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 $(\boldsymbol{x}^{(n)}, y^{(n)})$ 赋予一个权重 $r^{(n)}$,经验风险函数为

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

计算其最优参数 $\boldsymbol{w}^*$,并分析权重 $r^{(n)}$ 的作用. 对样本再多增一个特征

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 $\boldsymbol{X}\boldsymbol{X}^{\top}$ 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 $y \sim \mathcal{N}(\boldsymbol{w}^{\top}\boldsymbol{x}, \beta)$,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 $x^{(1)}, x^{(2)}, \cdots, x^{(N)}$ 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 $N \to \infty$ 时,最大后验估计趋向于最大似然估计.

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.

$\boxed{2.2}$ 记 $R = r^{(n)}$

$$\frac{\partial R(w)}{\partial w} = -XR(y - X^T w) = 0$$

$\boxed{2.3}$

$N < D+1$

$X \in R^{D \times N}$ 又 $r(X) \leq N$

$XX^T \in R^{D \times D}$ $\therefore r(XX^T) \leq r(X) \leq N$

$\boxed{2.4}$

$$R(w) = \frac{1}{2}\|y - X^T w\|^2 + \frac{1}{2}\lambda \|w\|^2$$

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

$$= -Xy + (XX^T + \lambda I) w = 0$$

$$w = (XX^T + \lambda I)^{-1} Xy$$

$\boxed{2.5}$

$$p(y|w; x; \beta) = \prod_{i}^{N} \frac{1}{\sqrt{2\pi}\beta} \exp\left(-\frac{(y - w^T x)^2}{2\beta^2}\right)$$

$$\log p(y|X; w, \beta) = N\log \frac{1}{\sqrt{2\pi}\beta} - \frac{\|y - X^T w\|^2}{2\beta^2}$$

$$\frac{\partial \log p(y|X; w, \beta)}{\partial w} = \frac{X}{\beta^2}(y - X^T w) = 0$$

$$w^{ML} = (XX^T)^{-1} Xy$$

$\boxed{2.6}$

1)

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

$$L = \prod_{i=1}^{N} p(x) = \prod_{i=1}^{N} N(\mu, \sigma^2)$$

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

$$\log L = -N \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{N} \frac{(x^{(i)} - \mu)^2}{2\sigma^2}$$

$$\frac{\partial \log L}{\partial \mu} = \sum_{i=1}^{N} \frac{1}{\sigma^2}(x^{(i)} - \mu) = 0.$$

$$\sum_{i=1}^{N} x^{(i)} - N\mu = 0$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

$$= \frac{X}{N} = \bar{X}$$

2)

$$p(\mu; \mu_0, \sigma_0^2) = N(\mu_0, \sigma_0^2)$$

$$L \propto p(x; \mu, \sigma) p(\mu; \mu_0, \sigma_0^2)$$

$$\log L \propto \log p(X; \mu, \sigma) + \log p(\mu; \mu_0, \sigma_0^2)$$

$$= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x^{(n)} - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2$$

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x^{(n)} - \mu) - \frac{1}{\sigma_0^2}(\mu - \mu_0)$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} x^{(n)} - \frac{N\mu}{\sigma^2}) - \frac{\mu}{\sigma_0^2} + \frac{\mu_0}{\sigma_0^2} = 0$$

$$\frac{\sum_{n=1}^{N} x^{(n)}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu$$

$$\mu^{MAP} = \frac{\sigma_0^2 \sum_{n=1}^{N} x^{(n)} + \sigma^2 \mu_0}{(N\sigma_0^2 + \sigma^2)}$$

后验 $\propto$ 似然 $\times$ 先验

似然函数的 可特性
来源于 联合概率，又x独立同分布，
故可以分别为 连乘的形式

最大似然估计： $p(y|X;w,\sigma)$ max
即是对着 $p(y^{(n)}|x^{(n)};w,\sigma)$ max
即 给定 $x^{(n)}$ 下 正确 $y^{(n)}$ 标签的概率最大

最大后验估计 = 结构风险最小化 $(XX^T+\lambda I)^{-1}Xy$
添加了正则化项 相当于增加 对 w的先验约束

2.7 $N\to\infty$
$$\mu^{MAP} \to \frac{\sigma_0^2\sum_{n=1}^{N}x^{(n)}}{N\sigma_0^2} = \bar{x}$$

2.8

$R(f) = E_{(x,y)\sim p_r(x,y)}[(y-f(x))^2]$ 联合分布

$= E_{x\sim p_r(x)}[E_{y\sim p_r(y|x)}[(y-f(x))^2]]$ 分开计算

$= E_{x\sim p_r(x)}[E_{y\sim p_r(y|x)}[(y-E_{y\sim p_r(y|x)}[y]+E_{y\sim p_r(y|x)}[y]-f(x))^2]]$ 添加常量 $E_{y\sim p_r(y|x)}[y]$

$= E_{x\sim p_r(x)}[\underbrace{E_y[(y-E_y[y])^2]}_{①}+\underbrace{E_y[(E_y[y]-f(x))^2]}_{②}+2\underbrace{E_y[(y-E_y[y])(E_y[y]-f(x))]}_{③}]$

① 与x无关

③ 有 $E_y[(y-E_y[y])] = E_y[y]-E_y[y]=0$ （$E(E(a))=a$），故③ =0

故要使 $R(f)$ min，找到一组 x，故只有使 $f(x)=E_y[y]$，即② =0

$\therefore$ 有 $f^*(x)=E_{y\sim p_r(y|x)}[y]$ 时，$R(f)$ min.

$\boxed{2.9}$

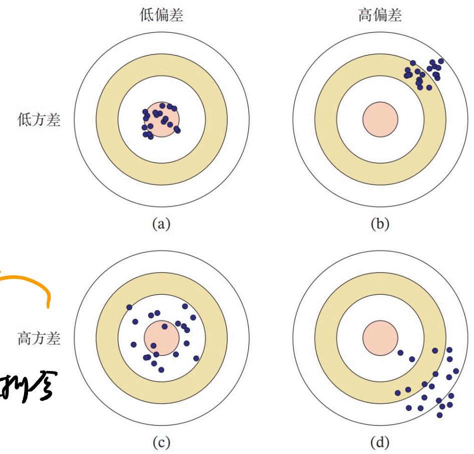

图 2.6　机器学习模型的四种偏差和方差组合情况

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.

$\boxed{2.10}$

$$E_D[(f_D(x)-f^*(x))^2]$$

$$= E_D[(f_D(x) - E_D[f_D(x)] + E_D[f_D(x)] - f^*(x))^2]$$

$$= \underbrace{E_D[(f_D(x)-E_D[f_D(x)])^2]}_{方差} + \underbrace{E_D[(E_D[f_D(x)]-f^*(x))^2]}_{偏差} + 2E_D[(f_D(x)-E_D[f_D(x)])(\cdots)]$$

其中 $E_D[(f_D(x)-E_D[f_D(x)])] = 0$ 得 0 　　$f^*(x)$ 常量, $E_D[f_D(x)]$ 常量.

$$= E_D[(E_D[f_D(x)]-f^*(x))^2] = (E_D[f_D(x)]-f^*(x))^2$$

$$= (bias.x)^2$$

$\boxed{2.11}$

一元

"$" "我" "打了" "张三" "#"

x_1: "我打了张三"
x_2: "张三打了我"

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

一元无法反映词序信息

二元

"$我" "$张三" "我打了" "张三打了" "打了张三" "打了我" "张三#" "我#"

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

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#"

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

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.

$$\frac{\frac{2}{3}\times 6}{\frac{7}{6}}$$

$\boxed{2.12}$

混淆

CM:

真实	$\hat{y}=c$	$\hat{y}\neq c$
$y=c$	TP_c	TN_c
$y\neq c$	FP_c	FN_c

$C=1$

1	1
1	6

$C=2$

2	1
2	4

$C=3$

2	2
1	4

$C=1$ 时，$P_c = \frac{1}{2}$，$R_c = \frac{1}{2}$，$F1_c = \frac{2\times\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}\times 1+\frac{1}{2}} = \frac{1}{2}$

$C=2$ 时，$P_c = \frac{1}{2}$，$R_c = \frac{2}{3}$，$F1_c = \frac{2\times\frac{1}{2}\times\frac{2}{3}}{1\times\frac{1}{2}+\frac{2}{3}} = \frac{4}{7}$

$C=3$ 时，$P_c = \frac{2}{3}$，$R_c = \frac{1}{2}$，$F1_c = \frac{2\times\frac{1}{2}\times\frac{2}{3}}{\frac{2}{3}+\frac{1}{2}} = \frac{4}{7}$

宏平均：$P_{marco} = \frac{1}{3}\times(1+\frac{2}{3}) = \frac{5}{9}$，$R_m = \frac{5}{9}$，$F1_m = \frac{2\times\frac{5}{9}\times\frac{5}{9}}{\frac{10}{9}} = \frac{5}{9}$