

第 11 章：模型知识产权保护

思考题 11.1

首先，根据指示函数 $\mathbb{I}(Y_{T_k} \neq M(X_{T_k}))$ 计算每个样本的预测错误情况。当真实标签与预测结果不一致时计为 1，一致时计为 0：

第 1 个样本： $\mathbb{I}(1 \neq 1) = 0$

第 2 个样本： $\mathbb{I}(0 \neq 1) = 1$

第 3 个样本： $\mathbb{I}(1 \neq 1) = 0$

因此，计算错误样本比例，即求误差的期望值 \mathbb{E} ，将所有样本的计算结果相加并除以总样本数：

$$\mathbb{E} = \frac{\sum \mathbb{I}(Y_{T_k} \neq M(X_{T_k}))}{|T|}$$

代入数值得

$$\mathbb{E} = \frac{0 + 1 + 0}{3} = \frac{1}{3} \approx 0.333$$

所以在当前触发集下，计算得出的错误样本比例 0.333

根据式 (11-7) 判断计算出的错误比例是否满足给定的验证条件，题目给定的判定阈值为

$$\sigma = 0.4$$

验证条件为：

$$\mathbb{E}(\mathbb{I}(Y_{T_k} \neq M(X_{T_k}))) < \sigma$$

代入数值得

$$0.333 < 0.4$$

显然该条件成立。

因此，根据式 (11-7)，判断水印验证结果为 TRUE

可见，因为计算得出的错误样本比例小于阈值 σ ，满足验证条件，可以认定该模型是一个被嵌入了水印的模型。

思考题 11.2

根据式 (11-5)，交叉熵正则化项的计算公式为

$$\mathcal{L}_R = - \sum_{j=1}^T \{b_j \log(y_j) + (1 - b_j) \log(1 - y_j)\}$$

将各项的具体数值依次代入公式中展开计算：

对于第 1 项 ($j = 1$)， $b_1 = 1$ ， $y_1 = 0.9$ ，代入得

$$1 \times \log(0.9) + (1 - 1) \times \log(1 - 0.9) = \log(0.9)$$

对于第 2 项 ($j = 2$)， $b_2 = 0$ ， $y_2 = 0.2$ ，代入得

$$0 \times \log(0.2) + (1 - 0) \times \log(1 - 0.2) = \log(0.8)$$

对于第 3 项 ($j = 3$), $b_3 = 1$, $y_3 = 0.8$, 代入得

$$1 \times \log(0.8) + (1 - 1) \times \log(1 - 0.8) = \log(0.8)$$

对于第 4 项 ($j = 4$), $b_4 = 0$, $y_4 = 0.1$, 代入得

$$0 \times \log(0.1) + (1 - 0) \times \log(1 - 0.1) = \log(0.9)$$

将上述各项结果相加, 总和更新为

$$\mathcal{L}_R = -(\log(0.9) + \log(0.8) + \log(0.8) + \log(0.9))$$

进一步合并同类项, 化简得

$$\mathcal{L}_R = -(2\log(0.9) + 2\log(0.8))$$

若以自然对数 \ln 代入计算具体数值 (保留三位小数), 具体计算数值约为

$$0.656$$

思考题 11.3

(a) 包含水印样本的损失函数 $\mathcal{L}(\mathbf{W})$ 需要将原始数据的均方误差与水印样本的均方误差结合起来。原始训练数据集的损失函数为:

$$\mathcal{L}_D(\mathbf{W}) = \frac{1}{1000} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|^2$$

对于 10 个水印样本集合 S_w , 其损失函数为:

$$\mathcal{L}_{S_w}(\mathbf{W}) = \frac{1}{10} \sum_{i=1}^{10} (\mathbf{W}^\top \mathbf{x}_i^w - y_i^w)^2$$

其中 $\mathbf{x}_i^w = (x_{i,\text{diet}}^w, x_{i,\text{weight}}^w, x_{i,\text{age}}^w)^\top$ 。为了显式体现预设参数 \mathbf{W}^* 的引导作用, 将水印标签的生成公式 $y_i^w = \mathbf{W}^{*\top} \mathbf{x}_i^w + \epsilon_i$ 代入上式中 (其中 \mathbf{x}_i^w 为水印特征向量)。因此, 包含水印样本的完整损失函数可以表示为:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{1000} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|^2 + \frac{1}{10} \sum_{i=1}^{10} \left(\mathbf{W}^\top \mathbf{x}_i^w - (\mathbf{W}^{*\top} \mathbf{x}_i^w + \epsilon_i) \right)^2$$

(b) 引入正则化参数 $\lambda = 0.1$ 来平衡水印项对主任务的影响, 构建的正则化损失函数 $\mathcal{L}_{\text{reg}}(\mathbf{W})$ 为:

$$\mathcal{L}_{\text{reg}}(\mathbf{W}) = \mathcal{L}_D(\mathbf{W}) + \lambda \mathcal{L}_{S_w}(\mathbf{W})$$

在梯度下降过程中, 参数 \mathbf{W} 的更新规则为沿着损失函数的负梯度方向以学习率 η 进行步进:

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla_{\mathbf{W}} \mathcal{L}_{\text{reg}}(\mathbf{W}^{(t)})$$

将其展开，推导出的梯度下降更新规则为：

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \left(\nabla_{\mathbf{W}} \mathcal{L}_D(\mathbf{W}^{(t)}) + \lambda \nabla_{\mathbf{W}} \mathcal{L}_{S_w}(\mathbf{W}^{(t)}) \right)$$

若写成分量形式，对体重参数 w 有：

$$w^{(t+1)} = w^{(t)} - \eta \left(\frac{\partial \mathcal{L}_D}{\partial w} + \lambda \frac{\partial \mathcal{L}_{S_w}}{\partial w} \right)$$

(c) 根据 (b) 中的更新规则，体重参数 w 的更新量 Δw 计算公式为：

$$\Delta w = -\eta \left(\frac{\partial \mathcal{L}_D}{\partial w} + \lambda \frac{\partial \mathcal{L}_{S_w}}{\partial w} \right)$$

题干中已经明确给定了体重特征在当前状态下的对应梯度项：

$$\frac{\partial \mathcal{L}_D}{\partial w} = -1.2, \quad \frac{\partial \mathcal{L}_{S_w}}{\partial w} = -0.4$$

已知学习率 $\eta = 0.01$ ，正则化参数 $\lambda = 0.1$ 。将数值代入公式：

$$\Delta w = -0.01 \times (-1.2 + 0.1 \times (-0.4)) = 0.0124$$

所以，体重参数 w 的更新量为 0.0124

思考题 11.4

(a) 根据线性缩放策略公式 $\mathbf{x}_i^{w,1'} = \alpha_1 \odot \mathbf{x}_i^{w,1} + \beta_1$ ，对其求均值可得：

$$\mu_{S_w^{(1)'}} = \alpha_1 \odot \mu_{S_w^{(1)}} + \beta_1$$

若要求调整后的均值与本地数据一致，即 $\mu_{S_w^{(1)'}} = \mu_{D_1}$ ，则推导位移量 β_1 的计算公式为

$$\beta_1 = \mu_{D_1} - \alpha_1 \odot \mu_{S_w^{(1)}}$$

代入给定的缩放系数 $\alpha^{(1)} = 0.6, \alpha^{(2)} = 0.9$ 及对应的均值数值得

$$\beta^{(1)} = 0.7 - 0.6 \times 1.0 = 0.1$$

$$\beta^{(2)} = 1.3 - 0.9 \times 0.8 = 0.58$$

因此，当均值完全对齐时，计算对应的位移量分别为 $\beta^{(1)} = 0.1$ 和 $\beta^{(2)} = 0.58$

(b) 在线性变换中，方差的变化规律取决于缩放系数的平方。因此，调整后的方差公式为

$$\sigma_{S_w^{(1)'}}^2 = (\alpha_1)^2 \odot \sigma_{S_w^{(1)}}^2$$

代入 (a) 中给定的缩放系数 $\alpha_1 = [0.6, 0.9]$ 及原始方差数值得

$$\sigma_{S_w^{(1)}}^{2(1)'} = 0.6^2 \times 0.25 = 0.36 \times 0.25 = 0.09$$

$$\sigma_{S_w^{(1)}}^{2(2)'} = 0.9^2 \times 0.16 = 0.81 \times 0.16 = 0.1296$$

因此，在 (a) 的条件下，计算调整后的方差结果为 $[0.09, 0.1296]$

(c) 为了求解 $\min_{\alpha_1} \mathcal{L}_1$, 即让优化目标 $\mathcal{L}_1 = \|\mu_{S_w^{(1)'} - \mu_{D_1}}\|_2^2 + 0.5 \|\sigma_{S_w^{(1)'} - \sigma_{D_1}}^2\|_2^2$ 最小化, 需要均值差异 $\|\mu_{S_w^{(1)'} - \mu_{D_1}}\|_2^2$ 和方差差异 $\|\sigma_{S_w^{(1)'} - \sigma_{D_1}}^2\|_2^2$ 均趋于最优状态。由 (a) 问可知, 总可以通过使 $\beta_1 = \mu_{D_1} - \alpha_1 \odot \mu_{S_w^{(1)}}$ 消除均值差异。为了使方差差异最优, α_1 应满足方差完全对齐:

$$(\alpha_1)^2 \odot \sigma_{S_w^{(1)}}^2 = \sigma_{D_1}^2$$

推导得出最优缩放系数公式为

$$\alpha_1 = \sqrt{\frac{\sigma_{D_1}^2}{\sigma_{S_w^{(1)}}^2}}$$

代入目标方差与原始方差数值得最优的 $\alpha_{1,opt} = [\alpha_{opt}^{(1)}, \alpha_{opt}^{(2)}]$, 其中 $\alpha_{opt}^{(1)} = \sqrt{\frac{0.04}{0.25}} = 0.4$, $\alpha_{opt}^{(2)} = \sqrt{\frac{0.09}{0.16}} = 0.75$ 。

若代入题目给定的假设值 $\alpha^{(1)} = 0.8$ 来验证第一维的方差:

$$0.8^2 \times 0.25 = 0.16 \neq 0.04$$

产生的方差结果与目标方差存在显著差异, 无法使损失函数达到最小。

可见, 验证的结果表明当 $\alpha^{(1)} = 0.8, \alpha^{(2)} = 1.0$ 时不是最优解