《人工智能与机器学习基础》第一次作业

PB24000150 李欣宸

2025 年 10 月 31 日

1 第一题

1.1 (a)

证明. 考虑对一个固定的 $\boldsymbol{x}$，推导过程如下：

$$\begin{aligned}
&\mathbb{E}_{\text{train}}[\mathbb{E}_{y|\boldsymbol{x}}[(y - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2]] \\
&= \mathbb{E}_{\text{train}}[\mathbb{E}_{y|\boldsymbol{x}}[((y - f_{\text{true}}) + (f_{\text{true}} - f_{\hat{w}(\text{train})}))^2]] \\
&= \underbrace{\mathbb{E}_{y|\boldsymbol{x}}[\epsilon^2]}_{\sigma^2 \text{ 的定义}} + \underbrace{2\mathbb{E}_{\text{train}}[\mathbb{E}_{y|\boldsymbol{x}}[\epsilon(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))]]}_{\epsilon \text{ 和 } f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}) \text{ 相互独立}} + \mathbb{E}_{\text{train}}[\mathbb{E}_{y|\boldsymbol{x}}[(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2]] \\
&= \sigma^2 + 2\underbrace{\mathbb{E}_{y|\boldsymbol{x}}[\epsilon]}_{=0}\mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))] + \mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2] \\
&= \sigma^2 + \mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2]
\end{aligned} \tag{1}$$

而对于 $\mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2]$，我们又有：

$$\begin{aligned}
&\mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2 \\
&= \mathbb{E}_{\text{train}}[((f_{\text{true}}(\boldsymbol{x}) - \overline{f}(\boldsymbol{x})) + (\overline{f}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x})))^2] \\
&= \underbrace{\mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - \overline{f}(\boldsymbol{x}))^2]}_{\text{Bias}^2 \text{ 的定义}} + 2(f_{\text{true}}(\boldsymbol{x}) - \overline{f}(\boldsymbol{x}))\underbrace{\mathbb{E}_{\text{train}}[(\overline{f}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))]}_{\text{由 } \overline{f} \text{ 的定义，}=0} \\
&\quad + \underbrace{\mathbb{E}_{\text{train}}[(\overline{f}(\boldsymbol{x}) - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2]}_{\text{Var}} \\
&= \text{Bias}^2 + \text{Var}
\end{aligned} \tag{2}$$

综合 (1) 和 (2)，我们得到：

$$\mathbb{E}_{\text{train}}[\mathbb{E}_{y|\boldsymbol{x}}[(y - f_{\hat{w}(\text{train})}(\boldsymbol{x}))^2]] = \sigma^2 + \text{Bias}^2 + \text{Var} \tag{3}$$

□

1.2 (b)

由 Bias^2 的定义 $\text{Bias}^2 = \mathbb{E}_{\text{train}}[(f_{\text{true}}(\boldsymbol{x}) - \overline{f}(\boldsymbol{x}))^2]$，由于平方项的存在，$\text{Bias}^2 \geq 0$ 恒成立。

当 $f_{\text{true}} \neq \overline{f}$ 时，即模型的表示力不足以表达数据产生的规律时，Bias^2 永远大于 0，一个直观的例子就是，一条直线不能拟合不在同一直线上的 3 个点。

1.3 (c)

- Bias^2 只与我们建立的数学模型能否表示数据真实产生的机制有关，与 N 和 n 并无直接关联；但是若采取题目中 $\overline{f}$ 的定义，Bias^2 与 n 的关系是：随着 n 的增大逐渐向下趋于一个定值，这个定值与模型的表示力和 N 的大小有关；
- $\text{Var} = \mathcal{O}(\frac{1}{N})$

模型越复杂，Bias^2 越低，Var 越高；反之亦然。

2 第二题

2.1 (a)

证明. 根据 Hoeffding 不等式，对于独立有界随机变量 $Z_i \in [a, b]$，有：

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right| > \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \tag{4}$$

取 $Z_i = I(h(x_i) \neq y_i)$，由定义，它是一个有界随机变量，$Z_i \in [0, 1]$。将 $a = 0, b = 1$ 带入 (4) 得：

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right| > \epsilon\right] \leq 2\exp(-2n\epsilon^2) \tag{5}$$

而由 Z_i 定义，经验误差 $\text{err}_S(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i$，期望误差 $\text{err}_D(h) = \mathbb{E}[Z_i]$，则：

$$\Pr[(\text{err}_S(h) - \text{err}_D(h)) > \epsilon] \leq 2\exp(-2n\epsilon^2) \tag{6}$$

□

2.2 (b)

证明.

$$\begin{aligned}\Pr[\exists h \in H, (\text{err}_S(h) - \text{err}_D(h)) > \epsilon] &\leq \sum_{h\in H} \Pr[(\text{err}_S(h) - \text{err}_D(h)) > \epsilon] \\ &\leq |H| \cdot 2\exp(-2n\epsilon^2) \\ &= 2|H|\exp(-2n\epsilon^2)\end{aligned} \tag{7}$$

□

2.3 (c)

我们的数据偏差与训练集大小 n，模型类的大小 $|H|$ 有关：

- 训练集大小 n：由 (7)，其他条件不变时，h 的泛化性不好的概率随 n 的增长指数下降。这说明 n 的大小对模型的泛化性非常显著；
- 模型类的大小 $|H|$：由于 $|H|$ 的大小随模型参数量的增加指数增长；由 (7)，h 泛化性不好的概率随模型类的大小线性增长，随模型参数量的增长指数增长。

这说明，训练参数量越大的模型，需要数据集大小也越多；如果认为 (7) 给出的上界是一个紧的界，那么需要的数据集大小和参数量之间大概呈现出一个正比例关系。

为了减缓数据偏差带来的影响，我们要根据拥有数据集的大小选择合适的模型大小和正则化参数。

3 第三题

3.1 (a)

$$
\begin{aligned}
\mathcal{L}(\mu^1,\mu^2,\Sigma) &= \prod_{i=1}^{n_1} p(x_i^1|y_1) \prod_{i=1}^{n_2} p(x_i^2|y_2) \\
\implies \ln\mathcal{L} &= \sum_{i=1}^{n_1} \ln p(x_i^1|y_1) + \sum_{i=1}^{n_2} \ln p(x_i^2|y_2) \\
&= -\frac{n_1+n_2}{2}\ln 2\pi - \frac{n_1+n_2}{2}\ln|\Sigma| \\
&\quad -\frac{1}{2}\sum_{i=1}^{n_1}(x_i^1-\mu^1)^T\Sigma^{-1}(x_i^1-\mu^1) - \frac{1}{2}\sum_{i=1}^{n_2}(x_i^2-\mu^2)^T\Sigma^{-1}(x_i^2-\mu^2)
\end{aligned} \tag{8}
$$

由于 $\arg\max\mathcal{L} = \arg\max\ln\mathcal{L}$，我们要最大化 $\ln\mathcal{L}$，则：

$$
\begin{aligned}
\frac{\partial\ln\mathcal{L}}{\partial\mu^1} &= -\frac{1}{2}\sum_{i=1}^{n_1}\frac{\partial(x_i^1-\mu^1)^T\Sigma^{-1}(x_i^1-\mu^1)}{\partial\mu^1} \\
&= -\frac{1}{2}\sum_{i=1}^{n_1} -2\Sigma^{-1}(x_i^1-\mu^1) \\
&= \sum_{i=1}^{n_1}\Sigma^{-1}(x_i^1-\mu^1) = 0 \\
\implies \quad \mu^1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1 \\
\text{同理}, \mu^2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2 \\
\frac{\partial\ln\mathcal{L}}{\partial\Sigma} &= -\frac{n_1+n_2}{2}\frac{\partial\ln|\Sigma|}{\partial\Sigma} - \frac{1}{2}\sum_{i=1}^{n_1}\frac{\partial(x_i^1-\mu^1)^T\Sigma^{-1}(x_i^1-\mu^1)}{\partial\Sigma} \\
&= -\frac{1}{2}\sum_{i=1}^{n_2}\frac{\partial(x_i^2-\mu^2)^T\Sigma^{-1}(x_i^2-\mu^2)}{\partial\Sigma} \\
&= -\frac{n_1+n_2}{2}\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{n_1}\Sigma^{-1}(x_i^1-\mu^1)(x_i^1-\mu^1)^T\Sigma^{-1} \\
&\quad + \frac{1}{2}\sum_{i=1}^{n_2}\Sigma^{-1}(x_i^2-\mu^2)(x_i^2-\mu^2)^T\Sigma^{-1} \\
&= \frac{1}{2}\Sigma^{-1}\left(-(n_1+n_2)\Sigma + \sum_{i=1}^{n_1}(x_i^1-\mu^1)(x_i^1-\mu^1)^T + \sum_{i=1}^{n_2}(x_i^2-\mu^2)(x_i^2-\mu^2)^T\right)\Sigma^{-1} = 0 \\
\implies \quad \Sigma &= \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(x_i^1-\mu^1)(x_i^1-\mu^1)^T + \sum_{i=1}^{n_2}(x_i^2-\mu^2)(x_i^2-\mu^2)^T\right)
\end{aligned}
$$

综上，显式解为：

$$
\mu^1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1, \quad \mu^2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2, \tag{9}
$$

$$
\Sigma = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(x_i^1-\mu^1)(x_i^1-\mu^1)^T + \sum_{i=1}^{n_2}(x_i^2-\mu^2)(x_i^2-\mu^2)^T\right) \tag{10}
$$

3.2 (b)

考虑将所有的数据堆叠成批，可以得到输入矩阵：

$$X^1 = \begin{pmatrix} 2.0 & 2.5 & 3.0 & 2.2 & 2.8 \\ 2.5 & 2.8 & 2.7 & 3.0 & 2.6 \\ 2.0 & 2.2 & 2.5 & 2.3 & 2.4 \end{pmatrix}, \quad X^2 = \begin{pmatrix} 3.5 & 3.2 & 3.8 & 3.0 & 4.0 \\ 3.8 & 4.0 & 3.5 & 3.9 & 3.6 \\ 3.2 & 3.5 & 3.7 & 3.3 & 3.9 \end{pmatrix} \tag{11}$$

带入 (9),(10) 可得：

$$\mu^1 = \begin{pmatrix} 2.5 \\ 2.72 \\ 2.28 \end{pmatrix}, \quad \mu^2 = \begin{pmatrix} 3.5 \\ 3.76 \\ 3.52 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.136 & -0.032 & 0.064 \\ -0.032 & 0.032 & -0.0114 \\ 0.064 & -0.0114 & 0.0476 \end{pmatrix} \tag{12}$$

由于 $x|y_i \sim \mathcal{N}(\mu^i, \Sigma)$，$i = 1, 2$，则有：

$$\begin{aligned} p(x|y_1) &= \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu^1)^T\Sigma^{-1}(x-\mu^1)\right\} \\ p(x|y_2) &= \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu^2)^T\Sigma^{-1}(x-\mu^2)\right\} \end{aligned} \tag{13}$$

代入 $x = \begin{pmatrix} 2.7 & 2.9 & 3.5 \end{pmatrix}^T$，可得 $p(x|y_1) = 1.3115 \times 10^{-15}$，$p(x|y_2) = 5.9952 \times 10^{-14}$，由于输入数据中两个分类类别各半，取 $p(y_1) = p(y_2) = 0.5$，则由 Bayes 公式可得：

$$p(y_1|x) = \frac{p(x|y_1)p(y_1)}{p(x|y_1)p(y_1) + p(x|y_2)p(y_2)} \approx 0.0214 = 2.14\,\% \tag{14}$$

$$p(y_2|x) = \frac{p(x|y_2)p(y_2)}{p(x|y_1)p(y_1) + p(x|y_2)p(y_2)} \approx 0.9786 = 97.86\,\% \tag{15}$$

由 (15) 可知，该样本更有可能属于类别 y_2（即标签 1），概率为约 97.86 %。

3.3 (c)

$$\begin{aligned} z &= \ln \frac{p(x|y_1)p(y_1)}{p(x|y_2)p(y_2)} \\ &= \ln \frac{n_1}{n_2} + \ln \frac{\exp\left\{-\frac{1}{2}(x-\mu^1)^T\Sigma^{-1}(x-\mu^1)\right\}}{\exp\left\{-\frac{1}{2}(x-\mu^2)^T\Sigma^{-1}(x-\mu^2)\right\}} \\ &= \ln \frac{n_1}{n_2} - \frac{1}{2}\left((x-\mu^1)^T\Sigma^{-1}(x-\mu^1) - (x-\mu^2)^T\Sigma^{-1}(x-\mu^2)\right) \\ &= \ln \frac{n_1}{n_2} - \frac{1}{2}\left(x^T\Sigma^{-1}x - 2(\mu^1)^T\Sigma^{-1}x + (\mu^1)^T\Sigma^{-1}\mu^1 - x^T\Sigma^{-1}x + 2(\mu^2)^T\Sigma^{-1}x - (\mu^2)^T\Sigma^{-1}\mu^2\right) \\ &= \left((\mu^1-\mu^2)^T\Sigma^{-1}\right)x + \frac{1}{2}\left((\mu^2)^T\Sigma^{-1}\mu^2 - (\mu^1)^T\Sigma^{-1}\mu^1\right) + \ln\frac{n_1}{n_2} \end{aligned} \tag{16}$$

观察到 (16) 符合 $z = w \cdot x + b$ 的形式，其中：

$$\begin{aligned} w &= (\mu^1-\mu^2)^T\Sigma^{-1} \\ b &= \frac{1}{2}\left((\mu^2)^T\Sigma^{-1}\mu^2 - (\mu^1)^T\Sigma^{-1}\mu^1\right) + \ln\frac{n_1}{n_2} \end{aligned} \tag{17}$$

4 第四题

4.1 (a)

$$x = \begin{pmatrix} 3 \\ 14 \end{pmatrix} \tag{18}$$

$$a^1 = f^1(z) = \max\{0, (W^1)^T x + w_0^1\} = \begin{pmatrix} 2 & 13 & 0 & 0 \end{pmatrix}^T \tag{19}$$

$$a^2 = f^2(z) = \mathsf{Softmax}((W^2)^T a^1 + w_0^2) = \mathsf{Softmax}(\begin{pmatrix} 15 \\ -13 \end{pmatrix}) = \begin{pmatrix} 1 - e^{-28} \\ e^{-28} \end{pmatrix} \approx \begin{pmatrix} 1 - 6.91 \times 10^{-13} \\ 6.91 \times 10^{-13} \end{pmatrix} \tag{20}$$

4.2 (b)

$$X = \begin{pmatrix} 0.5 & 0 & -3 \\ 0.5 & 2 & 0.5 \end{pmatrix} \tag{21}$$

$$f^1(Z^1) = \max\{0, (W^1)^T X + w_0^1\} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \tag{22}$$

$$\tag{23}$$

4.3 (c)

先求梯度:

$$\frac{\partial \mathcal{L}}{\partial z^2} = \mathsf{Softmax}(z^2) - y = \begin{pmatrix} 1 - e^{-28} \\ e^{-28} - 1 \end{pmatrix} \approx \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{24}$$

$$\frac{\partial z^2}{\partial W^2} = a^1 \implies \frac{\partial \mathcal{L}}{\partial W^2} = a^1 \cdot \big(\mathsf{Softmax}(z^2) - y\big)^T = \begin{pmatrix} 2 & -2 \\ 13 & -13 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{25}$$

$$\frac{\partial z^2}{\partial w_0^2} = 1 \implies \frac{\partial \mathcal{L}}{\partial w_0^2} = \frac{\partial \mathcal{L}}{\partial z^2} \cdot \frac{\partial z^2}{\partial w_0^2} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{26}$$

$$\frac{\partial z^2}{\partial a^1} = W^2 \implies \frac{\partial \mathcal{L}}{\partial a^1} = \frac{\partial z^2}{\partial a^1} \cdot \frac{\partial \mathcal{L}}{\partial z^2} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \end{pmatrix} \tag{27}$$

$$\frac{\partial a^1}{\partial z^1} = \mathrm{diag}(I_{z^1>0}) \implies \frac{\partial \mathcal{L}}{\partial z^1} = \frac{\partial a^1}{\partial z^1} \cdot \frac{\partial \mathcal{L}}{\partial a^1} = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix} \tag{28}$$

$$\frac{\partial z^1}{\partial W^1} = x \implies \frac{\partial \mathcal{L}}{\partial W^1} = x \cdot \left(\frac{\partial \mathcal{L}}{\partial z^1}\right)^T = \begin{pmatrix} 6 & 6 & 0 & 0 \\ 28 & 28 & 0 & 0 \end{pmatrix} \tag{29}$$

$$\frac{\partial z^1}{\partial w_0^1} = 1 \implies \frac{\partial \mathcal{L}}{\partial w_0^1} = \frac{\partial \mathcal{L}}{\partial z^1} \cdot \frac{\partial z^1}{\partial w_0^1} = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix} \tag{30}$$

再根据学习率 $\eta = 0.1$ 更新参数:

$$W^1 \leftarrow W^1 - \eta \frac{\partial \mathcal{L}}{\partial W^1} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} - 0.1 \begin{pmatrix} 6 & 6 & 0 & 0 \\ 28 & 28 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.4 & -0.6 & -0.1 & 0 \\ -2.8 & -1.8 & 0 & -0.1 \end{pmatrix} \tag{31}$$

$$w_0^1 \leftarrow w_0^1 - \eta \frac{\partial \mathcal{L}}{\partial w_0^1} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} - 0.1 \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -1.2 \\ -1.2 \\ -1 \\ -1 \end{pmatrix} \tag{32}$$

$$W^2 \leftarrow W^2 - \eta \frac{\partial \mathcal{L}}{\partial W^2} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} - 0.1 \begin{pmatrix} 2 & -2 \\ 13 & -13 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.8 & -0.8 \\ -0.3 & 0.3 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \tag{33}$$

$$w_0^2 \leftarrow w_0^2 - \eta \frac{\partial \mathcal{L}}{\partial w_0^2} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} - 0.1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.1 \\ 2.1 \end{pmatrix} \tag{34}$$