

课后练习1

-

1 问题一

1.1

$$\text{令 } x = \begin{bmatrix} 0 & 1 & 1 \\ 2 & 2 & 1 \\ 3 & 4 & 1 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}, \hat{w} = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix}, \text{可以计算 } x^T x = \begin{bmatrix} 13 & 16 & 5 \\ 16 & 21 & 7 \\ 5 & 7 & 3 \end{bmatrix}$$

$$\text{那么 } \hat{w} = (x^T x)^{-1} x^T y = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \text{故 } w_1 = 1, w_2 = 0, b = 1$$

1.2

$$\text{此时 } x = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 2 & 2 & 4 & 1 \\ 3 & 4 & 8 & 1 \end{bmatrix}, \hat{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ b \end{bmatrix}, \text{计算 } x^T x = \begin{bmatrix} 13 & 16 & 32 & 5 \\ 16 & 21 & 42 & 7 \\ 32 & 42 & 84 & 14 \\ 5 & 7 & 14 & 3 \end{bmatrix}, \text{可以发现这是一}$$

个奇异矩阵(有重复的行), 没有逆矩阵, 因此此时线性回归没有唯一解

1.3

$$\text{已知 } \lambda = 1, \text{故 } x^T x + \lambda I = \begin{bmatrix} 14 & 16 & 32 & 5 \\ 16 & 22 & 42 & 7 \\ 32 & 42 & 85 & 14 \\ 5 & 7 & 14 & 4 \end{bmatrix}$$

$$\text{那么 } \hat{w} = (x^T x + \lambda I)^{-1} x^T y = \begin{bmatrix} 0.378 \\ 0.140 \\ 0.280 \\ 0.304 \end{bmatrix}, \text{因此可得 } w_1 = 0.378, w_2 = 0.140, w_3 =$$

0.280, $b = 0.304$

2 问题二

2.1

令 $X = \begin{bmatrix} x_1^T, 1 \\ \vdots \\ x_n^T, 1 \end{bmatrix}$, $R = \begin{bmatrix} r_1 & & \\ & \ddots & \\ & & r_n \end{bmatrix}$, 则我们可以将损失函数写成矩阵的形式:

$$L(\hat{w}) = (y - X\hat{w})^T R (y - X\hat{w}) + \lambda \hat{w}^T \hat{w}$$

计算偏导数:

$$\begin{aligned} \frac{\partial L(\hat{w})}{\partial \hat{w}} &= \frac{y^T R y - y^T R X \hat{w} - \hat{w}^T X^T R y + \hat{w}^T X^T R X \hat{w} + \lambda \hat{w}^T \hat{w}}{\partial \hat{w}} \\ &= \frac{y^T R y - 2y^T R X \hat{w} + \hat{w}^T X^T R X \hat{w} + \lambda \hat{w}^T \hat{w}}{\partial \hat{w}} \\ &= -2X^T R y + 2X^T R X \hat{w} + 2\lambda I \hat{w} \\ &= 0 \end{aligned}$$

解得:

$$\hat{w} = (X^T R X + \lambda I)^{-1} X^T R y$$

2.2

通过人为加入权重可以对不同的数据的重要性进行评估和加权, 能够更准确地进行评估和预测, 这样更重要的数据权重更大, 能更好地反映实际情况

加权计算和直接将 (x_i, y_i) 复制 r_i 次的结果不一样, 因为复制数据后 X 矩阵会有重复行, 虽然有L2正则化, 但是矩阵也和加权计算的不一样, 最后结果也会不一样

3 问题三

3.1

$$Cross - Entropy Loss = - \sum_{i=1}^n (y_i \log(\hat{t}_i) + (1 - y_i) \log(1 - \hat{t}_i))$$

3.2

对于 x_i 来说, 它的似然函数是 $L_i(\theta) = t_i^{y_i} + (1 - t_i)^{1-y_i}$, 其中 t_i 是一个0到1之间的实数, 表示 $y_i = 1$ 的概率, 我们依然希望这个似然函数尽量大, 使得模型能够更好地预测 y_i , 因此我们要取log再取负, 得到我们可以优化的损失函数

4 问题四

4.1

Proof 假设对于任意一个超平面 $w^T x + b = 0$, 存在最大的 k 使得 $f(x) = (kw)^T x + kb$ 是一个最大似然解, 那么对于损失函数 $L_k = -\sum_{i=1}^n (y_i \log(\sigma(z)) + (1 - y_i)(1 - \log(\sigma(z))))$, 其中 $z = (kw)^T x + kb$, 一定有 $|\sigma((kw)^T x + kb)| < |\sigma(((k+1)w)^T x + (k+1)b)|$, 此时显然有 $L_{k+1} < L_k$, 这说明 $k+1$ 相比 k 是一个更好地选择, 说明假设不成立, 不存在一个最大的 k 使得损失函数达到最小值, 因此有 $k \rightarrow \infty$ 时, $\|kw\| \rightarrow \infty$, $f(x)$ 都是一个最大似然解

4.2

L2正则化相当于对参数施加了一个惩罚, 使得参数不能过分地大, 否则就会导致损失函数变大. 模型相当于会学习一个交叉熵和正则化之间的“平衡值”, 这个值一定是有限且唯一的, 因为交叉熵希望参数尽可能大, 而正则化希望参数都为0, 二者共同作用之下, 就能确定一个唯一的解