数据科学笔记

高金成　2023233003

二〇二四年三月十一日

前言

这本是用的是周志华的《机器学习理论导引》

一些参考资料

GitHub 上大佬的补充概念

这个博主做过部分习题

我并不建议大家钻研这本书, 对于数学系来说, 这本书尽管看起来严谨, 但是却还不够严谨, 我身边的同学在阅读本书的过程中很容易钻牛角尖.

对于计算机系同学来说, 这些概念和理论推导, 似乎对帮助理解算法, 或者可解释性方面帮助不大, 这本书更像是从数值优化收敛性角度出发的. 食之无味.

而且习题很难, 且对科研和学习帮助不大.

本文习题答案仅作参考, 我想大部分都是不严格的, 甚至是不对的. 仅仅作为思路供大家参考.

1 第一章习题答案

例题 1.1

试分析下面这个函数的凸性

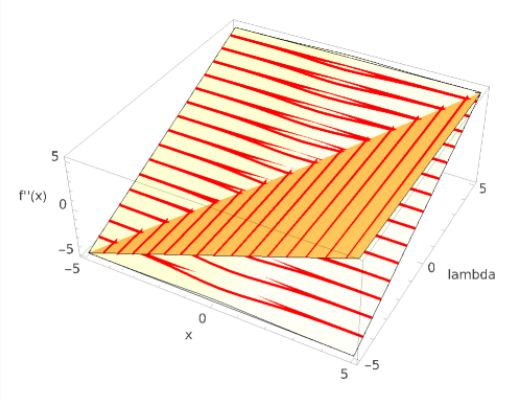
$$f(x) = \log(1 + e^{-x}) + \frac{\lambda}{2}x^2$$

该函数的二阶导函数为:

$$f''(x) = \frac{e^{-x}}{(1 + e^{-x})^2} + \lambda$$

1. 当 $\lambda > 0$ 时,$f''(x)$ 显然是正的, 所以是凸函数.
2. 当 $\lambda = 0$ 时,$f''(x) = \frac{e^{-x}}{(e^{-x}+1)^2}$ 也是正的, 所以是凸函数.
3. 当 $\lambda < 0$ 时,$f''(x)$ 的正负无法判断, 因此无法判断原函数的凸性.

给出 $\lambda, x, f(x)$ 的 3D 图像



例题 1.2

试推导下面这个函数的共轭函数 $f(x) = log(1 + e^{-x})$

共轭函数 $f^*(y)$ 的定义是:

$$f^*(y) = \sup_{x \in \mathrm{dom} f} (xy - f(x))$$

给定函数 $f(x) = \log(1 + e^{-x})$，需要找到 $f^*(y)$。

构建优化问题:

$$f^*(y) = \sup_x \left(xy - \log(1 + e^{-x})\right)$$

为了找到 $f^*(y)$，需要求解优化问题的最优解。这通常通过求解一阶条件（KKT 条件）来完成。

一阶导数:

$$\frac{d}{dx}\left(xy - \log(1 + e^{-x})\right) = y + \frac{1}{1 + e^x}$$

如果 $y \geq 0$, 一阶导数恒大于 0, 那么 $f^*(y) \to +\infty$
如果 $y \leq -1$, 一阶导数恒小于 0, 当 $y = -1$ 时,$f^*(y) = 0$, 当 $y < -1$ 时,$f^*(y) \to +\infty$
如果 $-1 < y < 0$, 令一阶导数等于 0, 解得 $x = \log \frac{1-y}{y}$, 代入原式得到 $f^*(y) = y \log \frac{1-y}{y} + \log y$

例题 1.3

试基于 Markov 不等式给出 Chebyshev 不等式和 Cantelli 不等式的证明

Markov 不等式

$$P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}$$

Chebyshev 不等式的证明考虑 $Y = (X - E[X])^2$。利用 Markov 不等式，得到：

$$P(Y \geq \varepsilon^2) \leq \frac{E[Y]}{\varepsilon^2}$$
$$P((X - E[X])^2 \geq \varepsilon^2) \leq \frac{V[X]}{\varepsilon^2}$$

证毕
Cantelli 不等式的证明:
设 $Y = X - E[X]$，则对于所有 $\lambda \geq 0$ 有：

$$\begin{aligned} P(X - E[X] \geq \varepsilon) &= P(Y + \lambda \geq \varepsilon + \lambda) \\ &= P((Y + \lambda)^2 \geq (\varepsilon + \lambda)^2) \\ &\leq \frac{E[(Y + \lambda)^2]}{(\varepsilon + \lambda)^2} = \frac{V[X] + \lambda^2}{(\varepsilon + \lambda)^2} \end{aligned}$$

通过对上界优化，上式中最优的 $\lambda = \frac{V[X]}{\varepsilon}$ 可得到：

$$P(X - E[X] \geq \varepsilon) \leq \frac{V[X]}{V[X] + \varepsilon^2}$$

证毕

例题 1.4

试给出 Bernstein 不等式以概率 $1 - \delta (0 < \delta < 1)$ 成立的表达形式

首先，我们回顾 Bernstein 不等式，其表达形式为：

$$P(\bar{X} \geq E[X] + \varepsilon) \leq \exp\left(-\frac{m\varepsilon^2}{2V[X] + 2b\varepsilon}\right)$$

为了将其转化为概率 $1 - \delta$ 的形式，我们将不等式右边设为 δ，从而得到：

$$\exp\left(-\frac{m\varepsilon^2}{2V[X] + 2b\varepsilon}\right) = \delta$$

对上述方程取自然对数，得到：

$$-\frac{m\varepsilon^2}{2V[X]+2b\varepsilon}=\ln(\delta)$$

重排此方程，我们得到 ε 的表达式：

$$\varepsilon^2+\frac{2b\ln(\delta)}{m}\varepsilon+\frac{2V[X]\ln(\delta)}{m}=0$$

解上述方程将给出与给定的 δ 相对应的 ε。

$$\varepsilon=\sqrt{\frac{b^2\ln^2\delta}{m^2}-\frac{2V[x]\ln\delta}{m}}-\frac{b\ln\delta}{m}$$

那么就有了 $1-\delta$ 概率下

$$\bar{X}\leq E[X]+\sqrt{\frac{b^2\ln^2\delta}{m^2}-\frac{2V[x]\ln\delta}{m}}-\frac{b\ln\delta}{m}$$

例题 1.5

试基于 MCDiarmid 不等式证明 Hoeffding 不等式

首先，回顾 McDiarmid 不等式：

$$P\left(f(X_1,\ldots,X_m)-\mathbb{E}[f(X_1,\ldots,X_m)]\geq\varepsilon\right)\leq e^{-2\varepsilon^2/\sum_{i=1}^m c_i^2}$$

其中，对所有的 $i\in[m]$ 和所有的 $x_1,\ldots,x_m,x_i'\in\chi$，满足

$$|f(x_1,\ldots,x_i,\ldots,x_m)-f(x_1,\ldots,x_i',\ldots,x_m)|\leq c_i$$

考虑 $X_1,X_2,\ldots,X_m$ 是独立同分布的随机变量，其取值在 $[0,1]$ 上，其均值为

$$\bar{X}=\frac{1}{m}\sum_{i=1}^m X_i$$

我们希望利用 McDiarmid 不等式来得到 Hoeffding 不等式。

对于函数

$$f(x_1,x_2,\ldots,x_m)=\frac{1}{m}\sum_{i=1}^m x_i$$

任意两组数 x_i 和 x_i' 之间的差异最大为 $1/m$。因此，对于所有 i，有 $c_i=1/m$。

代入 McDiarmid 不等式，我们有：

$$P(\bar{X}-\mathbb{E}[\bar{X}]\geq\varepsilon)\leq e^{-2m\varepsilon^2}$$

证毕

例题 1.6

试证明 0/1 损失函数 $\ell l_{0/1}(x)=\mathbb{I}(x<0)$ 非凸，而 hinge 损失函数 $\ell l_{hinge}(x)=\max(0,1-x)$ 是凸函数

0/1 损失函数的非凸性证明:

0/1 损失函数定义为:

$$\ell_{0/1}(x) = \mathbb{I}(x < 0)$$

这是一个分段常数函数，其在 $x = 0$ 处是不连续的。因此，它不能是凸函数。

Hinge 损失函数的凸性证明:

Hinge 损失函数定义为:

$$\ell_{\text{hinge}}(x) = \max(0, 1 - x)$$

要证明这是一个凸函数，我们可以使用二阶导数判断。如果一个函数的二阶导数在其定义域上都是非负的，则该函数是凸的。

对 $\ell_{\text{hinge}}(x)$ 求一阶导:

$$\text{当 } x < 1, \quad \ell'_{\text{hinge}}(x) = -1$$
$$\text{当 } x > 1, \quad \ell'_{\text{hinge}}(x) = 0$$

在 $x = 1$，该函数不可导。

对其求二阶导，得:

$$\ell''_{\text{hinge}}(x) = 0 \quad \text{除了在} x = 1 \text{这一不可导点}$$

由于二阶导数在其定义域上都是非负的，所以 Hinge 损失函数是凸的。

例题 1.7

给定训练集 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, 其中 $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$. 对率回归的优化问题为:

$$\min_{\omega \in \mathbb{R}^d} \sum_{i=1}^{m} \log\left(1 + e^{-y_i x_i^T \omega}\right) + \frac{\lambda}{2} \|\omega\|_2^2$$

其中 $\lambda > 0$ 是超参数, 试推导上述问题的对偶问题

首先，定义拉格朗日函数 $L(\omega, \alpha)$ 为:

$$L(\omega, \alpha) = \sum_{i=1}^{m} \log\left(1 + e^{-y_i x_i^T \omega}\right) + \frac{\lambda}{2} \|\omega\|_2^2$$

其中 α 是拉格朗日乘子, 但是本题没有约束, 所以没有拉格朗日乘子.

然后，对偶函数 $g(\alpha)$ 定义为:

$$g(\alpha) = \inf_{\omega} L(\omega, \alpha)$$

对偶问题则为:

$$\max_{\alpha} \quad g(\alpha) = \inf_{\omega} L(\omega)$$

$$s.t. \quad \alpha \geq 0$$

所以该对偶问题只需要寻找 $L(\omega)$ 的下确界就好了.

(笔者注: 这种无约束优化的对偶问题似乎没什么意义的样子, 一般用梯度下降直接求解)

例题 1.8

给定数据集 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, 其中 $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$, 其中 $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$. 软间隔支持向量机的优化问题

$$\begin{aligned} \min_{\omega, b, \xi_i} & \frac{1}{2} \|\omega\|^2 + \beta \sum_{i=1}^{m} \xi_i \\ s.t. \quad & y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi \geq 0 \quad (i \in [m]) \end{aligned}$$

缺省地认为: 无论错误发生在哪一类样本上, 所需付出的" 代价" 是相同的. 然而在现实应用中, 不同类别的错误代价往往不同. 不妨假设正例出错的代价是反例出错代价的 k 倍, 试给出相应的软间隔支持向量机的优化问题和对偶问题.

优化问题:

在考虑不同类别的错误代价时，我们可以引入两个不同的惩罚参数 β_1 和 β_2，其中 β_1 是正例出错的代价，β_2 是反例出错的代价。假设正例出错的代价是反例出错代价的 k 倍，即 $\beta_1 = k\beta_2$。

优化问题可以重新定义为:

$$\begin{aligned} \min_{\omega, b, \xi_i} & \frac{1}{2} \|\omega\|^2 + \sum_{i=1}^{m} \beta_i \xi_i \\ \text{s.t.} \quad & y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad (i \in [m]) \end{aligned}$$

其中当 $y_i = +1$ 时 $\beta_i = \beta_1$ ，当 $y_i = -1$ 时 $\beta_i = \beta_2$。

对偶问题:

对于这个优化问题，其拉格朗日函数为:

$$L(\omega, b, \xi, \alpha, \mu) = \frac{1}{2} \|\omega\|^2 + \sum_{i=1}^{m} \beta_i \xi_i - \sum_{i=1}^{m} \alpha_i \left(y_i(\omega^T \phi(x_i) + b) - 1 + \xi_i \right) - \sum_{i=1}^{m} \mu_i \xi_i$$

其中 $\alpha_i \geq 0$ 和 $\mu_i \geq 0$ 是拉格朗日乘子。

对偶问题为:

$$\max_{\alpha, \mu} \min_{\omega, b, \xi} L(\omega, b, \xi, \alpha, \mu)$$

满足 $0 \leq \alpha_i \leq \beta_i$ 和 $\mu_i \geq 0$。

整理一下得到最后的结果为:

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{m} (\beta_i - \mu_i - \alpha_i) \xi_i$$

满足:

$$0 \leq \alpha_i \leq \beta_i$$
$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

其中，当 $y_i = +1$ 时，$\beta_i = \beta_1$，当 $y_i = -1$ 时，$\beta_i = \beta_2$。

2 第二章习题答案

例题 2.1

证明: 若训练集 D 含有 m 个从分布 D 上独立同分布抽取的样本，且

$$0 < \varepsilon < 1$$

对任意假设 $h \in \mathcal{H}$，有

$$P\left(\hat{E}(h) - E(h) \geq \varepsilon\right) \leq \exp(-2m\varepsilon^2)$$
$$P\left(E(h) - \hat{E}(h) \geq \varepsilon\right) \leq \exp(-2m\varepsilon^2)$$
$$P\left(|\hat{E}(h) - E(h)| \geq \varepsilon\right) \leq 2\exp(-2m\varepsilon^2)$$

设：存在随机变量 X_i 的 $I(h(x_i) \neq y_i)$，则 $\hat{E}(h) = \frac{1}{m}\sum_{i=1}^{m} I(h(x_i) \neq y_i) = X$，

$$E(h) = E_{(x,y)\sim D} I(h(x_i) \neq y_i) = E[X]$$

由 Hoeffding 不等式:

$$P(X - E[X] > \varepsilon) \leq e^{-2m\varepsilon^2}$$
$$P(X - E[X] < -\varepsilon) \leq e^{-2m\varepsilon^2}$$

代入得到:

$$P(\hat{E}(h) - E(h) > \varepsilon) \leq e^{-2m\varepsilon^2}$$
$$P(E(h) - \hat{E}(h) > \varepsilon) \leq e^{-2m\varepsilon^2}$$
$$P(|\hat{E}(h) - E(h)| > \varepsilon) = P(\hat{E}(h) - E(h) > \varepsilon) \cup P(E(h) - \hat{E}(h) > \varepsilon)$$
$$\leq 2e^{-2m\varepsilon^2}$$

例题 2.2

令 $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ 表示一组阈值函数 $h_a(x) = \mathbb{I}(x < a)$ 形成的假设空间。假设目标概念 $c \in \mathcal{H}$。试证明这个无限假设空间 $\mathcal{H}$ 是 PAC 可学的。

设: 设 R 为目标概念的区间，R^D 为与目标所选正例的最小区间若 R 的错误率是 $P(R) < \varepsilon$，则 R^D 错误率不超过 ε 若 $P(R) > \varepsilon$，对 r 为 a 的左邻或右邻，错误率是少于 ε 如果 R^D 和 r 相交，则 r 是错误的。R^D 的所有这样的错误邻域，即 $E(R^D) < \varepsilon$ 如果 $E(R^D) > \varepsilon$，则 r 与 R^D 不相交，$P(E(R^D) > \varepsilon) \leq (1-\varepsilon)^m \leq e^{-m\varepsilon}$ 因为 $e^{-m\varepsilon} < \delta$，得 $P(E(R^D) < \varepsilon) = 1 - P(E(R^D) > \varepsilon) \geq 1 - \delta$

由此得到:

$$m \geq -\frac{1}{\varepsilon}\ln\frac{1}{\delta}$$

故: 假设空间 $\mathcal{H}$ 是 PAC 可学的

例题 2.3

令 $\mathcal{H}=\{h_r : r \in \mathbb{R}_+\}$ 表示函数 $h_r(x)=\mathbb{I}(||x|| \leq r)$ 形成的假设空间。假设目标概念 $c \in \mathcal{H}$。试证明 $\mathcal{H}$ 是 PAC 可学的。

设目标概念 $c \in \mathcal{H}_\iota$。试证明 $\mathcal{H}$ 是 PAC 可学的。

设：设 R 为目标概念的区间，R^D 与之有所交叉，且为在所选正例上的最小区间。若 R 的错误率为 $P(R)<\varepsilon$，则 R^D 错误率不超过 ε。若 $P(R)>\varepsilon$，设 r 为 R 的邻近区间，且与 R 之间无更小的目标概念错误率 e。如果 R^D 与 r 相交，则 r 是错误的。R^D 的所有这样的错误邻域，即 $E(R^D)<\varepsilon$。如果 $E(R^D)>\varepsilon$，则 r 与 R^D 不相交，$P(E(R^D)>\varepsilon) \leq (1-\varepsilon)^m \leq e^{-m\varepsilon}$。因为 $e^{-m\varepsilon}<\delta$，有 $P(E(R^D)<\varepsilon)=1-P(E(R^D)>\varepsilon) \geq 1-\delta$。

由此得到：

$$m \geq -\frac{1}{\varepsilon}\ln\frac{1}{\delta}$$

故：假设空间 $\mathcal{H}$ 是 PAC 可学的。

例题 2.4

试证明实数集合 $\mathbb{R}$ 上任意两个闭区间之并（形如 $[a_1,a_2]\cup[b_1,b_2]$）作为概念形成的概念类是 PAC 可学的。

设：区间 $[a_1,a_2]$ 和 $[b_1,b_2]$ 对实例上的概念集 D 分别不相交，分别为区间-正区间-负区间，此概念类也不两个区间的概念不为集中的最小假设。若有错误率最大的两个区间 PAC 可学的概念，将各个区间划分为 $\frac{\varepsilon}{4}$ 的区域，得到区间 PAC 可学的结论可得：

$$P(E(R^D)>\varepsilon) \leq 4(1-\varepsilon)^m \leq 4e^{-m\varepsilon/4}$$

$$\Rightarrow 4e^{-m\varepsilon/4}<\delta, \text{则} P(E(R^D)<\varepsilon)=1-P(E(R^D)>\varepsilon) \geq 1-\delta$$

由此得到：

$$m \geq \frac{4}{\varepsilon}\ln\frac{4}{\delta}$$

所以假设空间上的概念在所选实例上的假设是 PAC 可学的。

3 第三章习题答案

例题 3.1

试证明:

1. 轴平行矩形的假设空间的 VC 维为 4
2. R^d 中轴平行多面体的假设空间 $H = \{h(a,b)(x) : a_i \leq b_i, i \in [d]\}$ 的 VC 维为 $2d(d > 2)$
3. 决策树分类器的假设空间的 VC 维可以为无穷大
4. 1− 近邻分类假设空间的 VC 维可以为无穷大

1. 轴平行矩形的假设空间的 VC 维为 4。

证明：设数据集 $D = \{(1,1),(1,-1),(-1,1),(-1,-1)\}$，我们可以找到轴平行矩形 h，使得 $h(x) = 1$ 当 $x \in \{(1,1),(1,-1)\}$，并且 $h(x) = -1$ 当 $x \in \{(-1,1),(-1,-1)\}$。因此，数据集 D 能被轴平行矩形的假设空间打散。但是，对于任意包含 5 个或更多点的数据集，我们总可以找到一个点 x，使得它被包含在由其他点构成的最小轴平行矩形中。因此，该点的标签不能被正确分类。所以轴平行矩形的假设空间的 VC 维是 4。

2. R^d 中轴平行多面体的假设空间 $H = \{h(a,b)(x) : a_i \leq b_i, i \in [d]\}$ 的 VC 维为 $2d(d > 2)$。

证明：对于每个维度，我们可以选择 2 个平面来构成一个轴平行的多面体。因此，对于 d 维空间，我们有 $2d$ 个平面。由于多面体的每一面都可以是边界，所以我们至少需要 $2d$ 个点来确定一个多面体。因此，VC 维至少为 $2d$。

3. 决策树分类器的假设空间的 VC 维可以为无穷大。

证明：决策树的 VC 维与其叶子结点的个数有关。当叶子结点数恰好等于数据集的大小时，决策树的假设空间能够打散该数据集。由于我们可以不限制决策树的叶子结点的数量，因此其 VC 维可以为无穷大。

4. 1− 近邻分类假设空间的 VC 维可以为无穷大。

证明：对于 1− 近邻分类器，其决策边界由 Voronoi 图来表示。Voronoi 图能够将任意数据集划分为相邻的区域，每个区域中的点都比其他区域中的点更接近某个给定的数据点。由于 Voronoi 图能够适应任意的数据分布，因此 1− 近邻分类器的 VC 维可以为无穷大。

例题 3.2

考虑 VC 维为 d 的假设空间 $\mathcal{H}$, 其中 $h \in \mathcal{H} : \mathcal{X} \mapsto \{-1,+1\}$, 令 $\mathcal{M}$ 表示由 $\mathcal{H}$ 中任意 $k \geq 1$ 个假设依据多数投票法生成的假设所组成的假设空间, 即

$$\mathcal{M} = \left\{ h(x) = argmax_{y \in \{-1,+1\}} \sum_{i=1}^{k} \mathbb{I}(h_i(x) = y) : h_1, \cdots, h_k \in \mathcal{H} \right\}$$

若 $kd \geq 4$, 试证明 $\mathcal{M}$ 的 VC 维有上界 $(kd \ln(kd))$

证: 可以将 k 个假设的融合看成笛卡尔积, 即 $\mathcal{F} = \mathcal{F}^{(1)} \times \cdots \times \mathcal{F}^{(k)}$, 有:

$$
\begin{aligned}
II_{\mathcal{F}}(m) &\leq \prod_{i=1}^{k} II_{\mathcal{F}^{(i)}}(m) \\
&\leq \prod_{i=1}^{k} \left(\frac{e \cdot m}{d}\right)^{d} \\
&\leq (e \cdot m)^{kd}
\end{aligned}
$$

设 $\mathcal{M}$ 的 VC 维为 d', 有 $2^{d'} \leq (e \cdot d')^{kd}$
$d' \leq kd \log_2(e \cdot d') = kd \log_2 e + kd \log_2 d'$
$\frac{d'}{\log_2 d'} = O(kd)$
又 $\log_2 \frac{d'}{\log_2 d'} = \log_2 d' - \log_2 \log_2 d' = O(\log_2(kd))$
$\log_2 d' = O(\log_2(kd))$
$d' = \frac{d'}{\log_2 d'} \cdot \log_2 d' = O(kd \log_2(kd)) = O(kd \ln(kd))$

例题 3.3

考虑假设空间 $\mathcal{H}$, 其中 $h \in \mathcal{H} : \mathcal{X} \mapsto \{-1, +1\}$, 给定一个大小为 m 的集合 D, 试证明 $\mathcal{H}$ 关于 D 的经验 Rademacher 复杂度 $\hat{\mathcal{R}}_D(\mathcal{H})$

1. 若假设空间 $\mathcal{H}$ 只包含单个假设, 即 $\mathcal{H} = \{h\}$, 则有 $\hat{\mathcal{R}}_D(\mathcal{H}) = 0$
2. 若假设空间 $\mathcal{H}$ 满足 $|\mathcal{H}_{|D}| = 2^m$ 则有 $\hat{\mathcal{R}}_D(\mathcal{H}) = 1$

1. $\hat{\mathcal{R}}_D(\mathcal{H}) = E_\sigma \left[\frac{1}{m} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)\right] = \frac{1}{m} E_\sigma \left[\sum_{i=1}^{m} \sigma_i h(x_i)\right]$
$E[\sigma_i] = 0$
$\hat{\mathcal{R}}_D(\mathcal{H}) = 0$
2. 当 $h(x_i) = \sigma_i$ 时,$\sum_{i=1}^{m} \sigma_i h(x_i)$ 最大
$\hat{\mathcal{R}}_D(\mathcal{H}) = E_\sigma \left[\frac{1}{m} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)\right] = \frac{1}{m} E[\sum_{i=1}^{m} \sigma_i^2] = 1$

例题 3.4

令 $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ 表示一维阈值函数 $h_a(x) = \mathbb{I}(x \leq a)$ 构成的假设空间, 集合 D 包含实数轴上 m 个不同的点, 试证明 $\mathcal{H}$ 关于 D 的经验 Rademacher 复杂度 $\hat{\mathcal{R}}_D(\mathcal{H}) = O(\frac{1}{\sqrt{m}})$

$$
\begin{aligned}
\hat{\mathcal{R}}_D(\mathcal{H}) &= \frac{1}{m} E_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \mathcal{I}(x_i \le a) \right] \\
&\le \frac{1}{m} E_\sigma \left[\left| \sum_{i=1}^{m} \sigma_i (x_i \le a) \right| \right] \\
&\le \frac{1}{m} \left[E_\sigma \left[\left| \sum_{i=1}^{m} \sigma_i \mathbb{I}(x_i \le a) \right|^2 \right] \right]^{\frac{1}{2}} \\
&= \frac{1}{m} \left[E_\sigma \left[\sum_{i,j=1}^{m} \sigma_i \sigma_j \mathbb{I}(x_i \le a)^2 \mathbb{I}(x_j \le a) \right] \right]^{\frac{1}{2}} \\
&= \frac{1}{m} \left[\sum_{i=1}^{m} \mathbb{I}^2(x_i \le a) \right]^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{m}}
\end{aligned}
$$

$\hat{\mathcal{R}}_D(\mathcal{H}) = O(\frac{1}{\sqrt{m}})$

4 第四章习题答案

例题 4.1

试给出轴平行矩形假设空间基于 VC 维的泛化误差界, 并于 (2.23) 进行比较

VC 维度（轴平行矩形） $= \mathrm{d} = 4$, 根据公式 4.22 得到

$$\epsilon \leq \sqrt{\frac{16\ln(em/2) + 8\ln(4/\delta)}{m}}$$

2.23 为 $m \geq \frac{4}{\epsilon}\ln\frac{4}{\delta}$

例题 4.2

若假设空间 $\mathcal{H}$ 的 VC 维为 d

1. 试证明对任一大小为 m 的集合 $D, \hat{\mathscr{R}}_d(\mathcal{H}) \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$, 进一步, 对任一分布 $D, \mathscr{R}_m(\mathcal{H}) \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$
2. 试利用基于 Rademacher 复杂度的泛化误差界 (定理 4.5) 和 (1) 中的结果推到基于 VC 维的泛化误差界, 并与定理 4.3 进行比较

(1) 首先证明对任一大小为 m 的集合 $D, \hat{\mathscr{R}}_d(\mathcal{H}) \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$。

由 VC 维的定义，我们知道对任意大小为 d 的数据集 D，假设空间 $\mathcal{H}$ 可以打散 D，即 $\mathcal{H}$ 可以实现 D 的所有可能的标签组合。根据 Sauer 引理，对于大小为 m 的数据集 D，假设空间 $\mathcal{H}$ 的打散能力被限制在 $\sum_{i=0}^{d}\binom{m}{i}$ 中。

因此，我们可以得到

$$\hat{\mathscr{R}}_d(\mathcal{H}) = \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i) \leq \sqrt{\frac{2\ln(\sum_{i=0}^{d}\binom{m}{i})}{m}}$$

使用组合数的上界估计 $\binom{m}{i} \leq (\frac{em}{i})^i$，我们有

$$\begin{aligned}
\hat{\mathscr{R}}_d(\mathcal{H}) &\leq \sqrt{\frac{2\ln(\sum_{i=0}^{d}(\frac{em}{i})^i)}{m}} \\
&\leq \sqrt{\frac{2\ln(\sum_{i=0}^{d}(\frac{em}{d})^d)}{m}} \\
&= \sqrt{\frac{2\ln((d+1)(\frac{em}{d})^d)}{m}} \\
&\leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}
\end{aligned}$$

因此，我们得到对任一大小为 m 的集合 $D, \hat{\mathscr{R}}_d(\mathcal{H}) \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$。

接下来，我们证明对任一分布 $D, \mathscr{R}_m(\mathcal{H}) \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$。

由 Rademacher 复杂度的定义，我们有

$$\mathscr{R}_m(\mathcal{H}) = \mathbb{E}_{D^m}\hat{\mathscr{R}}_d(\mathcal{H}) \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$$

(2) 使用定理 4.5 和 (1) 中的结果，我们可以推导出基于 VC 维的泛化误差界。

对于任意 $h \in \mathcal{H}$，由定理 4.5 和 (1) 的结果，我们有

$$E(h) \leq \hat{E}(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}} \leq \hat{E}(h) + \sqrt{\frac{2d\ln(\frac{em}{d})}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

和

$$E(h) \leq \hat{E}(h) + \mathcal{R}_D(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \leq \hat{E}(h) + \sqrt{\frac{2d\ln(\frac{em}{d})}{m}} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

这就是基于 VC 维的泛化误差界。这个界比定理 4.3 中的界更加紧密，因为它考虑了假设空间的复杂度。

例题 4.3

若假设空间 $\mathcal{H}$ 满足 $|\mathcal{H}| \geq 3$, 试证明对于任意学习算法 $\mathcal{L}$ 存在分布 $\mathcal{D}$ 和目标概念 $c \in \mathcal{H}$, 使得至少需要 $\Omega(\frac{1}{\epsilon}\ln\frac{1}{\delta})$ 个样本才有

$$P(E_D(h_D, c) \leq \epsilon) \geq 1 - \delta$$

其中 h_D 为 $\mathcal{L}$ 基于从 $\mathcal{D}$ 独立同分布采样得到的训练集 D 输出的假设

霍夫丁不等式（Hoeffding's Inequality）是这样描述的：对于独立同分布的随机变量 $X_1, X_2, \ldots, X_m$，其中每个 X_i 取值于 [0, 1] 且它们的期望是 $E[X_i] = \mu$，有：

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right).$$

考虑到学习的问题，每个样本点上的错误可以被视为一个随机变量，其取值为 0（无误差）或 1（有误差）。应用霍夫丁不等式，我们可以得到：

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - E_D(h_i, c)\right| > \frac{\epsilon}{2}\right) \leq 2\exp\left(-m\frac{\epsilon^2}{2}\right).$$

为了保证上述的概率小于 δ，我们可以得到：

$$2\exp\left(-m\frac{\epsilon^2}{2}\right) \leq \delta.$$

从上式中解出 m，我们得到：

$$m \geq \frac{2}{\epsilon^2}\ln\left(\frac{2}{\delta}\right).$$

这意味着为了满足特定的误差 ϵ 和概率 δ，学习算法所需的样本数至少与 $\frac{1}{\epsilon}\ln\frac{1}{\delta}$ 成正比，也即：

$$m = \Omega\left(\frac{1}{\epsilon}\ln\frac{1}{\delta}\right).$$

例题 4.4

4.3 节分析实例中给出了二分类问题中支持向量机的泛化误差界. 对于多分类问题, 可以定义打分函数 $h(x,y): X \times Y \mapsto \mathbb{R}$ 实现分类结果 $argmax_{y\in\mathcal{Y}} h(x,y)$, 其中 $\mathcal{Y} = \{0, \cdots, K-1\}$

1. 定义打分函数 h 在点 (x,y) 处的间隔为 $\tau_h(x,y) = h(x,y) - max_{y'\neq y} h(x,y'), h$ 在 $D = \{(x_1,y_1), \cdots, (x_m,y_m)\}$ 上的经验间隔损失为 $\hat{E}_{D,\rho}(h) \leq \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\tau_h(x_i,y_i))$, 试证明:

$$\hat{E}_{D,\rho}(h) \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(\tau_h(x_i,y_i) \leq \rho)$$

2. 定义 $\tau_{\theta,h}(x,y) = \min_{y'\in\mathcal{Y}} = \min_{y'\in\mathcal{Y}}(h(x,y) - h(x,y') + \theta\mathbb{I}(y' \neq y))$, 其中 $\theta > 0$ 为任意常数,$(x,y) \sim \mathcal{D}$, 试证明:

$$\mathbb{E}_{\mathcal{D}}[\mathbb{I}(\tau_h(x,y) \leq 0)] \leq \mathbb{E}_{\mathcal{D}}[\mathbb{I}(\tau_{\theta,h}(x,y) \leq 0)]$$

3. 令 $\mathcal{H} \subset \mathbb{R}^{X\times Y}$ 表示函数 h 构成的集合, 固定 $\rho > 0$, 考虑假设空间 $\tilde{H} = \{(x,y) \mapsto \tau_{\theta,h}(x,y) : h \in \mathcal{H}\}$, 其中 $\theta = 2\rho, h$ 的泛化误差表示为 $E(h) = \mathbb{E}_D[\mathbb{I}(\tau_h(x,y) \leq 0)]$, 试证明对 $0 < \delta < 1, h \in \mathcal{H}$, 以至少 $1-\delta$ 的概率有

$$E(h) \leq \hat{E}_{D,\rho}(h) + \frac{2}{\rho}\mathscr{R}_m(\tilde{\mathcal{H}}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

1.

给定的经验间隔损失为

$$\hat{E}_{D,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\tau_h(x_i,y_i))$$

其中 $\Phi_\rho(t) = \max\{0, \rho - t\}$。

显然，如果 $\tau_h(x_i,y_i) \leq \rho$，那么 $\Phi_\rho(\tau_h(x_i,y_i)) \geq 1$。因此，有

$$\hat{E}_{D,\rho}(h) \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(\tau_h(x_i,y_i) \leq \rho)$$

其中 $\mathbb{I}$ 是指示函数，当给定条件为真时其值为 1，否则为 0。

2.

给定

$$\tau_{\theta,h}(x,y) = \min_{y'\in\mathcal{Y}}(h(x,y) - h(x,y') + \theta\mathbb{I}(y' \neq y))$$

如果 $\tau_h(x,y) \leq 0$, 那么存在某个 y' 使得 $h(x,y) \leq h(x,y')$, 即使我们为 $h(x,y)$ 添加一个 θ 的正偏差, $\tau_{\theta,h}(x,y)$ 仍然会小于或等于 0。因此，

$$\mathbb{E}_{\mathcal{D}}[\mathbb{I}(\tau_h(x,y) \leq 0)] \leq \mathbb{E}_{\mathcal{D}}[\mathbb{I}(\tau_{\theta,h}(x,y) \leq 0)]$$

3.

我们知道经验风险的一个界是

$$E(h) \leq \hat{E}_{D,\rho}(h) + \mathscr{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

其中 $\mathscr{R}_m(\mathcal{H})$ 是假设空间 $\mathcal{H}$ 在 m 个样本上的 Rademacher 复杂度。

给定新的假设空间 $\tilde{H}$，我们可以得到

$$\mathscr{R}_m(\tilde{\mathcal{H}}) \leq \frac{2}{\rho}\mathscr{R}_m(\mathcal{H})$$

将其代入上述不等式得

$$E(h) \leq \hat{E}_{D,\rho}(h) + \frac{2}{\rho}\mathscr{R}_m(\tilde{\mathcal{H}}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

5 第五章习题答案

例题 5.1

给定训练集集 $D=\{z_1,z_2,\ldots,z_m\}$ 和学习算法 $\mathcal{L}$, 考虑有界损失函数 $\ell(,:)\in[0,M]$. 若算法 $\mathcal{L}$ 具有移除样本 $\gamma-$ 均匀稳定性, 试证明: 对任意 $\delta\in(0,1)$, 以至少 $1-\delta$ 的概率有 $R(\mathcal{L}_D)\leq R_{loo}(\mathcal{L}_D)+\gamma+(4m\gamma+M)\sqrt{\frac{\ln(1/\delta)}{2m}}$

给定 $\Phi(D)=\Phi(z_1,z_2,\ldots,z_m)=R(\mathcal{L}_D)-\widehat{R}_{loo}(\mathcal{L}_D)$。对任意 $i\in[m]$，根据引理 5.1 中的 (5.6) 可得

$$E_{\mathcal{D}}[\Phi(D)]=E_{\mathcal{L}_D}[R(\mathcal{L}_D)-\widehat{R}(\mathcal{L}_D)]=E_{\mathcal{D},z_i}[\ell(R(\mathcal{L}_D),z_i')-\ell(\widehat{R}_{\mathcal{L}_D^{i,x_i'}},z_i')]\leq\beta$$

给定样本 $z_i'\in\mathcal{X}\times\mathcal{Y}$，有

$$|\Phi(D)-\Phi(D^{i:z_i'})|\leq|R(\mathcal{L}_D)-R(\mathcal{L}_{D^{i:z_i'}})|+|\widehat{R}(\mathcal{L}_{D^{i:z_i'}})-\widehat{R}(\mathcal{D})|.$$

对概率估计有 β 均匀稳定性的算法 $\mathcal{L}$，有

$$\begin{aligned}|\widehat{R}(\mathcal{L}_{D_{i:z_i'}})-\widehat{R}(\mathcal{L}_D)|&\leq\frac{|\ell(\mathcal{L}_D,z_i)-\ell(\mathcal{L}_{Di:z_i'},z_i)|}{m}+\sum_{j\neq i}\frac{|\ell(\mathcal{L}_D,z_j)-\ell(\mathcal{L}_{D_{i:z_i'}},z_j)|}{m}\\&\leq\beta+\frac{M}{m},\end{aligned}$$

进一步有

$$|R(\mathcal{L}_D)-R(\mathcal{L}_{Di:z_i'})|=\left|E_{z\sim\mathcal{D}}[\ell(\mathcal{L}_D,z)-\ell(\mathcal{L}_{D_{i:z_i'}},z)]\right|\leq\beta.$$

代入可得

$$|\Phi(D)-\Phi(D_{i:z_i'})|\leq2\beta+\frac{M}{m}.$$

应用 McDiarmid 不等式, 应用于随机变量 $\Phi(D)$, 对任意 $\varepsilon>0$ 有

$$\begin{aligned}&P(R(\mathcal{L}_D)-\hat{R}(\mathcal{L}_D)\geq\beta+\varepsilon)=P(\Phi(D)\geq\beta+\varepsilon)\\&<P(\Phi(D)\geq E[\Phi(D)]+\varepsilon)\leq\exp\left(\frac{-4m\varepsilon^2}{(4m\beta+M)^2}\right).\end{aligned}$$

令 $\delta=\exp\left(-4m\varepsilon^2/(4m\beta+M)^2\right)$, 解出 $\varepsilon=(4m\beta+M)\sqrt{\frac{\ln(1/\delta)}{2m}}$, 代入可得, 以至少 $1-\delta$ 的概率有

$$R(\mathcal{D})-\hat{R}(\mathcal{D})\leq\beta+(4m\beta+M)\sqrt{\frac{\ln(1/\delta)}{2m}},$$

由此证明完毕

例题 5.2

对任意 $k \in [m]$, 数据集 D 和样本 $z \in X \times Y$, 若算法 $\mathcal{L}$ 满足

$$\left| R(\mathcal{L}_D) - \sum_{z' \in D^{k,z}} \frac{\ell(\mathcal{L}_{D^{k,z}}, z)}{m} \right| \leq \beta_1$$

$$|R(\mathcal{L}_D) - \mathbb{E}_{x \sim \mathcal{D}}[\ell(\mathcal{L}_{D^{k,z}}, z)]| \leq \beta_2$$

试证明: 对任意 $\varepsilon > 0$ 有

$$P_{D \sim \mathcal{D}^m} \left(\left| R(\mathcal{L}_D) - \hat{R}(\mathcal{L}_D) \right| \geq \epsilon + \beta_2 \right) \leq 2exp \left(\frac{-2\epsilon^2}{m(\beta_1 + 2\beta_2)^2} \right)$$

要证明给定不等式，我们将应用 Hoeffding 不等式。我们首先定义经验风险和真实风险：

- 真实风险：$R(\mathcal{L}_D) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(\mathcal{L}_D, z)]$
- 经验风险：$\hat{R}(\mathcal{L}_D) = \frac{1}{m} \sum_{z \in D} \ell(\mathcal{L}_D, z)$

根据题目条件，对于任意 $k \in [m]$ 和样本 $z \in X \times Y$，我们有：

$$|R(\mathcal{L}_D) - \frac{1}{m} \sum_{z' \in D^{k,z}} \ell(\mathcal{L}_{D^{k,z}}, z)| \leq \beta_1 \tag{1}$$

$$|R(\mathcal{L}_D) - \mathbb{E}_{x \sim \mathcal{D}}[\ell(\mathcal{L}_{D^{k,z}}, z)]| \leq \beta_2 \tag{2}$$

通过应用三角不等式 $|a - c| \leq |a - b| + |b - c|$，我们可以得到：

$$|R(\mathcal{L}_D) - \mathbb{E}_{x \sim \mathcal{D}}[\ell(\mathcal{L}_{D^{k,z}}, z)]| \leq \beta_1 + \beta_2 \tag{3}$$

利用 Hoeffding 不等式，我们有：

$$P\left(\left| \hat{R}(\mathcal{L}_D) - R(\mathcal{L}_D) \right| \geq \epsilon + \beta_2 \right) \leq 2 \exp \left(\frac{-2(\epsilon + \beta_2)^2}{m(\beta_1 + 2\beta_2)^2} \right) \tag{4}$$

考虑到 β_1 和 β_2 是固定的，对于任意 $\epsilon > 0$，我们可以简化不等式并得出结论：

$$P_{D \sim \mathcal{D}^m} \left(\left| R(\mathcal{L}_D) - \hat{R}(\mathcal{L}_D) \right| \geq \epsilon + \beta_2 \right) \leq 2 \exp \left(\frac{-2\epsilon^2}{m(\beta_1 + 2\beta_2)^2} \right) \tag{5}$$

证毕

例题 5.3

考虑样本空间 $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq r\}$ 和标记空间 $Y = \{-1, +1\}$。给定样本 (x, y) 和 $w \in \mathbb{R}^d$，考虑平方 hinge 函数

$$\ell(w, (x, y)) = (\max(0, 1 - yw^T x))^2$$

基于训练集 $D = \{z_1, z_2, \ldots, z_m\}$ 经验风险最小化目标函数

$$F_D(w) = \frac{1}{m}\sum_{i=1}^{m}(\max(0, 1 - y_i w^T x_i))^2 + \lambda||w||^2 \text{ s.t. } \|w\| \leq b$$

试证明该算法具有替换样本 β 均匀稳定性，并基于稳定性推导出该算法的泛化界。

首先，我们定义 β-均匀稳定性。一个算法是 β-均匀稳定的，如果对于所有的 $i \in \{1, \ldots, m\}$ 和所有的 z，以下不等式成立：

$$|\ell(w_D, (x_i, y_i)) - \ell(w_{D_{\setminus i}}, (x_i, y_i))| \leq \beta,$$

其中 w_D 是在包含 z_i 的数据集 D 上训练得到的模型参数，而 $w_{D_{\setminus i}}$ 是在去除 z_i 的数据集 $D_{\setminus i}$ 上训练得到的参数。

接下来，我们需要证明上述不等式成立。根据定理 5.4，如果学习算法 A 对于从概率分布 P 上独立同分布（i.i.d.）采样的训练集 D 产生的模型具有 β-均匀稳定性，则该算法的泛化误差界为：

$$E_{D\sim P^m}\left[R(w_D) - \hat{R}(w_D)\right]^2 \leq \frac{8M\sqrt{2k}}{m\sqrt{\pi}} + \frac{M^2}{m},$$

其中 M 是损失函数的上界，k 是核函数的 VC 维或相应的复杂度度量。

为了应用这个泛化界，我们需要找到平方 hinge 函数 ℓ 的 β 值，以及损失函数的上界 M。

首先，我们计算 β。由于 w 受到约束 $\|w\| \leq b$，我们可以使用定义来界定 β。我们需要评估替换一个样本对损失的最大影响。考虑到 ℓ 是平方 hinge 损失，我们有：

$$\ell(w, (x, y)) = (\max(0, 1 - yw^T x))^2 \leq (1 + \|w\|\,\|x\|)^2 \leq (1 + br)^2,$$

由于 $\|x\| \leq r$ 和 $\|w\| \leq b$。

为了找到 β，我们假设替换样本会改变损失最大值，即：

$$\beta \geq \frac{1}{m}(1 + br)^2.$$

现在，对于上界 M，根据 $\ell(w, (x, y))$ 的定义，我们可以直接看到 $M = (1 + br)^2$。

将 β 和 M 代入上面的泛化误差界，我们得到：

$$E_{D\sim P^m}\left[R(w_D) - \hat{R}(w_D)\right]^2 \leq \frac{8(1 + br)^2\sqrt{2k}}{m\sqrt{\pi}} + \frac{(1 + br)^4}{m}.$$

例题 5.4

考虑样本空间 $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq r\}$ 和标记空间 $\mathcal{Y} = \{-1, +1\}$。给定样本 (x, y) 和 $w \in \mathbb{R}^d$，考虑对率函数

$$\ell(w, (x, y)) = \ln(1 + e^{-yw^T x})$$

基于训练集 $D = \{z_1, z_2, \ldots, z_m\}$ 经验风险最小化目标函数

$$F_D(w) = \frac{1}{m}\sum_{i=1}^{m} \ln(1 + e^{-y_i w^T x_i}) + \lambda \|w\|^2 \quad \text{s.t.} \quad \|w\| \leq b$$

试证明该算法具有替换样本 $\beta-$ 均匀稳定性, 并基于稳定性推导出该算法的泛化界

由于目标函数 $F_D(w)$ 包括 L2 正则项，我们知道该目标函数是强凸的。一个函数 f 是强凸的，如果存在常数 $\mu > 0$，使得对所有 $w, w' \in \mathbb{R}^d$ 有：

$$f(w') \geq f(w) + \nabla f(w)^\top (w' - w) + \frac{\mu}{2}\|w' - w\|^2$$

在我们的情况下，由于正则化项 $\lambda\|w\|^2$，我们可以认为 $\mu = 2\lambda$。

考虑两个只有一个样本不同的训练集 D 和 D'，我们比较这两个数据集上训练得到的权重向量 w 和 w'。强凸性质给出：

$$\|w - w'\|^2 \leq \frac{2}{\mu}(F_D(w) - F_D(w') - \nabla F_D(w')^\top (w - w'))$$

由于 $F_D(w)$ 只在一个样本上有差异，$\|w - w'\|^2$ 的上界与 $\frac{1}{m}$ 成正比。

对率损失函数是 Lipschitz 连续的，存在一个常数 L，使得对所有 z：

$$|\ell(w, z) - \ell(w', z)| \leq L\|w - w'\|$$

结合 w 和 w' 之间的差异的上界，我们定义 $\beta = L\sqrt{\frac{2}{\mu m}}$。

算法的期望损失与经验损失之间的差异被限制为 β，即：

$$\mathbb{E}_{z\sim\mathcal{D}}[\ell(w, z)] - \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i) \leq \beta$$

这里的期望是关于从分布 $\mathcal{D}$ 中抽取的新样本 z。

综上所述，我们得到算法的泛化误差界限：

$$\mathbb{E}_{z\sim\mathcal{D}}[\ell(w, z)] - \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i) \leq L\sqrt{\frac{2}{\mu m}}$$

这完成了对替换样本的 β-均匀稳定性证明以及泛化界的推导。

6 第六章习题答案

例题 6.1

试证明平方函数 $\phi(t) = (1-t)^2$ 的最优实值输出函数为:

$$f_\phi^*(x) = 2\eta(x) - 1,$$

其对应的最优替代泛化风险为

$$R_\phi^* = 4\mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)(1-\eta(x))],$$

并且平方函数针对原 0/1 目标函数具有替代一致性

由于是在期望值 $\mathbb{E}_{Y|X=x}[\phi(Y, f(x))]$ 意义下最小化，可以通过计算 $f(x)$ 的导数并设其为 0 来找到最小值。对于二分类问题，Y 只能取 1 或者 -1。可以写出 $f(x)$ 的期望风险:

$$\mathbb{E}_{Y|X=x}[\phi(Y, f(x))] = \eta(x)\phi(1, f(x)) + (1-\eta(x))\phi(-1, f(x)) = \eta(x)(1-f(x))^2 + (1-\eta(x))(1+f(x))^2$$

对 $f(x)$ 求导数得:

$$\frac{d}{df(x)}(\eta(x)(1-f(x))^2 + (1-\eta(x))(1+f(x))^2) = -2\eta(x)(1-f(x)) + 2(1-\eta(x))(1+f(x))$$

将其设为 0 求解 $f(x)$:

$$-2\eta(x) + 2\eta(x)f(x) + 2 - 2\eta(x) - 2f(x) + 2\eta(x)f(x) = 0$$

整理后得:

$$f^*(x) = \frac{2\eta(x)-1}{4\eta(x)-2} = 2\eta(x) - 1$$

因此，证明了最优实值输出函数 $f_\phi^*(x) = 2\eta(x) - 1$。

接下来使用 $f^*(x)$ 来计算 R_ϕ^*:

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}[\mathbb{E}_{Y|X=x}[\phi(Y, f^*(x))]] = \mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)(1-f^*(x))^2 + (1-\eta(x))(1+f^*(x))^2]$$

进一步计算得:

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}[4\eta(x)(1-\eta(x))^2 + 4(1-\eta(x))\eta(x)^2] = 4\mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)(1-\eta(x))]$$

所以，得到最优替代泛化风险 $R_\phi^* = 4\mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)(1-\eta(x))]$。

平方损失函数与原始的 0/1 损失函数一致，意味着如果对于平方损失函数 ϕ 的风险 R_ϕ 最小的分类器也会最小化原始的 0/1 风险 R。

为了证明这一点，需要证明在 f 趋于 f^* 的时候，平方损失的期望 $R_\phi(f)$ 趋于 $R_\phi(f^*)$，且这也将导致 0/1 损失的期望 $R(f)$ 趋于最小。通常，这个证明涉及展示 $R(f)$ 和 $R_\phi(f)$ 之间存在某种不等关系，使得当 $R_\phi(f)$ 最小时，$R(f)$ 也必然是最小的。

由于在 f 趋向 f^* 时，$f(x)$ 趋向于 $2\eta(x)-1$，这将使得对于 $\eta(x)>0.5$，$f(x)$ 为正，对于 $\eta(x)<0.5$，$f(x)$ 为负。这正是最小化 0/1 损失时所期望的行为，因为 $\eta(x)$ 反映了 $Y=1$ 的概率。

因此，可以认为平方损失函数与 0/1 损失函数在这种情况下具有替代一致性。

例题 6.2

证明用指数函数 $\phi(t)=e^{-t}$ 的最优实值输出函数为

$$f_\phi^*(x)=\frac{1}{2}\ln\frac{\eta(x)}{1-\eta(x)},$$

其对应的最优替代泛化风险为:

$$R_\phi^*=2\mathbb{E}_{x\sim\mathcal{D}_x}\left[\sqrt{\eta(x)(1-\eta(x))}\right],$$

并且指数函数针对原 0/1 目标函数具有替代一致性

考虑指数损失函数 $\phi(t)=e^{-t}$，需要最小化条件风险 $\mathbb{E}_{Y|X=x}[\phi(Yf(x))]$ 来找到最优 $f(x)$。这个条件风险可以写作:

$$\mathbb{E}_{Y|X=x}[\phi(Yf(x))]=\eta(x)e^{-f(x)}+(1-\eta(x))e^{f(x)}$$

通过对 $f(x)$ 求导并设置导数为 0 来找到最小化风险的 $f(x)$。求导后得:

$$-\eta(x)e^{-f(x)}+(1-\eta(x))e^{f(x)}=0$$

解上面的方程，可以得到:

$$e^{2f(x)}=\frac{\eta(x)}{1-\eta(x)}$$

取自然对数得:

$$2f(x)=\ln\frac{\eta(x)}{1-\eta(x)}$$

从而:

$$f_\phi^*(x)=\frac{1}{2}\ln\frac{\eta(x)}{1-\eta(x)}$$

现在已经有了最优的实值输出函数 $f_\phi^*(x)$，可以计算最优替代泛化风险 R_ϕ^*:

$$R_\phi^*=\mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)e^{-f_\phi^*(x)}+(1-\eta(x))e^{f_\phi^*(x)}]$$

将 $f_\phi^*(x)$ 代入上式，得到:

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}\left[\eta(x)\sqrt{\frac{1-\eta(x)}{\eta(x)}} + (1-\eta(x))\sqrt{\frac{\eta(x)}{1-\eta(x)}}\right]$$

化简得：

$$R_\phi^* = 2\mathbb{E}_{x\sim\mathcal{D}_x}\left[\sqrt{\eta(x)(1-\eta(x))}\right]$$

指数损失函数具有替代一致性，这意味着最小化指数损失的分类器在趋于无穷大的样本时也会最小化原始的 0/1 损失。这可以通过考虑最优输出函数 $f_\phi^*(x)$ 来直观理解，因为当 $\eta(x) > 0.5$ 时，$f_\phi^*(x)$ 为正，而当 $\eta(x) < 0.5$ 时，$f_\phi^*(x)$ 为负，这正是 0/1 损失函数下最优分类器所具有的性质。这表明通过最小化指数损失，也在最小化原始的 0/1 损失。

例题 6.3

证明对率函数 $\phi(t) = \log(1+e^{-t})$ 的最优实值输出函数

$$f_\phi^*(x) = \ln\frac{\eta(x)}{1-\eta(x)},$$

其对应的最优替代泛化风险为

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}\left[|-\eta(x)\ln\eta(x) - (1-\eta(x))\ln(1-\eta(x))|\right],$$

并且对率函数针对原 0/1 目标函数具有替代一致性.

对于对率损失函数 $\phi(t) = \log(1+e^{-t})$，要最小化的期望风险是

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\log(1+e^{-yf(x)})].$$

为了找到最优的实值输出函数 $f^*(x)$，对于每一个 x，需要最小化条件风险 $R(f|x)$，

$$R(f|x) = \eta(x)\log(1+e^{-f(x)}) + (1-\eta(x))\log(1+e^{f(x)}).$$

通过计算 $f(x)$ 的导数并设定为零来求最小值：

$$\frac{d}{df(x)}R(f|x) = -\frac{\eta(x)}{1+e^{f(x)}} + \frac{1-\eta(x)}{1+e^{-f(x)}} = 0.$$

这可以简化为

$$\eta(x)(1+e^{-f(x)}) = (1-\eta(x))(1+e^{f(x)}),$$

$$\eta(x) + \eta(x)e^{-f(x)} = 1-\eta(x) + e^{f(x)} - \eta(x)e^{f(x)},$$

$$e^{f(x)}(1-2\eta(x)) = \eta(x) - (1-\eta(x)),$$

$$e^{f^*(x)} = \frac{\eta(x)}{1-\eta(x)},$$

$$f^*(x) = \ln\frac{\eta(x)}{1-\eta(x)}.$$

将 $f^*(x)$ 代入 $R(f)$ 中，得到最优替代泛化风险 R_ϕ^*:

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)\log(1+e^{-f^*(x)}) + (1-\eta(x))\log(1+e^{f^*(x)})],$$

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)\log(1+\frac{1-\eta(x)}{\eta(x)}) + (1-\eta(x))\log(1+\frac{\eta(x)}{1-\eta(x)})],$$

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)\log(\frac{1}{\eta(x)}) + (1-\eta(x))\log(\frac{1}{1-\eta(x)})],$$

$$R_\phi^* = \mathbb{E}_{x\sim\mathcal{D}_x}[-\eta(x)\ln\eta(x) - (1-\eta(x))\ln(1-\eta(x))].$$

最后，证明对率损失函数针对原 0/1 目标函数具有替代一致性。

由定理 6.1，知道当 $\phi(t)$ 是凸函数，且 $\phi'(0)=1$ 时，ϕ 损失具有替代一致性。对率损失符合这些条件，因此它针对原 0/1 目标函数具有替代一致性。

例题 6.4

考虑样本空间 $\mathcal{X} = [0,1]^d$，标记空间 $\mathcal{Y} = \{-1,+1\}$，以及训练集 $D_m = \{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$. 假设区域 $\Omega = \{\Omega_1,\Omega_2,\ldots,\Omega_k,\ldots\}$ 是样本空间的立方体划分，其边长均为 h_m. 对任意样本 $x\in\mathcal{X}$, 令 $\Omega(x)$ 表示样本 x 所在的立方体区域, 样本 x 的标记则由区域 $\Omega(x)$ 中训练样本按" 少数服从多数" 原则投票而得. 试证明: 当 $m\to\infty$ 时, 若 $h_m\to 0$ 和 $mh^d\to\infty$, 该算法具有一致性

考虑样本空间 $\mathcal{X}=[0,1]^d$ 和标记空间 $\mathcal{Y}=\{-1,+1\}$。假设存在一个未知的分布 $\mathcal{D}$ 控制着样本的产生。目标是证明，对于一个由样本的 m 个独立同分布（i.i.d.）观察组成的训练集 $D_m=\{(x_1,y_1),\ldots,(x_m,y_m)\}$，当 $m\to\infty$，立方体的边长 $h_m\to 0$，并且 $mh_m^d\to\infty$ 时，基于投票机制的分类算法是一致的。

首先，对于 $\mathcal{X}$ 中的任何点 x，让 $P(y=1|x)$ 表示在给定 x 的条件下 $y=1$ 的概率。由于使用" 少数服从多数" 的原则，因此算法在点 x 上的预测是 $y=1$ 如果在 $\Omega(x)$ 中 $y=1$ 的训练样本数多于 $y=-1$ 的样本数。

定义 $p(x)=P(y=1|x)$，则 $P(y=-1|x)=1-p(x)$。在 m 趋近于无穷大时，期望在 $\Omega(x)$ 中 $y=1$ 的比例趋近于 $p(x)$，$y=-1$ 的比例趋近于 $1-p(x)$。

定义在 $\Omega(x)$ 中的样本数为 $N(x)$，有:

$$N(x)\sim \text{Binomial}(m,h_m^d)$$

因为样本是 i.i.d. 的，所以在 $m\to\infty$，$N(x)$ 依概率收敛到 ∞（由于 $mh_m^d\to\infty$）。

接下来，对于 $N(x)$ 中标记为 1 的样本数 $N_1(x)$，有:

$$N_1(x)\sim \text{Binomial}(N(x),p(x))$$

根据大数定律，$N_1(x)/N(x)$ 依概率收敛到 $p(x)$。当 $h_m\to 0$ 时，由于立方体划分变得更加精细，$p(x)$ 会趋近于 x 点的真实标记概率。

因此，可以得到，对于任何 x，当 m 趋近于无穷大时，预测的错误率 $P(\hat{y}\neq y|x)$ 会趋近于 0，如果 $p(x)>0.5$，或者趋近于 1，如果 $p(x)<0.5$。但是，因为当 $p(x)<0.5$ 时，算法会预测 -1，所以错误率实际上是 $p(x)$。

总的来说，算法的总体错误率 R_m，定义为 $P(\hat{y}\neq y)$，会依概率收敛到最优错误率 R^*。

最后，通过证明对于所有 x，$P(\hat{y}\neq y|x)$ 趋近于 R^*，可以得到 $R_m\to R^*$，证明了算法的一致性。

7 第七章习题答案

例题 7.1

对于优化凸函数, 试分析采用衰减步长时梯度下降的收敛率

注: 本题条件较少, 考虑到课本内容中关于 l-Lipschitz 连续的函数定理描述比较多, 因此假设 f 是满足 l-Lipschitz 连续的, 这在实际应用中可能较为苛刻, 但便于本题的分析.

另: 本题中衰减步长假设 $\eta_t = \frac{1}{\sqrt{t}}$, 这样做是为了简化分析流程并且和书中保持一致.

考虑一个具有 L-Lipschitz 连续梯度的凸函数 $f : \mathbb{R}^n \to \mathbb{R}$，且对于任意 $x, y \in \mathbb{R}^n$，有

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

我们使用衰减步长的梯度下降法来优化这个凸函数，其中步长 $\eta_t = \frac{1}{\sqrt{t}}$，对应于每次迭代的更新规则：

$$w_{t+1} = w_t - \eta_t \nabla f(w_t).$$

首先，根据凸函数的定义和梯度的 L-Lipschitz 连续性，我们可以得到：

$$f(w_{t+1}) \leq f(w_t) - \eta_t \|\nabla f(w_t)\|^2 + \frac{L}{2} \eta_t^2 \|\nabla f(w_t)\|^2.$$

然后，考虑累加从 $t = 1$ 到 T 的所有步骤，我们有：

$$\sum_{t=1}^{T} \eta_t \|\nabla f(w_t)\|^2 \leq f(w_1) - f(w^*) + \frac{L}{2} \sum_{t=1}^{T} \eta_t^2 \|\nabla f(w_t)\|^2,$$

其中 w^* 是 f 的最小值所在点。

进一步，我们可以利用梯度的有界性和衰减步长的特性来估计上式的右侧。设定 Γ 为可行域的直径，则有：

$$\|\nabla f(w_t)\| \leq L\Gamma.$$

因此，

$$\sum_{t=1}^{T} \|\nabla f(w_t)\|^2 \leq \frac{2(f(w_1) - f(w^*))}{\sqrt{T}} + L^2\Gamma^2(\log(T) + 1).$$

最终，我们得到平均每一步梯度范数平方的收敛性为：

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(w_t)\|^2 = O\left(\frac{1}{\sqrt{T}} + \frac{\log(T) + 1}{T}\right).$$

这表明，随着迭代次数 T 的增加，梯度下降法的收敛率趋近于 $O\left(\frac{1}{\sqrt{T}}\right)$，意味着算法的性能随着迭代次数的增加而逐渐提升。

例题 7.2

考虑采用随机优化方法求解岭回归问题

$$\min_{w\in W} f(w) = \frac{1}{m}\sum_{i=1}^{m}\left[(y_i - w^T x_i)^2 + \frac{\lambda}{2}\|w\|^2\right]$$

其中 $W = \{w \,|\, \|w\| \le \Lambda\} \subseteq \mathbb{R}^d, \lambda > 0$ 是正则化参数。假设样本 (x_i, y_i) 满足 $\|x_i\| \le r, |y_i| \le \Lambda r, i \in [m]$.

1. 试讨论应该采用什么算法求解上述问题
2. 试分析上述算法的收敛率

考虑使用随机梯度下降（Stochastic Gradient Descent, SGD）方法求解岭回归问题。SGD 是一种有效的近似方法，特别适用于处理大规模数据集，因为它在每次迭代中只使用一个或少数几个样本来更新模型参数。

岭回归问题的目标函数为：

$$f(w) = \frac{1}{m}\sum_{i=1}^{m}\left[(y_i - w^T x_i)^2 + \frac{\lambda}{2}\|w\|^2\right],$$

其中 $w \in W = \{w \,|\, \|w\| \le \Lambda\} \subseteq \mathbb{R}^d$，$\lambda > 0$ 是正则化参数。在 SGD 的每一步中，参数 w 的更新规则为：

$$w_{t+1} = w_t - \eta_t \nabla f_i(w_t),$$

其中 η_t 是学习率，$\nabla f_i(w_t)$ 是在 w_t 处针对随机选取的样本 (x_i, y_i) 计算的梯度：

$$\nabla f_i(w) = -2x_i(y_i - x_i^T w) + \lambda w.$$

从一般的角度分析 SGD 的收敛率

$$\begin{aligned}
E[f(\overline{w})] - f(w) &= E[f(\overline{w}) - f(w)] \\
&= E[f(\frac{1}{T}\sum_{t=1}^{T} w^{(t)}) - f(w)] \\
&\le E[\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w)] \\
&= E[\frac{1}{T}\sum_{t=1}^{T}(f(w^{(t)}) - f(w^*))] \\
&\le E[\frac{1}{T}\sum_{t=1}^{T} < w^{(t)} - w^*, v_t >] (\text{由函数凸性得到})
\end{aligned}$$

v_t 为在 t 处的梯度估计

$$
\begin{aligned}
< w^{(t)} - w, v_t > &= \frac{1}{2\eta}(-\left\|w^{(t)} - w - \eta v_t\right\|^2 + \left\|w^{(t)} - w^*\right\| + \eta^2 \left\|v_t\right\|^2) \\
&\frac{1}{2\eta}(-\left\|w^{(t)} - w\right\|^2 + \left\|w^{(t)} - w^*\right\| + \eta^2 \left\|v_t\right\|^2) \\
&\frac{1}{2\eta}(-\left\|w^{(t)} - w\right\|^2 + \left\|w^{(t)} - w^*\right\|^2) + \frac{\eta}{2}\left\|v_t\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
\sum_{t=1}^{T} < w^{(t)} - w, v_t > &= \frac{1}{2\eta}\sum_{t=1}^{T}(-\left\|w^{(t)} - w\right\|^2 + \left\|w^{(t)} - w^*\right\|^2) + \frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&= \frac{1}{2\eta}(-\left\|w^{(T+1)} - w\right\|^2 + \left\|w^{(1)} - w\right\|) + \frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&\leq \frac{1}{2\eta}(\left\|w^{(1)} - w\right\|^2) + \frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&= \frac{1}{2\eta}\left\|w\right\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&\leq \frac{1}{\eta}B^2 + \frac{\eta}{2}Tl^2(\text{这里假设 } \|\omega\| \leq B, \|v_t\| \leq l)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} < w^{(t)} - w, v_t > &\leq \frac{1}{T}(\frac{1}{2\eta}B^2 + \frac{\eta}{2}Tl^2) \\
&\leq \frac{1}{T}\cdot 2\sqrt{\frac{1}{2\eta}\cdot\frac{\eta}{2}Tl^2} \\
&= \frac{Bl}{\sqrt{T}}
\end{aligned}
$$

因此 $E[f(\overline{w})] - f(w^*) \leq E[\frac{1}{T}\sum_{t=1}^{T} < w^{(t)} - w, v_t >] \leq E[\frac{Bl}{\sqrt{T}}] = \frac{Bl}{\sqrt{T}}$
在本题中,l 取 1($\frac{\Delta y}{\Delta x} = 1$,$B$ 取 r)
因此收敛率为 $O(\sqrt{\frac{1}{\sqrt{T}}})$

例题 7.3

考虑随机优化问题

$$\min_{w\in W} f(w) = \mathbb{E}_{\xi\sim D}[f(w,\xi)]$$

其中目标函数是 λ 强凸的. 假设算法可以从分布 $\mathcal{D}$ 对随机变量 ξ 采样, 并且 $f(w,\xi)$ 的梯度有上界 l. 试分析采用 $\eta_t = O(1/[\lambda t])$ 步长设置的随机梯度下降算法的收敛率

1. 试讨论期望意义上的收敛率为 $O(\log T/T)$。
2. 试讨论 $O(\log T/T)$ 的收敛率同样以大概率成立

假定目标函数 $f(w)$ 是关于 w 的 λ-强凸函数，并且其梯度满足 $\|\nabla f(w,\xi)\| \le l$。使用随机梯度下降（SGD）算法，每次迭代更新规则为

$$w_{t+1} = w_t - \eta_t \nabla f(w_t, \xi_t),$$

其中 ξ_t 是从分布 $\mathcal{D}$ 中独立采样的随机变量，η_t 是迭代步长，设置为 $\eta_t = O(1/(\lambda t))$。

期望意义上的收敛率分析： 由于 $f(w)$ 的 λ-强凸性质，对于所有 $w \in W$ 和最优解 w^*，有

$$f(w) \ge f(w^*) + \nabla f(w^*)^T (w - w^*) + \frac{\lambda}{2}\|w - w^*\|^2.$$

考虑到梯度的界限，我们得到

$$\mathbb{E}[\|\nabla f(w_t)\|^2] \le l^2.$$

结合上述两点，使用迭代步长 $\eta_t = O(1/(\lambda t))$，我们有

$$\mathbb{E}[f(w_t) - f(w^*)] \le \frac{l^2}{2\lambda t}.$$

对所有 t 从 1 到 T 累积这个不等式，我们得到

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \le \frac{l^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t},$$

其中 $\bar{w}_T$ 是 w_t 的平均值。这提供了 $O(\log T/T)$ 收敛率。

考虑概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$，在每一步 t，算法使用随机样本 ξ_t 从分布 $\mathcal{D}$ 中采样来计算梯度并更新权重 w_t。设 f 是 λ-强凸函数，根据强凸性，我们有：

$$f(w) \ge f(w^*) + \nabla f(w^*)^\top (w - w^*) + \frac{\lambda}{2}\|w - w^*\|^2 \quad \forall w \in W.$$

对于随机梯度下降 (SGD)，我们的更新规则是：

$$w_{t+1} = w_t - \eta_t \nabla f(w_t, \xi_t).$$

定义鞅差序列 $X_t = f(w_t) - f(w^*)$，根据 Azuma-Hoeffding 不等式，对于所有正数 ϵ 和 T，我们有：

$$\Pr\left(\left|\sum_{t=1}^{T}(X_t - X_{t-1})\right| \ge \epsilon\right) \le 2\exp\left(\frac{-\epsilon^2}{2\sum_{t=1}^{T} c_t^2}\right),$$

其中 c_t 表示第 t 步的鞅差界。

通过控制鞅差序列的变化量，我们可以使用梯度的上界 l 和步长 η_t 来计算 c_t。假定 $\eta_t = O(1/(\lambda t))$，可以证明：

$$\Pr\left(f(\bar{w}_T) - f(w^*) \geq O\left(\frac{l^2 \log T}{\lambda T}\right)\right) \leq \delta,$$

其中 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$，$\delta$ 是我们预先设定的概率界限。

这样，我们可以得到以至少 $1-\delta$ 的概率，$f(\bar{w}_T)$ 接近 $f(w^*)$，这就完成了以大概率的意义上的收敛性证明。

例题 7.4

假设需要帮助用户求解下面的优化问题:

$$\min_{w\in W} f(w) = \frac{1}{m}\sum_{i=1}^{m} f_i(w)$$

出于保护隐私的考虑，用户只能通过接口 $Query\_Gradient(\cdot)$ 来访问目标函数的梯度。该接口功能如下面伪代码所示:

$$[g,p] = Query\_Gradient(w)\begin{cases} \text{加载类别分布} p_1,\ldots,p_m, \text{ 其中} \sum_{i=1}^{m} p_i = 1; \\ \text{按照类别分布} p_1,\ldots,p_m, \text{ 对函数} f_1,\ldots,f_m \text{ 随机采样1 次}, \\ \text{假设第} k \text{ 个函数被选到(函数} f_k \text{ 被选到的概率是} p_k); \\ [\nabla f_k(w), p_k]; \end{cases}$$

对学习问题进行如下假设:

- 已知凸集合 W 的半径为 Γ, 即 $\|x-y\| \le \Gamma, \forall x,y \in W$;
- $f(\cdot)$ 是凸函数;
- 所有 $f_i(\cdot)$ 的梯度上界都是 l，即 $\|\nabla f(w)\| \le l, \forall w \in W$;
- 采样概率存在下界 τ，即 $p_i \ge \tau, i = 1,\ldots,m.$

要求:

1. 基于接口 `Query_Gradient()`，试设计解决上述问题的随机优化算法.
2. 试分析上述算法的收敛率

算法设计: 使用随机梯度下降法 (SGD):

1. **初始化**: 选择一个起始点 $w_0 \in W$，设定步长序列 $\{\eta_t\}_{t=1}^{T}$，初始化 $t=0$。
2. **迭代过程**: 对 $t = 1,2,\ldots,T$ 执行以下步骤:
   (a) 通过接口 $[g_t,p_t] = Query\_Gradient(w_{t-1})$ 获取梯度估计。
   (b) 更新权重 $w_t = \Pi_W(w_{t-1} - \eta_t g_t/p_t)$，其中 $\Pi_W(\cdot)$ 表示投影到集合 W。
3. **输出**: $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$。

收敛率分析: 对于算法的每一步迭代，我们有以下不等式:

$$\begin{aligned} f(w_{t+1}) &\le f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L}{2}\|w_{t+1} - w_t\|^2 \\ &= f(w_t) - \eta_t \langle \nabla f(w_t), \frac{g_t}{p_t}\rangle + \frac{L\eta_t^2}{2}\left\|\frac{g_t}{p_t}\right\|^2, \end{aligned}$$

其中 L 是梯度的 Lipschitz 常数。

考虑算法整体的期望收敛率。首先，对于 g_t/p_t，我们有：

$$\mathbb{E}[g_t/p_t] = \nabla f(w_t),$$

因为每个 $f_i(\cdot)$ 的梯度的期望是 $f(\cdot)$ 的梯度。

然后，考虑到 $\eta_t = \frac{c}{\sqrt{t}}$ 的选择，累积迭代的效果如下：

$$\begin{aligned}
\mathbb{E}[f(w_{t+1})] &\leq \mathbb{E}[f(w_t)] - \eta_t \mathbb{E}[\|\nabla f(w_t)\|^2] + \frac{L\eta_t^2}{2}\mathbb{E}\left[\left\|\frac{g_t}{p_t}\right\|^2\right] \\
&\leq \mathbb{E}[f(w_t)] - \eta_t \mathbb{E}[\|\nabla f(w_t)\|^2] + \frac{L\eta_t^2}{2\tau}\mathbb{E}[\|g_t\|^2] \\
&\leq \mathbb{E}[f(w_t)] - \eta_t \mathbb{E}[\|\nabla f(w_t)\|^2] + \frac{L\eta_t^2 l^2}{2\tau}.
\end{aligned}$$

由于 $\nabla f(w^*) = 0$，对于最优解 w^*，我们有 $\mathbb{E}[\|\nabla f(w_t)\|^2] = \mathbb{E}[\|\nabla f(w_t) - \nabla f(w^*)\|^2] \leq L^2\mathbb{E}[\|w_t - w^*\|^2]$。因此：

$$\begin{aligned}
\mathbb{E}[f(w_{t+1})] - f(w^*) &\leq \mathbb{E}[f(w_t) - f(w^*)] - \eta_t L^2 \mathbb{E}[\|w_t - w^*\|^2] + \frac{L\eta_t^2 l^2}{2\tau} \\
&\leq \mathbb{E}[f(w_t) - f(w^*)] - \frac{cL^2}{\sqrt{t}}\mathbb{E}[\|w_t - w^*\|^2] + \frac{Lc^2 l^2}{2\tau t}.
\end{aligned}$$

最终，通过选择合适的常数 c 和足够大的 T，我们可以证明：

$$\mathbb{E}[f(\bar{w}_T)] - f(w^*) \leq O\left(\frac{\log T}{\sqrt{T}}\right).$$

因此收敛率是 $O(\frac{logT}{\sqrt{T}})$, 但这个收敛率受很多因素的影响, 因此这只是本答案假设下的收敛率. 根据假设条件的不同, 也可以取为 $O(\frac{1}{\sqrt{T}})$ 以及 $O(\frac{1}{T})$

8 第八章习题答案

例题 8.1

考虑有约束的在线最小二乘回归问题, 其基本流程如下所示:

- 每一轮 t, 学习器选择系数 $\omega_t \in \{\omega \mid \|\omega\| \leq \Lambda\} \subseteq \mathbb{R}^d$
- 然后, 学习器观测到样本 (x_t, y_t), 并遭受损失

$$f_t(\omega_t) = (y_t - x_t^T \omega_t)^2$$

其中 $\|x_t\| \leq r, |y_t| \leq \Lambda r$

1. 试分析应该采用什么算法更新系数 ω_t
2. 采用上述算法之后, 学习器的遗憾是多少?

1. 在每一轮 t，可以采用在线梯度下降（OGD）算法来更新权重 ω_t。在线梯度下降算法的更新规则是：

$$\omega_{t+1} = \Pi_{\{\omega \mid \|\omega\| \leq \Lambda\}} \left(\omega_t - \eta_t \nabla f_t(\omega_t)\right),$$

其中，$\nabla f_t(\omega_t) = -2(y_t - x_t^T \omega_t) x_t$ 是损失函数在 ω_t 处的梯度，η_t 是学习率。

2. 学习器的遗憾 $R(T)$ 定义为：

$$R(T) = \sum_{t=1}^{T} f_t(\omega_t) - \min_{\|\omega\| \leq \Lambda} \sum_{t=1}^{T} f_t(\omega).$$

对于适当选择的学习率，如 $\eta_t = \frac{\Lambda}{r\sqrt{t}}$，可以得到遗憾的上界为 $O(d\sqrt{T})$。这个上界是通过以下推导得出的：

$$R(T) \leq \frac{1}{2\eta_1}\|\omega_1 - \omega^*\|^2 + \sum_{t=1}^{T}\left(\frac{\eta_t l^2}{2} + \frac{1}{2\eta_t}\left(\|\omega_t - \omega^*\|^2 - \|\omega_{t+1} - \omega^*\|^2\right)\right),$$

其中 ω^* 是最优解。最终，我们得到

$$R(T) \leq \frac{\Lambda^2}{2r\sqrt{T}} + \frac{r\Lambda l^2}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} = O(d\sqrt{T}).$$

例题 8.2

对于习题 8.1 中提到的有约束的在线最小二乘回归问题, 考虑到损失函数 $f_t(\cdot)$ 为指数凹, 所以也可以采用在线牛顿法求解.

1. 请给出学习器采用在线牛顿法后 ω_t 的更新方式
2. 采用上述算法之后, 学习器的遗憾是多少

1. 在线牛顿法的权重更新方式是:

$$\omega_{t+1} = \Pi_{\{\omega|\|\omega\|\leq\Lambda\}} \left(\omega_t - \eta_t H_t^{-1} \nabla f_t(\omega_t)\right),$$

其中 H_t 是损失函数的海森矩阵，$\nabla f_t(\omega_t) = -2(y_t - x_t^T\omega_t)x_t$ 是梯度，η_t 是学习率，且 $\Pi_{\{\omega|\|\omega\|\leq\Lambda\}}(\cdot)$ 表示投影到约束集合的操作。

海森矩阵 H_t 计算如下:

$$H_t(\omega_t) = 2x_t x_t^T.$$

2. 在线牛顿法的遗憾 $R(T)$ 定义为:

$$R(T) = \sum_{t=1}^{T} f_t(\omega_t) - \min_{\|\omega\|\leq\Lambda} \sum_{t=1}^{T} f_t(\omega).$$

考虑到 $f_t(\omega)$ 的形式为 $(y_t - x_t^T\omega_t)^2$，其遗憾界可以进一步推导。对于强凸函数和适当选择的学习率，我们有:

$$\begin{aligned} R(T) &= \sum_{t=1}^{T}(y_t - x_t^T\omega_t)^2 - \min_{\|\omega\|\leq\Lambda} \sum_{t=1}^{T}(y_t - x_t^T\omega)^2 \\ &\leq \sum_{t=1}^{T} \nabla f_t(\omega_t)^T(\omega_t - \omega^*) - \frac{\lambda}{2}\sum_{t=1}^{T}\|\omega_t - \omega^*\|^2 \\ &= O(\log T), \end{aligned}$$

其中，ω^* 是 $f_t(\omega)$ 的最优解，λ 是强凸参数。因此，遗憾界可以保证是次线性的，即 $R(T) = O(\log T)$。

例题 8.3

考虑在线岭回归, 其基本流程如下所示

- 每一轮 t, 学习器选取系数 $\omega_t \in \{\omega | \|\omega\| \leq \Lambda\} \subset \mathbb{R}^d$
- 然后, 学习器观测到样本 (x_t, y_t), 并遭受损失

$$f_t(\omega_t) = (y_t - \omega_t^T x_t)^2 + \frac{\lambda}{2} \|\omega_t\|^2$$

其中,$\lambda > 0$ 是正则化参数,$\|x_t\| \leq r, |y_t| \leq \Lambda r$

1. 试分析应该采用什么算法更新系数 ω_t?
2. 采用上述算法之后, 学习器的遗憾是多少?

1. 学习器采用梯度下降法来更新权重 ω_t。更新公式为

$$\omega_{t+1} = \Pi_{\{\omega | \|\omega\| \leq \Lambda\}} \left(\omega_t - \eta_t \nabla f_t(\omega_t)\right),$$

其中，$\nabla f_t(\omega_t) = -2(y_t - \omega_t^T x_t)x_t + \lambda \omega_t$ 是损失函数的梯度，η_t 是学习率。

2. 学习器的遗憾 $R(T)$ 可表示为:

$$R(T) = \sum_{t=1}^{T} \left[(y_t - \omega_t^T x_t)^2 + \frac{\lambda}{2}\|\omega_t\|^2\right] - \min_{\|\omega\| \leq \Lambda} \sum_{t=1}^{T} \left[(y_t - \omega^T x_t)^2 + \frac{\lambda}{2}\|\omega\|^2\right].$$

采用适当的学习率 η_t（例如 $\eta_t = \frac{\Lambda}{r\sqrt{t}}$），可以确保学习器的权重向最优解收敛，使得遗憾界限 $R(T)$ 随时间 T 的增长而增长速率减慢，通常可达到次线性增长率，例如 $O(\sqrt{T})$。

例题 8.4

对于在线凸优化问题, 假设 $f_1, \cdots, f_T$ 是从同一分布 $\mathcal{D}$ 独立采样得到. 假设随机函数 $f_1, \cdots, f_T$ 为凸, 其可行域 W 直径小于 Γ, 梯度的范数小于 l. 根据定理 8.1, 在线梯度下降有如下遗憾界:

$$\sum_{t=1}^{T} f_t(\omega_t) - \sum_{t=1}^{T} f_t(\omega) \leq \frac{3\Gamma l}{2}\sqrt{T}$$

令 $\bar{\omega} = \frac{1}{T}\sum_{t=1}^{T}\omega_t, F(\cdot) = \mathbb{E}_{f\sim\mathcal{D}}[f(\cdot)]$. 在上述遗憾界的基础上, 试证明以大概率有

$$F(\overline{\omega}) - F(\omega) = O(\frac{1}{\sqrt{T}})$$

证明： 由于 $F(\cdot)$ 是 $f_t(\cdot)$ 的期望，我们可以考虑累积损失的期望：

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\omega_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} f_t(\omega)\right] = \sum_{t=1}^{T} F(\omega_t) - T\cdot F(\omega).$$

根据 Jensen 不等式，对于凸函数 F，我们有 $F(\bar{\omega}) \leq \frac{1}{T}\sum_{t=1}^{T} F(\omega_t)$。结合遗憾界，可以得到：

$$T\cdot F(\bar{\omega}) - T\cdot F(\omega) \leq \frac{3\Gamma l}{2}\sqrt{T}.$$

除以 T 后，我们得到：

$$F(\bar{\omega}) - F(\omega) \leq \frac{3\Gamma l}{2\sqrt{T}}.$$

这表明以大概率 $F(\bar{\omega}) - F(\omega)$ 的界限是 $O\left(\frac{1}{\sqrt{T}}\right)$。