

习题 2-1

平方损失函数如下

$$\mathcal{L}(y, f(x, \theta)) = \frac{1}{2} (y - f(x, \theta))^2$$

1. 使用平方损失函数即假设数据分布为正态分布
2. 很多情况下, $f(x, \theta)$ 是非线性的, 可能会导致优化困难

习题 2-2

Suppose: y has shape $(n, 1)$, x has shape (n, m) , w has shape $(m, 1)$

$$\begin{aligned}\mathcal{R}(w) &= \frac{1}{2} \sum_{n=1}^N r^{(n)} (y^{(n)} - w^T x^{(n)})^2 \\ &= \frac{1}{2} \sum_{n=1}^N (y^{(n)} - w^T x^{(n)}) r^{(n)} (y^{(n)} - w^T x^{(n)}) \\ &= \frac{1}{2} (y - xw)^T r (y - xw)\end{aligned}$$

and, $r = \text{diag}(r^{(1)}, \dots, r^{(n)})$

$$\begin{aligned}\nabla_w \mathcal{R} &= \frac{1}{2} \nabla_w (y^T r y - w^T x^T r y - y^T r x w + w^T x^T r x w) \\ &= \frac{1}{2} \nabla_w (y^T r y - 2y^T r x w + w^T x^T r x w) \\ &= \frac{1}{2} (-2y^T r x + 2w^T x^T r x) \\ &= w^T x^T r x - y^T r x\end{aligned}$$

Let $\nabla_w \mathcal{R} = 0$, then

$$\begin{aligned}w^T x^T r x &= y^T r x \\ w^T &= y^T r x (x^T r x)^{-1} \\ w^* &= (x^T r x)^{-1} x^T r y\end{aligned}$$

r 控制了每个测试数据对损失的权重, 故可以决定当前权重主要受哪几个测试数据影响. 事实上, $\mathcal{R}(w)$ 是 Locally weighted linear regression 的损失函数.

该方法的主要思路为: 对于每一个测试数据的预测结果, 其附近的训练数据对其的影响应该比远处的训练数据对其的影响大.

在实现中, 对每个输入 x , 我们都根据某个分布 (比如高斯分布) 得到单独的 r , 再通过 r 在训练数据上计算一个单独的 w , 最终求出 \hat{y} .

习题 2-3

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

习题 2-4

$$\begin{aligned}
 w^* &= \arg \min_w \mathcal{R}_D^{\text{struct}}(w) \\
 &= \arg \min_w \mathcal{R}_D^{\text{emp}}(w) + \frac{1}{2}\lambda\|w\|^2 \\
 &= \arg \min_w \frac{1}{2}\|Y - X^T w\|^2 + \frac{1}{2}\lambda\|w\|^2
 \end{aligned}$$

then we get $\mathcal{R}(w) = \frac{1}{2}\|Y - X^T w\|^2 + \frac{1}{2}\lambda\|w\|^2$

$$\begin{aligned}
 \nabla_w \mathcal{R}(w) &= \nabla_w \left(\frac{1}{2}\|Y - X^T w\|^2 + \frac{1}{2}\lambda\|w\|^2 \right) \\
 &= (Y - X^T w)^T (-X^T) + \lambda w^T \\
 &= w^T X X^T - Y^T X^T + \lambda w^T
 \end{aligned}$$

let $\nabla_w \mathcal{R}(w) = 0$, then

$$\begin{aligned}
 w^T X X^T - Y^T X^T + \lambda w^T &= 0 \\
 w^T X X^T + \lambda w^T &= Y^T X^T \\
 w^T (X X^T + \lambda I) &= Y^T X^T \\
 w^T &= Y^T X^T (X X^T + \lambda I)^{-1} \\
 w &= (X X^T + \lambda I)^{-1} X Y
 \end{aligned}$$

习题 2-5

we have $y \sim \mathcal{N}(w^T x, \beta)$, then

$$\begin{aligned}
 p(Y; w^T X, \beta) &= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\beta}} \exp\left(-\frac{(y^{(n)} - w^T x^{(n)})^2}{2\beta}\right) \\
 \log p(Y; w^T X, \beta) &= \log \prod_{n=1}^N \frac{1}{\sqrt{2\pi\beta}} \exp\left(-\frac{(y^{(n)} - w^T x^{(n)})^2}{2\beta}\right) \\
 &= \sum_{n=1}^N \left(-\frac{(y^{(n)} - w^T x^{(n)})^2}{2\beta} - \log \sqrt{2\pi\beta}\right) \\
 \nabla_w \log p(Y; w^T X, \beta) &= \nabla_w \sum_{n=1}^N \left(-\frac{(y^{(n)} - w^T x^{(n)})^2}{2\beta} - \log \sqrt{2\pi\beta}\right) \\
 &= \nabla_w \sum_{n=1}^N \left(-\frac{(y^{(n)} - w^T x^{(n)})^2}{2\beta}\right) \\
 &= -\frac{1}{2\beta} \nabla_w (Y - X^T w)^T (Y - X^T w) \\
 &= -\frac{1}{2\beta} \nabla_w (Y^T Y - 2Y^T X^T w + w^T X X^T w) \\
 &= -\frac{1}{\beta} (-Y^T X^T + w^T X X^T)
 \end{aligned}$$

let $\nabla_w \log p(Y; w^T X, \beta) = 0$, then

$$\begin{aligned}
 w^T X X^T &= Y^T X^T \\
 w^T &= Y^T X^T (X X^T)^{-1} \\
 w &= (X X^T)^{-1} X Y
 \end{aligned}$$

习题 2-6

(1)

we have $(x^{(1)}, \dots, x^{(n)}) \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned}
 p(X; \mu, \sigma) &= \prod_{n=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2}\right) \\
 \log p(X; \mu, \sigma) &= \log \prod_{n=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2}\right) \\
 &= \sum_{n=1}^N \left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2} - \log \sigma\sqrt{2\pi}\right) \\
 \nabla_{\mu} \log p(X; \mu, \sigma) &= \nabla_{\mu} \sum_{n=1}^N \left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2} - \log \sigma\sqrt{2\pi}\right) \\
 &= -\frac{1}{2\sigma^2} \nabla_{\mu} \sum_{n=1}^N ((x^{(n)} - \mu)^2) \\
 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x^{(n)} - \mu)
 \end{aligned}$$

let $\nabla_{\mu} \log p(X; \mu, \sigma) = 0$, then

$$\begin{aligned}
 \frac{1}{\sigma^2} \sum_{n=1}^N (x^{(n)} - \mu) &= 0 \\
 \sum_{n=1}^N x^{(n)} &= N\mu \\
 \mu^{ML} &= \frac{1}{N} \sum_{n=1}^N x^{(n)}
 \end{aligned}$$

(2)

we have

$$\begin{aligned}
 \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\
 x &\sim \mathcal{N}(\mu, \sigma^2)
 \end{aligned}$$

then

$$\begin{aligned}
p(\mu|X; \sigma, \mu_0, \sigma_0) &= \frac{p(X|\mu; \sigma)p(\mu; \mu_0, \sigma_0)}{p(X; \mu, \sigma)} \\
&\propto p(X|\mu; \sigma)p(\mu; \mu_0, \sigma_0) \\
&= \prod_{n=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2}\right) \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \\
\log p(\mu|X; \sigma, \mu_0, \sigma_0) &= \log \prod_{n=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2}\right) \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \\
&= \sum_{n=1}^N \left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2} - \log \sigma\sqrt{2\pi}\right) + \left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \log \sigma_0\sqrt{2\pi}\right) \\
\nabla_\mu \log p(\mu|X; \sigma, \mu_0, \sigma_0) &= \nabla_\mu \sum_{n=1}^N \left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2} - \log \sigma\sqrt{2\pi}\right) + \left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \log \sigma_0\sqrt{2\pi}\right) \\
&= \nabla_\mu \sum_{n=1}^N \left(-\frac{(x^{(n)} - \mu)^2}{2\sigma^2}\right) + \left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \\
&= \sum_{n=1}^N \left(\frac{1}{\sigma^2}(x^{(n)} - \mu)\right) + \left(-\frac{1}{\sigma_0^2}(\mu - \mu_0)\right)
\end{aligned}$$

let $\nabla_\mu \log p(\mu|X; \sigma, \mu_0, \sigma_0) = 0$, then

$$\begin{aligned}
0 &= \sum_{n=1}^N \left(\frac{1}{\sigma^2}(x^{(n)} - \mu)\right) + \left(-\frac{1}{\sigma_0^2}(\mu - \mu_0)\right) \\
\frac{1}{\sigma^2} \sum_{n=1}^N x^{(n)} - \frac{1}{\sigma^2} N\mu &= \frac{1}{\sigma_0^2} \mu - \frac{1}{\sigma_0^2} \mu_0 \\
\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \mu &= \frac{1}{\sigma^2} \sum_{n=1}^N x^{(n)} + \frac{1}{\sigma_0^2} \mu_0 \\
\mu^{MAP} &= \frac{\frac{1}{\sigma^2} \sum_{n=1}^N x^{(n)} + \frac{1}{\sigma_0^2} \mu_0}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}
\end{aligned}$$

习题 2-7

$$\begin{aligned}
\lim_{N \rightarrow \infty} \mu^{MAP} &= \lim_{N \rightarrow \infty} \frac{\frac{1}{\sigma^2} \sum_{n=1}^N x^{(n)} + \frac{1}{\sigma_0^2} \mu_0}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \\
&= \lim_{N \rightarrow \infty} \frac{\sum_{n=1}^N x^{(n)} + \frac{\sigma^2}{\sigma_0^2} \mu_0}{N + \frac{\sigma^2}{\sigma_0^2}} \\
&= \lim_{N \rightarrow \infty} \frac{\sum_{n=1}^N x^{(n)}}{N + \frac{\sigma^2}{\sigma_0^2}} + \frac{\frac{\sigma^2}{\sigma_0^2} \mu_0}{N + \frac{\sigma^2}{\sigma_0^2}} \\
&\approx \frac{1}{N} \sum_{n=1}^N x^{(n)}
\end{aligned}$$

习题 2-8

$$\begin{aligned}
\mathcal{R}(f) &= \mathbb{E}_{(x,y) \sim p_r(y|x)} [(y - f(x))^2] \\
&= \int \int (y - f(x))^2 p_r(x, y) dx dy \\
&= \int \int (y - f(x))^2 p_r(y|x) p_r(x) dx dy \\
&= \mathbb{E}_{x \sim p_r(x)} [\mathbb{E}_{y \sim p_r(y|x)} [y^2 - 2yf(x) + f^2(x)]] \\
&= \mathbb{E}_{x \sim p_r(x)} [\mathbb{E}_{y \sim p_r(y|x)} [y^2] - 2\mathbb{E}_{y \sim p_r(y|x)} [y] f(x) + f^2(x)] \\
&\geq \mathbb{E}_{x \sim p_r(x)} [\mathbb{E}_{y \sim p_r(y|x)}^2 [y] - 2\mathbb{E}_{y \sim p_r(y|x)} [y] f(x) + f^2(x)] \\
&= \mathbb{E}_{x \sim p_r(x)} [(\mathbb{E}_{y \sim p_r(y|x)} [y] - f(x))^2]
\end{aligned}$$

let $\mathcal{R}(f) \geq 0$, then

$$\begin{aligned}
(\mathbb{E}_{y \sim p_r(y|x)} [y] - f(x))^2 &\geq 0 \\
f^*(x) &= \mathbb{E}_{y \sim p_r(y|x)} [y]
\end{aligned}$$

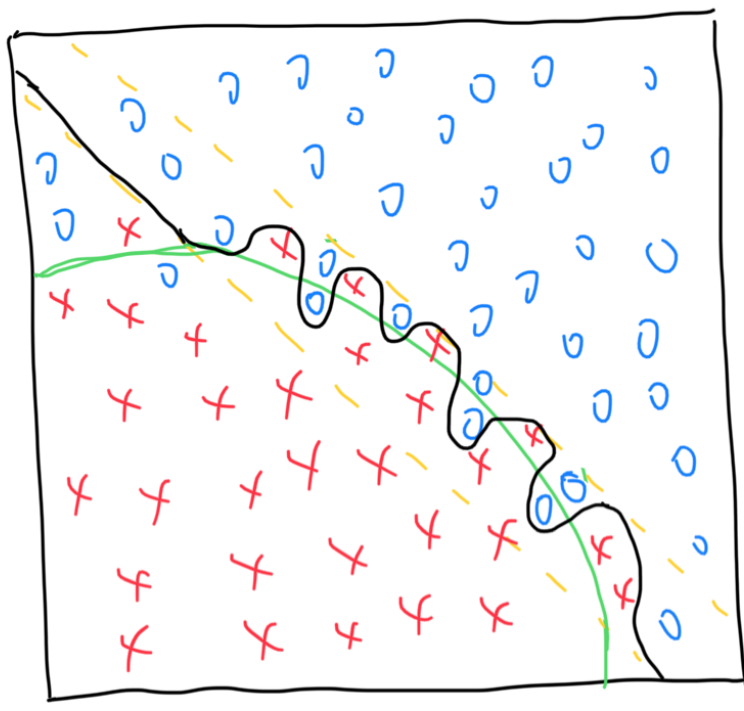


图 1: 高偏差 + 高方差

习题 2-9

高偏差: 模型具有很强的 (错误的) 先验知识 (偏见)

高方差: 模型过于依赖训练数据

高偏差 + 高方差: 即有很强的先验知识, 又过于依赖数据

如图1所示, 该图描述了一个二分类问题. 其中, 红 X 和蓝 O 是需区分的两类, 绿色的曲线为理想的决策曲线, 黄色的线段为人为规定的决策区域 (模型只能在其中进行决策), 黑线为模型最终训练得到的决策曲线. 在这个模型中, 人为规定的决策区域给模型提供了一个很强的先验 (偏见), 而模型在其中又过分拟合每个训练数据. 最终, 得到了一个高偏差 + 高方差的模型.

习题 2-10

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

习题 2-11

一元

特征单元: 我, 打了, 张三

$$v_1 = [1, 1, 1]^T$$

$$v_2 = [1, 1, 1]^T$$

二元

特征单元: \$ 我, 我打了, 打了张三, 张三 #, \$ 张三, 张三打了, 打了我, 我 #

$$v_1 = [1, 1, 1, 1, 0, 0, 0, 0]^T$$

$$v_2 = [0, 0, 0, 0, 1, 1, 1, 1]^T$$

三元

特征单元: \$ 我打了, 我打了张三, 打了张三 #, \$ 张三打了, 张三打了我, 打了我 #

$$v_1 = [1, 1, 1, 0, 0, 0]^T$$

$$v_2 = [0, 0, 0, 1, 1, 1]^T$$

习题 2-12

查准率

$$\mathcal{P}_c = \frac{TP_c}{TP_c + FP_c}$$

$$\mathcal{P}_1 = \frac{1}{1+1} = \frac{1}{2}$$

$$\mathcal{P}_2 = \frac{2}{2+2} = \frac{1}{2}$$

$$\mathcal{P}_3 = \frac{2}{2+1} = \frac{2}{3}$$

查全率

$$\mathcal{R}_c = \frac{TP_c}{TP_c + FN_c}$$

$$\mathcal{R}_1 = \frac{1}{1+1} = \frac{1}{2}$$

$$\mathcal{R}_2 = \frac{2}{2+1} = \frac{2}{3}$$

$$\mathcal{R}_3 = \frac{2}{2+2} = \frac{1}{2}$$

F1 值

$$\mathcal{F}_c = \frac{(1+1^2) * \mathcal{P}_c * \mathcal{R}_c}{1^2 * \mathcal{P}_c + \mathcal{R}_c}$$

$$\mathcal{F}_1 = \frac{2 * \frac{1}{2} * \frac{1}{2}}{1 * \frac{1}{2} + \frac{1}{2}} = \frac{1}{2}$$

$$\mathcal{F}_2 = \frac{2 * \frac{1}{2} * \frac{2}{3}}{1 * \frac{1}{2} + \frac{2}{3}} = \frac{4}{5}$$

$$\mathcal{F}_3 = \frac{2 * \frac{2}{3} * \frac{1}{2}}{1 * \frac{2}{3} + \frac{1}{2}} = \frac{4}{7}$$

宏平均

$$\begin{aligned}\mathcal{P}_{macro} &= \frac{1}{C} \sum_{c=1}^C \mathcal{P}_c = \frac{1}{3} \left(\frac{1}{2} + \frac{1}{2} + \frac{2}{3} \right) = \frac{5}{9} \\ \mathcal{R}_{macro} &= \frac{1}{C} \sum_{c=1}^C \mathcal{R}_c = \frac{1}{3} \left(\frac{1}{2} + \frac{2}{3} + \frac{1}{2} \right) = \frac{5}{9} \\ \mathcal{F1}_{macro} &= \frac{2 * \mathcal{P}_{macro} * \mathcal{R}_{macro}}{\mathcal{P}_{macro} + \mathcal{R}_{macro}} = \frac{2 * \frac{5}{9} * \frac{5}{9}}{\frac{5}{9} + \frac{5}{9}} = \frac{5}{9}\end{aligned}$$

微平均

$$\begin{aligned}\mathcal{P}_{micro} &= \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C TP_c + \sum_{c=1}^C FP_c} = \frac{5}{5+4} = \frac{5}{9} \\ \mathcal{P}_{micro} &= \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C TP_c + \sum_{c=1}^C FN_c} = \frac{5}{5+4} = \frac{5}{9} \\ \mathcal{F1}_{micro} &= \frac{2 * \mathcal{P}_{micro} * \mathcal{R}_{micro}}{\mathcal{P}_{micro} + \mathcal{R}_{micro}} = \frac{2 * \frac{5}{9} * \frac{5}{9}}{\frac{5}{9} + \frac{5}{9}} = \frac{5}{9}\end{aligned}$$