

计算智能

作业本

习题

习题 2-1 分析为什么平方损失函数不适用于分类问题.

习题 2-2 在线性回归中, 如果我们给每个样本 $(\mathbf{x}^{(n)}, y^{(n)})$ 赋予一个权重 $r^{(n)}$, 经验风险函数为

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N r^{(n)} \left(y^{(n)} - \mathbf{w}^\top \mathbf{x}^{(n)} \right)^2, \quad (2.91)$$

计算其最优参数 \mathbf{w}^* , 并分析权重 $r^{(n)}$ 的作用.

习题 2-3 证明在线性回归中, 如果样本数量 N 小于特征数量 $D + 1$, 则 $\mathbf{X}\mathbf{X}^\top$ 的秩最大为 N .

习题 2-4 在线性回归中, 验证岭回归的解为结构风险最小化准则下的最小二乘法估计, 见公式 (2.44).

习题 2-5 在线性回归中, 若假设标签 $y \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \beta)$, 并用最大似然估计来优化参数, 验证最优参数为公式 (2.52) 的解.

习题 2-6 假设有 N 个样本 $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}$ 服从正态分布 $\mathcal{N}(\mu, \sigma^2)$, 其中 μ 未知. 1) 使用最大似然估计来求解最优参数 μ^{ML} ; 2) 若参数 μ 为随机变量, 并服从正态分布 $\mathcal{N}(\mu_0, \sigma_0^2)$, 使用最大后验估计来求解最优参数 μ^{MAP} .

习题 2-7 在习题 2-6 中, 证明当 $N \rightarrow \infty$ 时, 最大后验估计趋于最大似然估计.

习题 2-8 验证公式 (2.61).

习题 2-9 试分析什么因素会导致模型出现图 2.6 所示的高偏差和高方差情况.

<https://nnndl.github.io/>

习题 2-10 验证公式 (2.66).

习题 2-11 分别用一元、二元和三元特征的词袋模型表示文本“我打了张三”和“张三打了我”, 并分析不同模型的优缺点.

习题 2-12 对于一个三分类问题, 数据集的真实标签和模型的预测标签如下:

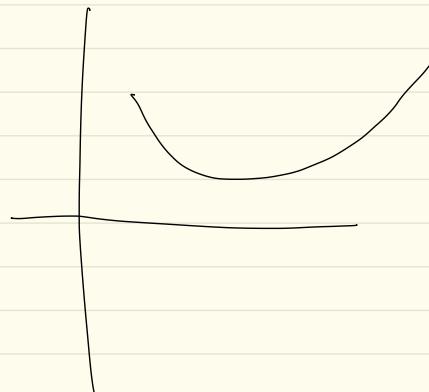
真实标签	1	1	2	2	2	3	3	3	3
预测标签	1	2	2	2	3	3	3	1	2

分别计算模型的精确率、召回率、F1 值以及它们的宏平均和微平均.

2.1 使用平方损失函数意味着模型输出是以预测值为均值的高斯分布，损失函数是在这个预测分布下的真实值的似然度，即假设了数据服从正态分布，而实际上数据并不服从正态分布。（二分类问题服从伯努利分布）

在分类问题中，标签没有连续的概念，one-hot 作为标签的一种表达方式，标签间的距离也是没有意义的，所以平方损失函数不能很好的反应分类模型的优劣程度。另外，

MSE 函数对于分类问题是非常的，用 MSE 函数不一定能将损失函数最小化。



习题2-2 在线性回归中, 如果我们给每个样本 $(\mathbf{x}^{(n)}, y^{(n)})$ 赋予一个权重 $r^{(n)}$, 经验风险函数为

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N r^{(n)} (y^{(n)} - \mathbf{w}^\top \mathbf{x}^{(n)})^2, \quad (2.91)$$

计算其最优参数 \mathbf{w}^* , 并分析权重 $r^{(n)}$ 的作用.

$$\begin{aligned} \mathcal{R}(\mathbf{w}) &= \sum_{n=1}^N \mathcal{L}(y^{(n)}, f(\mathbf{x}^{(n)}; \mathbf{w})) \\ &= \frac{1}{2} \sum_{n=1}^N (y^{(n)} - \mathbf{w}^\top \mathbf{x}^{(n)})^2 \\ &= \frac{1}{2} \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2, \end{aligned}$$

其中 $\mathbf{y} = [y^{(1)}, \dots, y^{(N)}]^\top \in \mathbb{R}^N$ 是由所有样本的真实标签组成的列向量, 而 $\mathbf{X} \in \mathbb{R}^{(D+1) \times N}$ 是由所有样本的输入特征 $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ 组成的矩阵;

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{(1)} & \mathbf{x}_1^{(2)} & \cdots & \mathbf{x}_1^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_D^{(1)} & \mathbf{x}_D^{(2)} & \cdots & \mathbf{x}_D^{(N)} \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

风险函数 $\mathcal{R}(\mathbf{w})$ 是关于 \mathbf{w} 的凸函数, 其对 \mathbf{w} 的偏导数为

$$\begin{aligned} \frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} &= \frac{1}{2} \frac{\partial \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2}{\partial \mathbf{w}} \\ &= -\mathbf{X}(\mathbf{y} - \mathbf{X}^\top \mathbf{w}), \end{aligned}$$

$\hat{\mathbf{w}} = \frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} = 0$, 得到最优的参数 \mathbf{w}^*

$$\begin{aligned} \mathbf{w}^* &= (\mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X} \mathbf{y} \\ &= \left(\sum_{n=1}^N \mathbf{x}^{(n)} (\mathbf{x}^{(n)})^\top \right)^{-1} \left(\sum_{n=1}^N \mathbf{x}^{(n)} y^{(n)} \right). \end{aligned}$$

$$\begin{aligned} \mathcal{R}(\mathbf{w}) &= \frac{1}{2} \sum r^{(n)} (y^{(n)} - \mathbf{w}^\top \mathbf{x}^{(n)})^2 \\ &= \frac{1}{2} \sum \left(\sqrt{r^{(n)}} (y^{(n)} - \mathbf{w}^\top \mathbf{x}^{(n)}) \right)^2 \\ &= \frac{1}{2} \left| \sqrt{\mathbf{R}} (\mathbf{y} - \mathbf{X}^\top \mathbf{w}) \right|^2 \end{aligned}$$

$$\frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} = -\sqrt{\mathbf{R}} \sqrt{\mathbf{R}}^\top (\mathbf{y} - \mathbf{X}^\top \mathbf{w}) = 0$$

$$\sqrt{\mathbf{R}} \sqrt{\mathbf{R}}^\top \mathbf{y} = \sqrt{\mathbf{R}} \sqrt{\mathbf{R}}^\top \mathbf{X}^\top \mathbf{w}$$

$$\mathbf{w}^* = (\sqrt{\mathbf{R}} \sqrt{\mathbf{R}}^\top \mathbf{X}^\top)^{-1} \sqrt{\mathbf{R}} \sqrt{\mathbf{R}}^\top \mathbf{y}$$

权重 $r^{(n)}$ 的作用: 是对数据的一个权重, 权重越大, 此数据对结果的影响越大, 因此, 可根据数据重要性来设置权重.

最大为N.

$$\text{rank}(AB) \leq \min \{ \text{rank}(A), \text{rank}(B) \}$$

$$N < D+1$$

X 为 $(D+1) \times N$ 的矩阵

$$\text{rank}(X) \leq \min \{ \text{rank}(D+1), \text{rank}(N) \} = N$$

$$\text{rank}(XX^T) \leq \min \{ \text{rank}(X), \text{rank}(X^T) \} = N$$

所以 XX^T 的秩最大为 N

习题 2-4 在线性回归中,验证岭回归的解为结构风险最小化准则下的最小二乘法估计,见公式(2.44).

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^T + \lambda I)^{-1} \mathbf{X}\mathbf{y}, \quad (2.43)$$

其中 $\lambda > 0$ 为预先设置的超参数, I 为单位矩阵.

岭回归的解 \mathbf{w}^* 可以看作结构风险最小化准则下的最小二乘法估计,其目标函数可以写为

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}^T \mathbf{w}\|^2 + \frac{1}{2} \lambda \|\mathbf{w}\|^2, \quad (2.44) \quad \text{参见习题 2-4.}$$

其中 $\lambda > 0$ 为正则化系数.

2.3.1.3 最大似然估计

$$\frac{\partial \mathcal{R}(\mathbf{w})}{\mathbf{w}} = -\mathbf{X}(\mathbf{y} - \mathbf{X}^T \mathbf{w}) + \lambda \mathbf{w} = 0$$

$$\lambda \mathbf{w} = \mathbf{X}\mathbf{y} - \mathbf{X}\mathbf{X}^T \mathbf{w}$$

$$\mathbf{w}^* = (\lambda + \mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y}$$

习题2-5 在线性回归中,若假设标签 $y \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$,并用最大似然估计来优化参数,验证最优参数为公式(2.52)的解.

最大似然估计 (Maximum Likelihood Estimation, MLE) 是指找到一组

参数 \mathbf{w} 使得似然函数 $p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \sigma^2)$ 最大, 等价于对数似然函数 $\log p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \sigma^2)$

最大.

令 $\frac{\partial \log p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \sigma^2)}{\partial \mathbf{w}} = 0$, 得到

即 \mathbf{w} 服从均值为 $\mathbf{0}$, 方差为 σ^2 的高斯分布. 这样, \mathbf{y} 服从均值为 $\mathbf{w}^\top \mathbf{x}$, 方差为 σ^2 的高斯分布:

$$\mathbf{w}^{ML} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}.$$

可以看出,最大似然估计的解和最小二乘法的解相同.

$$p(\mathbf{y}|\mathbf{x}; \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{y}; \mathbf{w}^\top \mathbf{x}, \sigma^2) \quad (2.47)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{w}^\top \mathbf{x})^2}{2\sigma^2}\right). \quad (2.48)$$

$$p(\mathbf{y}|\mathbf{w}) = \frac{1}{\sqrt{2\pi\beta}} \exp\left(-\frac{(\mathbf{y} - \mathbf{w}^\top \mathbf{x})^2}{2\beta^2}\right)$$

$$\log p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \log \prod p(y_i|x_i, \mathbf{w})$$

$$= \sum \log p(y_i|x_i, \mathbf{w})$$

$$= \sum -\log(\sqrt{2\pi\beta}) - \log \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\beta^2}$$

$$\sum \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\beta^2} \leftarrow \text{最小值.}$$

$$\frac{\partial}{\partial \mathbf{w}} \sum \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\beta^2} = 0$$

$$\frac{\partial}{\partial \mathbf{w}} \sum (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = 0$$

$$2 \mathbf{x}^\top (\mathbf{y} - \mathbf{x}^\top \mathbf{w}) = 0$$

$$\mathbf{x}^\top (\mathbf{y} - \mathbf{x}^\top \mathbf{w}) = 0$$

$$\begin{aligned} \mathbf{X}\mathbf{Y} &= \mathbf{X}\mathbf{X}^\top \mathbf{w} \\ \mathbf{w}^* &= (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{Y}. \end{aligned}$$

$$= \mathbf{w}^{ML}$$

参数,验证最优参数为公式(2.52)的解.

习题 2-6 假设有 N 个样本 $x^{(1)}, x^{(2)}, \dots, x^{(N)}$ 服从正态分布 $\mathcal{N}(\mu, \sigma^2)$, 其中 μ 未知. 1) 使用最大似然估计来求解最优参数 μ^{ML} ; 2) 若参数 μ 为随机变量, 并服从正态分布 $\mathcal{N}(\mu_0, \sigma_0^2)$, 使用最大后验估计来求解最优参数 μ^{MAP} .

$$1) P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mu^{ML} = \arg \max_{\mu} \log P$$

$$= \arg \max_{\mu} \log \prod_i P(x_i)$$

$$= \arg \max_{\mu} \sum_i \log P(x_i)$$

$$= \arg \max_{\mu} \sum_i \left[\log \frac{1}{\sqrt{2\pi}\sigma} - \log \frac{(x_i-\mu)^2}{2\sigma^2} \right]$$

$$= \arg \min_{\mu} \sum_i (x_i-\mu)^2$$

$$\frac{\partial}{\partial \mu} \sum_i (x_i-\mu)^2 = 0$$

$$\sum (x_i-\mu) = 0$$

$$\mu^{ML} = \frac{1}{N} \sum_{i=1}^N x_i$$

习题2-6 假设有 N 个样本 $x^{(1)}, x^{(2)}, \dots, x^{(N)}$ 服从正态分布 $\mathcal{N}(\mu, \sigma^2)$, 其中 μ 未知。
 1) 使用最大似然估计来求解最优参数 μ^{ML} ; 2) 若参数 μ 为随机变量, 并服从正态分布 $\mathcal{N}(\mu_0, \sigma_0^2)$, 使用最大后验估计来求解最优参数 μ^{MAP} .

$$2). \quad \mu^{MAP} = \arg \max_{\mu} \log p(\mu | X)$$

$$\mathbf{w}^{MAP} = \arg \max_{\mathbf{w}} p(\mathbf{y} | \mathbf{X}, \mathbf{w}; \sigma) p(\mathbf{w}; \nu),$$

$$= \arg \max_{\mu} \log \frac{p(x|\mu) p(\mu)}{p(x)}$$

$$\propto \arg \max_{\mu} [\log p(X|\mu) p(\mu)]$$

$$= \arg \max_{\mu} [\log p(X|\mu) + \log p(\mu)]$$

$$= \arg \max_{\mu} [\log \prod_i p(x_i|\mu) + \log p(\mu)]$$

$$= \arg \max_{\mu} \sum_i [\log p(x_i|\mu) + \log p(\mu)]$$

$$= \arg \max_{\mu} \sum_i \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right)$$

$$\propto \arg \max_{\mu} \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}$$

$$= \arg \min_{\mu} \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} + \frac{(\mu - \mu_0)^2}{2\sigma_0^2}$$

$$p(B_i|A) = \frac{p(B_i) P(A|B_i)}{\sum_{j=1}^k p(B_j) P(A|B_j)}$$

$$P(A|B) = \frac{p(B|A) P(A)}{p(B)}$$

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

$$\log \left(\frac{1}{\sqrt{2\pi}\sigma_0} \exp \left\{ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right\} \right)$$

$$\arg \min_{\mu} \sum \frac{(x_i - \mu)^2}{2\sigma^2} + \frac{(\mu - \mu_0)^2}{2\sigma_0^2}$$

$$\frac{\partial}{\partial \mu} \sum \frac{(x_i - \mu)^2}{2\sigma^2} + \frac{(\mu - \mu_0)^2}{2\sigma_0^2} = 0$$

$$\sum \frac{-(x_i - \mu)}{\sigma^2} + \frac{(\mu - \mu_0)}{\sigma_0^2} = 0$$

$$\sum \frac{\mu - x_i}{\sigma^2} + \frac{\mu}{\sigma_0^2} - \frac{\mu_0}{\sigma_0^2} = 0$$

$$\underline{N \frac{\mu}{\sigma^2} - \sum \frac{x_i}{\sigma^2} + \frac{\mu}{\sigma_0^2} - \frac{\mu_0}{\sigma_0^2}} = 0$$

$$N \frac{\mu}{\sigma^2} + \frac{\mu}{\sigma_0^2} = \sum \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}$$

$$6^2 N \mu + 6^2 \mu = 6^2 \sum x_i + 6^2 \mu_0$$

$$\mu = \frac{6^2 \sum x_i + 6^2 \mu_0}{N 6^2 + 6^2}$$

$$\mu^{\text{MAP}} = \frac{\sum_{i=1}^N x_i}{6^2} + \frac{\mu_0}{6^2}$$

$$\frac{N}{6^2} + \frac{1}{6^2}$$

习题2-7 在习题2-6中,证明当 $N \rightarrow \infty$ 时,最大后验估计趋于最大似然估计.

$$\mu^{MAP} = \frac{6^2 \sum x_i + 6^2 \mu_0}{N 6^2 + 6^2}$$

$$\mu^{ML} = \frac{1}{N} \sum x_i$$

$$\lim_{N \rightarrow \infty} \mu^{MAP} \underset{\sim}{\approx} \lim_{N \rightarrow \infty} \frac{6^2 \sum x_i}{N 6^2} = \frac{1}{N} \sum_{i=1}^N x_i = \mu^{ML}$$

以回归问题为例,假设样本的真实分布为 $p_r(\mathbf{x}, y)$,并采用平方损失函数,模型 $f(\mathbf{x})$ 的期望错误为

习题2-8 验证公式(2.61).

$$\mathcal{R}(f) = \mathbb{E}_{(\mathbf{x}, y) \sim p_r(\mathbf{x}, y)} [(y - f(\mathbf{x}))^2]. \quad (2.60)$$

那么最优的模型为

$$E[\bar{g}(\mathbf{x})] \geq g(E[\mathbf{x}])$$

$$f^*(\mathbf{x}) = \mathbb{E}_{y \sim p_r(y|\mathbf{x})}[y]. \quad (2.61)$$

Jense不等式 $E(E(g|\mathbf{x})) = E(g) \quad E[f(\mathbf{x})^2|\mathbf{x}] = f(\mathbf{x})^2$

$$\begin{aligned} R(f) &= E_{(\mathbf{x}, y)} [(\bar{y} - f(\mathbf{x}))^2] \\ &= E[E[(\bar{y} - f(\mathbf{x}))^2 | \mathbf{x}]] \\ &= E[E[\bar{y}^2 | \mathbf{x}] + E[\underbrace{f(\mathbf{x})^2}_{f(\mathbf{x})^2} | \mathbf{x}] - E[2\bar{y}f(\mathbf{x}) | \mathbf{x}]] \\ &= E[E[\bar{y}^2 | \mathbf{x}] + f(\mathbf{x})^2 - 2f(\mathbf{x}) E[\bar{y} | \mathbf{x}]] \\ &\geq E[E^2[\bar{y} | \mathbf{x}] + f(\mathbf{x})^2 - 2f(\mathbf{x}) E[\bar{y} | \mathbf{x}]] \\ &= E[(E[\bar{y} | \mathbf{x}] - f(\mathbf{x}))^2] \geq 0 \end{aligned}$$

令 $R(f)=0$. 则 $f(\mathbf{x}) = E[\bar{y} | \mathbf{x}]$

$$f(\mathbf{x}) = E[\bar{y} | \mathbf{x}] = E_{y \sim p_r(y | \mathbf{x})} [\bar{y}], \text{得证.}$$

习题2.9 试分析什么因素会导致模型出现图2.6所示的高偏差和高方差情况。

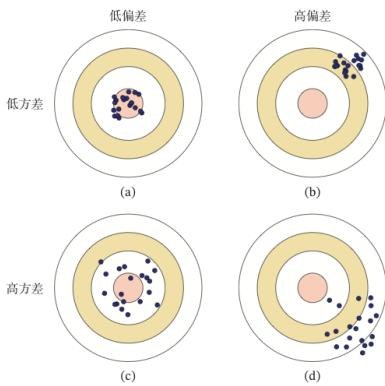


图 2.6 机器学习模型的四种偏差和方差组合情况

高偏差，拟合能力差，欠拟合，
高方差，泛化能力差，过拟合
↑
训练数据较少

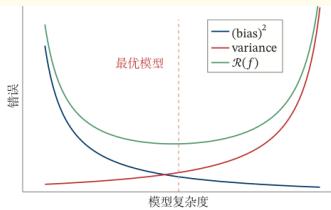


图 2.7 机器学习模型的期望错误、偏差和方差随复杂度的变化情况

模型简单，方差较大。

即模型复杂度低且训练数据较少时会偏向。

用 $y = wx + b$ 去预测

↑
时，偏差大，

方差也不小。

习题 2-10 验证公式(2.66).

对于单个样本 \mathbf{x} , 不同训练集 \mathcal{D} 得到模型 $f_{\mathcal{D}}(\mathbf{x})$ 和最优模型 $f^*(\mathbf{x})$ 的期望差距为

$$\mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - f^*(\mathbf{x}))^2 \right] \quad (2.65)$$

$$= \mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] + \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f^*(\mathbf{x}))^2 \right] \\ = \underbrace{\left(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f^*(\mathbf{x}) \right)^2}_{\text{(bias, x)}} + \underbrace{\mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2 \right]}_{\text{variance, x}}. \quad (2.66)$$

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - f^*(\mathbf{x}))^2 \right] = \mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] + \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f^*(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2 + (\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f^*(\mathbf{x}))^2 \right. \\ &\quad \left. + 2(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f^*(\mathbf{x})) \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2 + \mathbb{E}_{\mathcal{D}}^2[f_{\mathcal{D}}(\mathbf{x})] + f^*(\mathbf{x})^2 - 2f(\mathbf{x})\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] \right. \\ &\quad \left. + 2f_{\mathcal{D}}(\mathbf{x})\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - 2f_{\mathcal{D}}(\mathbf{x})f^*(\mathbf{x}) - 2\mathbb{E}_{\mathcal{D}}^2[f_{\mathcal{D}}(\mathbf{x})] + 2\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})]f^*(\mathbf{x}) \right] \\ &= \mathbb{E}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}[f_{\mathcal{D}}(\mathbf{x})])^2] + \mathbb{E}[E^2[f_{\mathcal{D}}(\mathbf{x})]] + \mathbb{E}[f^*(\mathbf{x})^2] \\ &\quad - 2\mathbb{E}[f^*(\mathbf{x})\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})]] + 2\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})]] \cancel{- 2f} \\ &\quad - 2\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})f^*(\mathbf{x})] - 2\mathbb{E}[E^2[f_{\mathcal{D}}(\mathbf{x})]] + 2\mathbb{E}[E[f_{\mathcal{D}}(\mathbf{x})]f^*(\mathbf{x})] \\ &= \mathbb{E}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}[f_{\mathcal{D}}(\mathbf{x})])^2] + \mathbb{E}^2[f_{\mathcal{D}}(\mathbf{x})] + f^*(\mathbf{x})^2 - 2f^*(\mathbf{x})\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})] \\ &\quad + 2\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})]] - 2f^*(\mathbf{x})\cancel{\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})]} - 2\cancel{\mathbb{E}^2[f_{\mathcal{D}}(\mathbf{x})]} + \cancel{2\mathbb{E}[f^*(\mathbf{x})\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})]]} \\ &= \mathbb{E}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}[f_{\mathcal{D}}(\mathbf{x})])^2] + \mathbb{E}^2[f_{\mathcal{D}}(\mathbf{x})] + f^*(\mathbf{x})^2 - 2f^*(\mathbf{x})\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})] \\ &= \left(\mathbb{E}[f_{\mathcal{D}}(\mathbf{x})] + f^*(\mathbf{x}) \right)^2 + \mathbb{E}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}[f_{\mathcal{D}}(\mathbf{x})])^2] \end{aligned}$$

$(\text{bias, x})^2$ variance, x

习题 2-11 分别用一元、二元和三元特征的词袋模型表示文本“我打了张三”和“张三打了我”，并分析不同模型的优缺点。

$$x_1 = \text{“我打了张三”} \quad x_2 = \text{“张三打了我”}$$

一元) 特征: $\{“我”, “打了”, “张三”, “”\}$ 无法表示语序, 单词相同, 向量便相同。

$$x_1 = [1, 1, 1]^T, \quad x_2 = [1, 1, 1]^T$$

二元) 特征: $\{“我”, “张三”, “我打了”, “张三打了”, “打了张三”, “打了我”, “张三#”, “我#”\}$

$$x_1 = [1, 0, 1, 0, 1, 0, 1, 0]^T \quad x_2 = [0, 1, 0, 1, 0, 1, 0, 1]^T$$

二元特征表示相邻单词间顺序, 不同语序向量不同。

三元) $\{“我打了”, “张三打了”, “我打了张三”, “张三打了我”, “打了张三#”, “打了我#”\}$

$$x_1 = [1, 0, 1, 0, 1, 0]^T \quad x_2 = [0, 1, 0, 1, 0, 1]^T$$

$N=1$, 无法表达语序。

N 过大时, 可能出现一个特征表示一个句子, 失去元信息。

且 N 过大时, N 元特征的数量会指数上升。

习题 2-12 对于一个三分类问题，数据集的真实标签和模型的预测标签如下：

真实标签	1	1	2	2	2	3	3	3	3
预测标签	1	2	2	2	3	3	3	1	2
	✓	✓	✓	✓	✓	✓	✓		

分别计算模型的精确率、召回率、F1 值以及它们的宏平均和微平均。

设 P_i 、 R_i 分别为类别 i 的精确率及召回率

$$P_1 = \frac{1}{2} \quad R_1 = \frac{1}{2} \quad F_1 = \frac{2PR}{P+R} = \frac{\frac{2}{2} \cdot \frac{1}{2}}{\frac{1}{2} + \frac{1}{2}} = \frac{1}{2}$$

$$P_2 = \frac{1}{2} \quad R_2 = \frac{2}{3} \quad F_1 = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} + \frac{2}{3}} = \frac{4}{7}$$

$$P_3 = \frac{2}{3} \quad R_3 = \frac{1}{2} \quad F_1 = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} + \frac{2}{3}} = \frac{4}{7}$$

宏平均： $P_{macro} = \frac{\frac{1}{2} + \frac{1}{2} + \frac{2}{3}}{3} = \frac{5}{9}$

$$R_{macro} = \frac{\frac{1}{2} + \frac{2}{3} + \frac{1}{2}}{3} = \frac{5}{9}$$

$$F_{macro} = \frac{2 \times \frac{5}{9} \times \frac{5}{9}}{\frac{5}{9} + \frac{5}{9}} = \frac{5}{9}$$

微平均：

$$P = R = \frac{5}{9}$$

$$F_1 = \frac{5}{9}$$