{pmatrix}$

### (f)

此时$z^L = \frac{1}{\frac{1}{e}+1+e}[-1, 2+e+\frac{1}{2e}, -1-e-\frac{1}{2e}] \approx [-0.245, 1.335, -1.090]$. 带入即知概率为0.772.

## 4 神经网络

### (a)

我们有$[max(0, 1{\cdot}3{+}014\grave{}1) = 2, max(0, 0{\cdot}3{+}1{\cdot}14\grave{}1) = 13, max(0, \grave{}1{\cdot}3{+}0{\cdot}14\grave{}1) = 0, max(0 \cdot 3 + (\grave{}1) \cdot 14\grave{}1) = 0]$，故为$[2, 13, 0, 0]$.

进而带入定义知$z_1^2 = 15$，$z_2^2 = -13$,故$[a_1^2, a_2^2] = [\frac{e^{15}}{e^{15}+e^{-13}}, \frac{e^{-13}}{e^{15}+e^{-13}}] \approx [1, 0]$

## (b)

类似的按照神经网络定义计算，可知结果为$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$.

## (c)

- 直接按照定义代入计算可得$[a_1^2, a_2^2] = [\frac{1}{1+e^2}, \frac{e^2}{1+e^2}] \approx [0.12, 0.88]$，分至第二类。
- 计算得$[a_1^2, a_2^2] = [\frac{1}{2}, \frac{1}{2}]$，恰在边界。
- 计算得$[a_1^2, a_2^2] = [\frac{e^3}{e^{-1}+e^3}, \frac{e^{-1}}{e^{-1}+e^3}] \approx [0.98, 0.02]$，分至第一类。

## (d)

隐藏层输出: $a_1 = (2, 13, 0, 0)$
输出层输入: $z_2 = (15, -13)$
输出层输出: $a_2 = (1 - 6.91 \times 10^{-13}, 6.91 \times 10^{-13})$
使用交叉熵损失函数: $L = -[y_1 \ln(a_{12}) + y_2 \ln(a_{22})] = -\ln(a_{22})$
计算输出层误差项:

$$\delta_j^{(2)} = a_j^{(2)} - y_j$$

$$\delta_1^{(2)} = 1 - 0 = 1$$

$$\delta_2^{(2)} = 0 - 1 = -1$$

对于 ReLU 激活函数，导数为：

$$f_1'(z) = \begin{cases} 1, & z > 0 \\ 0, & z \le 0 \end{cases}$$

计算隐藏层误差项:

$$\delta_i^{(1)} = f_1'(z_{1i}) \sum_j w_{ij}^{(2)}$$

$$\delta_1^{(1)} = 1 \times [1 \times w_{11}^{(2)} + (-1) \times w_{12}^{(2)}] = 1 \times (1 - (-1)) = 2$$

$$\delta_2^{(1)} = 1 \times [1 \times w_{21}^{(2)} + (-1) \times w_{22}^{(2)}] = 1 \times (1 - (-1)) = 2$$

$$\delta_3^{(1)} = 0$$

$$\delta_4^{(1)} = 0$$

输出层权重更新:

$$w_{ij}^{(2)} = w_{ij}^{(2)} - \eta \delta_j^{(2)} a_i^{(1)}$$

$$w_{0j}^{(2)} = w_{0j}^{(2)} - \eta \delta_j^{(2)}$$

计算更新值:

$$w_{11}^{(2)} = 1 - 0.1 \times 1 \times 2 = 0.8$$

$$w_{12}^{(2)} = -1 - 0.1 \times (-1) \times 2 = -0.8$$

$$w_{21}^{(2)} = 1 - 0.1 \times 1 \times 13 = -0.3$$

$$w_{22}^{(2)} = -1 - 0.1 \times (-1) \times 13 = 0.3$$

$w_{31}^{(2)}$, $w_{32}^{(2)}$, $w_{41}^{(2)}$ 和 $w_{42}^{(2)}$ 均不变

$$w_{0,1}^{(2)} = 0 - 0.1 \times 1 = -0.1$$
$$w_{0.2}^{(2)} = 2 - 0.1 \times (-1) = 2.1$$

隐藏层权重更新：

$$w_{ij}^{(1)} = w_{ij}^{(1)} - \eta \delta_j^{(1)} x_i$$

$$w_{0,j}^{(1)} = w_{0,j}^{(1)} - \eta \delta_j^{(1)}$$

计算更新值

$$w_{11}^{(1)} = 1 - 0.1 \times 2 \times 3 = 0.4$$

$$w_{12}^{(1)} = 0 - 0.1 \times 2 \times 3 = -0.6$$

$$w_{21}^{(1)} = 0 - 0.1 \times 2 \times 14 = -2.8$$

$$w_{22}^{(1)} = 1 - 0.1 \times 2 \times 14 = -1.8$$

$$w_{0,1}^{(1)} = -1 - 0.1 \times 2 = -1.2$$

$$w_{0,2}^{(1)} = -1 - 0.1 \times 2 = -1.2$$

其他权重和偏置不变

更新后的权重和偏置:

隐藏层权重矩阵:

$$\begin{pmatrix} w_{11}^{(1)} & w_{12}^{(1)} & w_{13}^{1)} & w_{14}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} & w_{23}^{(1)} & w_{24}^{(1)} \end{pmatrix} = \begin{pmatrix} 0.4 & -0.6 & -1 & 0 \\ -2.8 & -1.8 & 0 & 1 \end{pmatrix}$$

隐藏层偏置:

$$[w_{0,1}^{(1)}, w_{0,2}^{(1)}, w_{0,3}^{(1)}, w_{0,4}^{(1)}] = [-1.2, -1.2, -1, -1]$$

输出层权重矩阵：

$$\begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} \\ w_{31}^{(2)} & w_{32}^{(2)} \\ w_{41}^{(2)} & w_{42}^{(2)} \end{pmatrix} = \begin{pmatrix} 0.8 & -0.8 \\ -0.3 & 0.3 \\ 1 & -1 \\ 1 & -1 \end{pmatrix}$$

输出层偏置：

$$[w_{0,1}^{(2)}, w_{0,2}^{(2)}] = [-0.1, 2.1]$$