5 泛化性

5.1 引言

机器学习模型在未见数据上的预测能力通常可称为泛化性, 泛化性的好坏可由泛化误差界表示. 泛化误差界与上一章提到的假设空间复杂度密切相关. 上一章中提到过, 假设空间按照其中概念的数量是否有限可以分为两大类: 有限假设空间和无限假设空间.

本章组织结构如下: 5.2节介绍有限假设空间情况下的泛化误差界, 对于无限假设空间, 5.3节分别介绍基于 VC 维和 Rademacher 复杂度的泛化误差界, 最后 5.4节介绍一些常用机器学习算法（以 SVM 与 Adaboost 为例）的泛化误差界.

5.2 有限假设空间

有限假设空间, 顾名思义就是概念数目有限的假设空间. 概念的数目 $|\mathcal{H}|$ 能很好地刻画假设空间的复杂度, 在之后的泛化界推导的过程中, 将多次用到这一参数. 根据目标概念 c 是否在假设空间中, 泛化界的推导将分为可分情况与不可分情况.

5.2.1 可分情况

可分情况意味着目标概念 c 属于假设空间 $\mathcal{H}$, 即 $c \in \mathcal{H}$. 首先很容易想到, 对于可分的有限假设空间, 任何在训练集 D 上犯错的概念都肯定不是要找的目标概念. 由此可以剔除这些假设, 最终留下和数据集 D 一致的假设, 目标概念一定会存在于这些一致的假设中. 如果 D 足够大, 则最终剩下的一致的假设就很少, 从而能够以较大的概率找到目标概念的近似. 然而由于实际应用中数据集 D 通常只包括有限数量的样本, 所以假设空间 $\mathcal{H}$ 中会剩下不止一个和 D 一致的" 等效" 假设, 这些假设无法通过数据集 D 再进行区分. 通常无法强求通过训练集 D 精确找到目标概念, 在 PAC 框架中, 只要训练集 D 的规模能使学习算法 $\mathfrak{L}$ 以 $1-\delta$ 的概率找到目标假设的 ϵ 近似即可.

定理 5.1. *设 $\mathcal{H}$ 为有限假设空间, D 为包含 m 个从 $\mathcal{D}$ 独立同分布采样的样本, $\forall c \in \mathcal{H}$, 学习算法 $\mathfrak{L}$ 基于训练集 D 返回与训练集一致的假设为 h, 如果*

$$m \geq \frac{1}{\epsilon}(ln|\mathcal{H}| + ln\frac{1}{\delta})$$

则 $E(h) \leq \epsilon$ 以至少 $1-\delta$ 的概率成立, 即 $P(E(h) > \epsilon) < \delta$.

证明. 学习算法 $\mathfrak{L}$ 按照前面所说的方法可以找到与训练集一致的假设 $h \in \mathcal{H}$, 该假设的性能取决于使用的训练集 D, 我们期望的是找到的假设能够以较大的概率与目标概念 ϵ 近似. 首先, 假设 h 的泛化误差大于 ϵ 且与训练集一致, 那这样的假设出现的概率可以表示为:

$$P(h \in \mathcal{H} : E(h) > \epsilon \wedge \hat{E}(h) = 0),$$

我们只需证明该概率的不大于 δ 即可. 通过计算可得

$$\begin{aligned} P(h \in \mathcal{H} : E(h) > \epsilon \wedge \hat{E}(h) = 0) &\leq \sum_{h \in \mathcal{H}} P(E(h) > \epsilon \wedge \hat{E}(h) = 0) \\ &< |\mathcal{H}| P(E(h) > \epsilon \wedge \hat{E}(h) = 0) \\ &= |\mathcal{H}| P((h(\boldsymbol{x}_1) = y_1) \wedge \cdots \wedge (h(\boldsymbol{x}_m) = y_m)) \\ &= |\mathcal{H}|(1 - P(h(\boldsymbol{x}) \neq y))^m \\ &< |\mathcal{H}|(1 - \epsilon)^m. \end{aligned}$$

因此只需要保证上式最右端不大于 δ. 令 $|\mathcal{H}|(1-\epsilon)^m \leq \delta$ 可得:

$$m \geq \frac{1}{\epsilon}(\ln |\mathcal{H}| + \ln \frac{1}{\delta}).$$

■

这一定理表明假设空间 $\mathcal{H}$ 是有限可分时, 样本复杂度是关于 $1/\delta$ 和 $1/\epsilon$ 的多项式, 所以有限可分的假设空间 $\mathcal{H}$ 是 PAC 可学习的. 通过上述不等式还可以发现, 随着训练集中样本数目的逐渐增加, 泛化误差的上届逐渐趋近于 0, 收敛速率是 $O(1/m)$.

5.2.2 不可分情况

在不可分情况中, 目标概念不在假设空间中, 假设空间中的每个假设都会或多或少的出现分类错误, 我们的目标不再是希望找到目标概念的 ϵ 近似, 而是希望找到假设空间中泛化误差最小假设的 ϵ 近似. 首先考虑对于假设 h, 其经验误差与泛化误差之间的关系.

引理 5.1. ***Hoeffding* 不等式**: 对于 $\epsilon > 0$, 令 D 是大小为 m 且从 $\mathcal{D}$ 中独立同分布采样得到的训练集, 则对于任意假设 $h \in \mathcal{H}$, 有如下的不等式成立:

$$P\big(\hat{E}(h) - E(h) \geq \epsilon\big) \leq \exp(-2m\epsilon^2)$$

$$P\big(E(h) - \hat{E}(h) \geq \epsilon\big) \leq \exp(-2m\epsilon^2)$$

由此推出:

$$P\big(|\hat{E}(h) - E(h)| \geq \epsilon\big) \leq 2\exp(-2m\epsilon^2).$$

令上式右侧 $2\exp(-2m\epsilon^2)$ 等于 δ 可以得到如下推论

推论 5.1. 对于 $\epsilon > 0$, 令 D 是大小为 m 且从 $\mathcal{D}$ 中独立同分布采样得到的训练集, 则对于任意假设 $h \in \mathcal{H}$, 下式以至少 $1-\delta$ 的概率成立:

$$|\hat{E}(h) - E(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

这一推论表明了经验误差与泛化误差之间的关系, 当样本数目 m 较大时, h 的经验误差是泛化误差的较好近似.

定理 5.2. *$\mathcal{H}$ 为有限假设空间, 令 D 是大小为 m 且从 $\mathcal{D}$ 中独立同分布采样得到的训练集, 对于 $\delta > 0$ 下式成立:*

$$P\Big(\big|\hat{E}(h) - E(h)\big| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2m}}\Big) \geq 1 - \delta.$$

证明. 将 $\mathcal{H}$ 中的有限假设记为 $h_1, h_2 \ldots, h_{|\mathcal{H}|}$, 运用上述推论可得

$$\begin{aligned} P\Big(\exists h \in \mathcal{H} | \hat{E}(h) - E(h)| > \epsilon\Big) &= P\Big((|\hat{E}(h_1) - E(h_1)| > \epsilon) \vee \cdots \vee (|\hat{E}(h_{\mathcal{H}}) - E(h_{\mathcal{H}})| > \epsilon)\Big) \\ &\leq \sum_{h \in \mathcal{H}} P\Big(|\hat{E}(h) - E(h)| > \epsilon\Big) \\ &\leq 2|\mathcal{H}| \exp(-2m\epsilon^2). \end{aligned}$$

令 $2|\mathcal{H}| \exp(-2m\epsilon^2) = \delta$ 可证. ■

由上述定理可得 $E(h) \leq \hat{E}(h) + \sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}$ 以至少 $1-\delta$ 的概率成立. 由于 $\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}} = O(1/\sqrt{m})$, 所以在有限不可分情况下, 泛化误差的收敛速率为 $O(1/\sqrt{m})$. 与可分情况相比, 这一收敛速度较慢, 原来大小为 100 的训练集能达到训练效果, 在不可分情况下则需要大小为 10000 的训练集.

5.3 无限假设空间

对于无限假设空间, 需要从 VC 维和 Rademacher 复杂度的角度来分析其无限泛化误差界.

5.3.1 有限 VC 维的假设空间

首先回顾一下有限 VC 维, VC 维与增长函数密切相关, 对于增长函数则由如下定理成立:

定理 5.3. *对于假设空间 $\mathcal{H}$, $h \in \mathcal{H}$, $m \in \mathbb{N}$ 且 $0 < \epsilon < 1$, 有*

$$P\Big(|E(h) - \hat{E}(h)| > \epsilon\Big) \leq 4\Pi_{\mathcal{H}}(2m) \exp(-\frac{m\epsilon^2}{8})$$

.

再考虑到对于 VC 为 d 的假设空间 $\mathcal{H}$ 和任意整数 $m \geq d$, 有

$$\Pi_{\mathcal{H}}(m) \leq \Big(\frac{e \cdot m}{d}\Big)^d,$$

综上可得

定理 5.4. *假设空间 $\mathcal{H}$ 的 VC 维为 d, 对于任意 $m > d$, $0 < \delta < 1$ 和任意 $h \in \mathcal{H}$, 有*

$$P\Big(|E(h) - \hat{E}(h)| \leq \sqrt{\frac{8d \ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\Big) \geq 1 - \delta.$$

证明. 根据有限 VC 维的性质, 有如下不等式

$$4\Pi_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8}) \leq 4\Big(\frac{2em}{d}\Big)^d \exp(-\frac{m\epsilon^2}{8}).$$

令不等式右侧等于 δ 可得

$$\epsilon = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}},$$

从而定理得证. ■

根据上述定理可知 $E(h) \leq \hat{E}(h) + O(\sqrt{\frac{\ln(m/d)}{m/d}})$ 以至少 $1-\delta$ 的概率成立, 泛化误差的收敛速率为 $O(\sqrt{\frac{\ln(m/d)}{m/d}})$. 对于有限 VC 维的假设空间, 泛化误差的收敛速率与 VC 维的大小有关, VC 维越大, 假设空间越复杂, 泛化误差的收敛速率也越慢. 其次, 有限 VC 维的不可分假设空间比有限不可分假设空间更难收敛, 这也是有限假设空间与无限假设空间的区别.

由上述定理可得有限 VC 维假设空间的可学习性:

定理 5.5. *任何 VC 维有限的假设空间 $\mathcal{H}$ 都是 (不可知)PAC 可学习的.*

证明. 假设 $\mathfrak{L}$ 为满足经验风险最小化原则的算法, 即 $\mathfrak{L}$ 输出的假设 h 满足: $\hat{E}(h) = \min_{h'\in\mathcal{H}} \hat{E}(h')$. 令 g 表示 $\mathcal{H}$ 中具有最小泛化误差的假设, 即

$$E(g) = \min_{h\in\mathcal{H}} E(h).$$

令

$$\delta' = \frac{\delta}{2},$$
$$\sqrt{\frac{(\ln 2/\delta')}{2m}} \leq \frac{\epsilon}{2},$$

根据

$$P(|\hat{E}(g) - E(g)| \geq \epsilon) \leq 2\exp(-2m\epsilon^2)$$

可得

$$\hat{E}(g) - \frac{\epsilon}{2} \leq E(g) \leq \hat{E}(g) + \frac{\epsilon}{2}$$

以至少 $1-\delta/2$ 的概率成立. 再令

$$\frac{\epsilon}{2} \geq \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta'}}{m}}$$

并结合上述定理可知

$$P\Big(|E(h) - \hat{E}(h)| \leq \frac{\epsilon}{2}\Big) \geq 1 - \frac{\delta}{2}.$$

从而可知

$$\begin{aligned} E(h) - E(g) &\leq \hat{E}(h) + \frac{\epsilon}{2} - \Big(\hat{E}(g) - \frac{\epsilon}{2}\Big) \\ &= \hat{E}(h) - \hat{E}(g) + \epsilon \\ &\leq \epsilon \end{aligned}$$

以至少 $1-\delta/2$ 的概率成立. 所以 VC 维有限的假设空间是 (不可知)PAC 可学习的, 其中 m 可以通过 $\frac{\epsilon}{2} \geq \sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta'}}{m}}$ 求得. ■

5.3.2 基于 Rademacher 复杂度的泛化误差界

首先回顾一下 Rademacher 复杂度的定义, 函数空间 $\mathcal{F}$ 关于训练集 Z 的经验 Rademacher 复杂度是:

$$\hat{R}_Z(\mathcal{F}) = \mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{z}_i)\Big],$$

函数空间 $\mathcal{F}$ 关于 $\mathcal{Z}$ 上分布 $\mathcal{D}$ 的 Rademacher 复杂度:

$$R_m(\mathcal{F}) = \mathbb{E}_{Z\subset\mathcal{Z}:|Z|=m}\Big[\hat{R}_Z(\mathcal{F})\Big].$$

基于 Rademacher 复杂度可以得到关于函数空间 $\mathcal{F}$ 的泛化误差界 [Mohri et al., 2012]

定理 5.6. *对实值函数空间 $\mathcal{F}:\mathcal{Z}\to[0,1]$, 根据分布 $\mathcal{D}$ 从 $\mathcal{Z}$ 独立同分布采样得到示例集 $Z=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m\}$,$\boldsymbol{z}_i\in\mathcal{Z}$, $0<\delta<1$, 对任意 $f\in\mathcal{F}$, 下列不等式以至少 $1-\delta$ 的概率成立.*

$$\mathbb{E}[f(\boldsymbol{z})] \leq \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i) + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

$$\mathbb{E}[f(\boldsymbol{z})] \leq \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i) + 2\hat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

证明. 令

$$\hat{E}_Z(f) = \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i)$$

$$\Phi(Z) = \sup_{f\in\mathcal{F}}\mathbb{E}[f] - \hat{E}_Z(f)$$

同时, 令 Z' 为只与 Z 有一个示例不同的训练集, 不妨设 $\boldsymbol{z}_m\in Z$ 和 $\boldsymbol{z}'_m\in Z'$ 为不同示例, 可得

$$\begin{aligned}\Phi(Z')-\Phi(Z) &= \Big(\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\hat{E}_{Z'}(f)\Big)-\Big(\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\hat{E}_Z(f)\Big)\\ &\leq \sup_{f\in\mathcal{F}}\hat{E}_Z(f)-\hat{E}_{Z'}(f)\\ &= \sup_{f\in\mathcal{F}}\frac{f(\boldsymbol{z}_m)-f(\boldsymbol{z}'_m)}{m}\\ &\leq \frac{1}{m}.\end{aligned}$$

同理可得

$$\Phi(Z)-\Phi(Z') \leq \frac{1}{m},$$

$$|\Phi(Z)-\Phi(Z')| \leq \frac{1}{m}.$$

根据 McDiarmid 不等式可知, 对于任意 $\delta \in (0,1)$,

$$\Phi(Z) \leq \mathbb{E}_Z[\Phi(Z)] + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

以至少 $1-\delta$ 的概率成立. 下面来估计 $\mathbb{E}_Z[\Phi(Z)]$ 的上界

$$\begin{aligned}
\mathbb{E}_Z[\Phi(Z)] &= \mathbb{E}_Z\Big[\sup_{f\in\mathcal{F}} \mathbb{E}[f] - \hat{E}_Z(f)\Big] \\
&= \mathbb{E}_Z\Big[\sup_{f\in\mathcal{F}} \mathbb{E}_{Z'}[\hat{E}_{Z'}[f] - \hat{E}_Z(f)]\Big] \\
&\leq \mathbb{E}_{Z,Z'}\Big[\sup_{f\in\mathcal{F}} \hat{E}_{Z'}[f] - \hat{E}_Z(f)\Big] \\
&= \mathbb{E}_{Z,Z'}\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}(f(\boldsymbol{z}_i') - f(\boldsymbol{z}_i))\Big] \\
&= \mathbb{E}_{\boldsymbol{\sigma},Z,Z'}\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(\boldsymbol{z}_i') - f(\boldsymbol{z}_i))\Big] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},Z'}\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{z}_i')\Big] + \mathbb{E}_{\boldsymbol{\sigma},Z}\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}-\sigma_i f(\boldsymbol{z}_i)\Big] \\
&= 2\mathbb{E}_{\boldsymbol{\sigma},Z}\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{z}_i)\Big] \\
&= 2R_m(\mathcal{F}).
\end{aligned}$$

至此定理中第一个不等式得证. 令 Z' 为只与 Z 有一个示例不同的训练集, 不妨设 $\boldsymbol{z}_m \in Z$ 和 $\boldsymbol{z}_m' \in Z'$ 为不同示例, 由经验 Rademacher 复杂度定义可知

$$|\hat{R}_Z(f) - \hat{R}_{Z'}(f)| \leq \frac{1}{m}.$$

再由 McDiarmid 不等式可知

$$R_m(\mathcal{F}) \leq \hat{R}_Z(f) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

以 $1-\delta/2$ 的概率成立, 并且

$$\Phi(Z) \leq \mathbb{E}_Z[\Phi(Z)] + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

以至少 $1-\delta/2$ 的概率成立. 于是有

$$\Phi(Z) \leq 2\hat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

以至少 $1-\delta$ 的概率成立, 至此定理中第二个不等式得证. ■

上述定理的适用范围是实值函数空间 $\mathcal{F}: \mathcal{Z} \to [0,1]$, 一般用在回归问题中. 对于更加具体的分类问题, 有如下定理成立.

定理 5.7. *对于假设空间 $\mathcal{H}: \mathcal{X} \to \{-1,+1\}$, 根据分布 $\mathcal{D}$ 从 $\mathcal{X}$ 中独立同分布采样的得到示例集 $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$, $\boldsymbol{x}_i \in \mathcal{X}$, $0 < \delta < 1$, 对任意 $h \in \mathcal{H}$ 下列不等式以至少 $1-\delta$ 的概率成立.*

$$E(h) \leq \hat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

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

证明. 对于二分类问题的假设空间 $\mathcal{H}$, 令 $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$, 则 $\mathcal{H}$ 中的假设 h 可以变形为

$$f_h(\boldsymbol{z}) = f_h(\boldsymbol{x}, y) = \mathbb{I}(h(\boldsymbol{x}) \neq y).$$

于是值域为 $\{-1, +1\}$ 的假设空间 $\mathcal{H}$ 转化为值域为 $[0, 1]$ 的函数空间 $\mathcal{F}_{\mathcal{H}} = \{f_h : h \in \mathcal{H}\}$. 根据经验 Rademacher 复杂度的定义可知:

$$\begin{aligned}
\hat{R}_Z(\mathcal{F}_{\mathcal{H}}) &= \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f_h \in \mathcal{F}_{\mathcal{H}}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f_h(\boldsymbol{x}_i, y_i)\Big] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \mathbb{I}(h(\boldsymbol{x}_i) \neq y_i)\Big] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \frac{1 - y_i h(\boldsymbol{x}_i)}{2}\Big] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\frac{1}{m}\sum_{i=1}^{m} \sigma_i + \sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} (-y_i \sigma_i h(\boldsymbol{x}_i))\Big] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} (-y_i \sigma_i h(\boldsymbol{x}_i))\Big] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} (\sigma_i h(\boldsymbol{x}_i))\Big] \\
&= \frac{1}{2}\hat{R}_D(\mathcal{H}).
\end{aligned}$$

同时取期望可得

$$R_Z(\mathcal{F}_{\mathcal{H}}) = \frac{1}{2} R_D(\mathcal{H}).$$

结合上述定理, 定理得证. ■

5.3.3 基于 VC 维的泛化误差界与基于 Rademacher 复杂度的泛化误差界之间的关系

从前面的论述过程可知基于 VC 维的泛化误差界是分布无关且数据独立的, 仅与 VC 维和训练集的大小有关, 而基于 Rademacher 复杂度的泛化误差界与分布 $\mathcal{D}$ 有关 ($R_m(\mathcal{H})$ 项) 或与数据 D 有关 ($\hat{R}_D(\mathcal{H})$ 项). 换而言之, 基于 Rademacher 复杂度的泛化误差界依赖于具体学习问题上的数据及其分布, 是为具体的学习问题量身定制, 因此基于 Rademacher 复杂度的泛化误差界通常比基于 VC 维的泛化误差界更紧一些. 实际上通过 Rademacher 复杂度和增长函数的关系也可推导出基于 VC 维的泛化误差. 为此需要引入一个定理.

定理 5.8. ***Massart* 定理**: 令 $A \subset \mathbb{R}^m$ 为一个有限集合且 $r = \max_{\boldsymbol{x} \in A} \|\boldsymbol{x}\|_2$, 则有如下不等式成立.

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{\boldsymbol{x} \in A}\sum_{i=1}^{m} \sigma_i x_i\right] \leq \frac{r\sqrt{2\ln|A|}}{m}$$

其中 σ_i 为独立地且符合 $\{-1, +1\}$ 上均匀分布的随机变量, x_i 为向量 $\boldsymbol{x}$ 的分量.

证明. 对任意 $t>0$ 使用 Jensen 不等式可得如下不等式:

$$\begin{aligned}\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\le \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp(t\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i)\right]\\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\exp(t\sum_{i=1}^{m}\sigma_i x_i)\right]\\ &\le \sum_{\boldsymbol{x}\in A}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp(t\sum_{i=1}^{m}\sigma_i x_i)\right].\end{aligned}$$

接着使用 $\boldsymbol{\sigma}$ 的独立性及 Hoeffding 不等式可得

$$\begin{aligned}\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\le \sum_{\boldsymbol{x}\in A}\prod_{i=1}^{m}\mathbb{E}_{\sigma_i}\left[\exp(t\sigma_i x_i)\right]\\ &\le \sum_{\boldsymbol{x}\in A}\prod_{i=1}^{m}\exp\left[\frac{t^2(2x_i)^2}{8}\right]\\ &= \sum_{\boldsymbol{x}\in A}\exp\left[\frac{t^2}{2}\sum_{i=1}^{m}x_i^2\right]\\ &\le \sum_{\boldsymbol{x}\in A}\exp\left[\frac{t^2r^2}{2}\right]\\ &\le |A|e^{\frac{t^2r^2}{2}}.\end{aligned}$$

对不等式两边取对数可得

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \le \frac{\ln|A|}{t}+\frac{tr^2}{2}.$$

当 $t=\frac{\sqrt{2\ln|A|}}{r}$ 时, 右侧取最小值, 即

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \le r\sqrt{2\ln|A|}.$$

两边除以 m 定理得证. ∎

根据上述不等式可得如下推论:

推论 5.2. *假设空间 $\mathcal{H}$ 的 Rademacher 复杂度 $R_m(\mathcal{H})$ 与增长函数 $\Pi_{\mathcal{H}}(m)$ 满足*

$$R_m(\mathcal{H}) \le \sqrt{\frac{2\ln\Pi_{\mathcal{H}}(m)}{m}}.$$

证明. 对于给定数据集 $D=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\}$, $\mathcal{H}_{|D}$ 为假设空间在数据集 D 上的限制, $\mathcal{H}_{|D}$ 中的元素为 $(h(\boldsymbol{x}_1),\dots,h(\boldsymbol{x}_m))$ 形式的向量. 由于 $h\in\mathcal{H}$ 取值为 $\{-1,+1\}$, 则每个向量的模长为 $\sqrt{m}$, 运用 Massart 定理可得

$$\begin{aligned}R_m(\mathcal{H}) &= \mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{u\in\mathcal{H}_{|D}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i u_i\right]\right]\\ &\le \mathbb{E}_D\left[\frac{\sqrt{m}\sqrt{2\ln|\mathcal{H}_{|D}|}}{m}\right].\end{aligned}$$

又因为 $|\mathcal{H}_{|D}| \leq \Pi_{\mathcal{H}}(m)$, 有

$$R_m(\mathcal{H}) \leq \mathbb{E}_D \left[\frac{\sqrt{m}\sqrt{2\ln \Pi_{\mathcal{H}}(m)}}{m} \right] = \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}},$$

从而定理得证. ■

由上述定理, 再结合

$$E(h) \leq \hat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{e \cdot m}{d}\right)^d,$$

可得

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

以 $1-\delta$ 的概率成立, 并且这一泛化误差界比之前推导得到的更紧一些.

5.4 实例

本节将介绍一些常用的机器学习算法的泛化误差界, 具体包括 SVM, AdaBoost.

5.4.1 SVM 的泛化误差界

本节将介绍 SVM 的泛化误差界. SVM 通常会将原始特征空间映射到高维甚至无穷维的空间中, 这意味着 4.4.1 节得到 SVM 的 VC 维趋近于无穷. 显然, 这样的泛化误差界没有理论价值. 为了解决这一问题, 本节中将引入间隔理论, 使得 SVM 的泛化误差界不依赖于其特征空间的维数. 这样的泛化误差界的适用性更加广泛, 也更具理论价值.4.4.1节中提及过, SVM 的本质是线性分类器. 接下来, 我们先回顾一下线性分类器.

考虑输入空间 $\mathcal{X}$ 为 $\mathbb{R}^N$ 的子集, 其中 $N \geq 1$, 输出空间或者标记空间为 $\mathcal{Y} = \{-1, +1\}$, $c : \mathcal{X} \to \mathcal{Y}$ 为目标概念. 给定一个假设空间 $\mathcal{H}$, 即一族 $\mathcal{X}$ 映射到 $\mathcal{Y}$ 的函数, 和大小为 m 从分布 $\mathcal{D}$ 中独立同分布采样得到的数据集 $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, 其中 $y_i = c(\boldsymbol{x}_i)$, 学习器根据数据集从假设空间 $\mathcal{H}$ 中选取一个假设即二类分类器 h, 使得该分类器泛化误差较小, 其中泛化误差为

$$E(h) = P_{\boldsymbol{x}\sim\mathcal{D}}[h(\boldsymbol{x}) \neq c(\boldsymbol{x})].$$

可以选取不同的假设空间来完成这一分类任务, 根据之前关于假设空间复杂性的讨论可以知道, 其他条件不变的情况下, 较小的假设空间复杂度能为学习结果提供较高的保障. 常见的具有较小复杂度的假设空间是线性分类器, 或称线性超平面, 即

$$\mathcal{H} = \{\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b) : \boldsymbol{w} \in \mathbb{R}^N, b \in \mathbb{R}\}.$$

典型的线性超平面是将 $\boldsymbol{w}, b$ 放缩后, 满足 $\min_{(\boldsymbol{x},y)\in D} |\boldsymbol{w}\cdot\boldsymbol{x}+b| = 1$ 的 $(\boldsymbol{w}, b)$ 所构成的超平面. 一个形如 $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{w}\cdot\boldsymbol{x}+b)$ 的假设会将所有落在 $\boldsymbol{w}\cdot\boldsymbol{x}+b=0$ 一侧的点标记为正类, 另外一侧的点标记为负类. 该问题关注的是如何找到较好的线性超平面 $\boldsymbol{w}\cdot\boldsymbol{x}+b=0$, 即具有最大间隔的线性超平面:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$s.t.: y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b) \geq 1 \quad \forall i \in [1,m]$$

对于不可分情况则没有可分情况来得那么简单, 需要引入松弛变量即软间隔, 具体的分类问题也可以写成优化的形式:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i^p$$

$$s.t.: y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b) \geq 1-\xi_i \wedge \xi_i \geq 0, \forall i \in [1,m]$$

为了推导 SVM 的泛化误差界, 需要引入间隔理论首先了解一下关于间隔的定义:

定义 5.1. *一个标记为 y 的点 $\boldsymbol{x}$ 关于线性分类器 $h: \boldsymbol{x} \mapsto \boldsymbol{w}\cdot\boldsymbol{x}+b$ 的几何间隔 $\rho(\boldsymbol{x})$ 为该点到超平面 $\boldsymbol{w}\cdot\boldsymbol{x}+b=0$ 的距离*

$$\rho(\boldsymbol{x}) = \frac{y(\boldsymbol{w}\cdot\boldsymbol{x}+b)}{\|\boldsymbol{w}\|}$$

一个线性分类器关于数据集 $D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$ 的间隔是关于所有样本点的最小间隔

$$\rho = \min_{1\leq i\leq m} \frac{y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)}{\|\boldsymbol{w}\|}$$

之前提过 $\mathbb{R}^N$ 中超平面的 VC 维为 $N+1$, 因此由关于 VC 维的泛化误差界可得到如下结果: 对于任意 $\delta > 0$ 以至少 $1-\delta$ 的概率下式成立

$$E(h) \leq \hat{E}(h) + \sqrt{\frac{2(N+1)\ln\frac{em}{N+1}}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \quad \forall h \in \mathcal{H}$$

当特征空间 (输入空间) 的维数 N 相对于样本点很大时, 上式没有给出什么有价值的信息. 下面一个定理将不再依赖特征空间的维数 N, 而是依赖间隔以及包含所有样本点的球半径 r.

定理 5.9. *令 $D \subset \{\boldsymbol{x}: \|\boldsymbol{x}\| \leq r\}$, 典型超平面族 $\{\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{w}\cdot\boldsymbol{x}) : \min_{\boldsymbol{x}\in D}|\boldsymbol{w}\cdot\boldsymbol{x}| = 1 \wedge \|\boldsymbol{w}\| \leq \Lambda\}$ 的 VC 维 d 满足如下不等式*

$$d \leq r^2\Lambda^2.$$

证明. 令 $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$ 为能被打散的集合, 则对于任意 $y=(y_1,\ldots,y_m)\in\{-1,+1\}^d$ 存在 $\boldsymbol{w}$ 使得

$$\forall i \in [1,d], 1 \leq y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i).$$

对这些不等式求和可得

$$d \leq \boldsymbol{w}\cdot\sum_{i=1}^{d} y_i\boldsymbol{x}_i \leq \|\boldsymbol{w}\|\|\sum_{i=1}^{d} y_i\boldsymbol{x}_i\| \leq \Lambda\|\sum_{i=1}^{d} y_i\boldsymbol{x}_i\|.$$

由于上述不等式对任意 $y \in \{-1,+1\}^d$ 都成立, 对不等式两边按 $y_1,\ldots,y_m$ 服从 $\{-1,+1\}$ 独立且均匀的分布取期望, 由独立性假设可知 $\mathbb{E}[y_iy_j] = \mathbb{E}[y_i]\mathbb{E}[y_j]$, $i \neq j$. 再由均匀假设, 又可知 $\mathbb{E}[y_iy_j] = 0$, $i \neq j$, $\mathbb{E}[y_iy_j] = 1$, $i = j$,

$$\begin{aligned} d &\leq \Lambda \mathbb{E}_y[\|\sum_{i=1}^{d} y_i\boldsymbol{x}_i\|] \\ &\leq \Lambda \left[\mathbb{E}_y[\|\sum_{i=1}^{d} y_i\boldsymbol{x}_i\|^2]\right]^{1/2} \\ &= \Lambda \left[\sum_{i,j=1}^{d} \mathbb{E}_y[y_iy_j](\boldsymbol{x}_i\boldsymbol{x}_j)\right]^{1/2} \\ &= \Lambda[\sum_{i=1}^{d}(\boldsymbol{x}_i\boldsymbol{x}_i)]^{1/2} \\ &\leq \Lambda[dr^2]^{1/2} \\ &= \Lambda r\sqrt{d}, \end{aligned}$$

所以有 $\sqrt{d} \leq \Lambda r$, 定理得证. ■

对于典型超平面而言间隔 $\rho = \frac{1}{\|\boldsymbol{w}\|}$, 所以可以固定最大间隔 ρ, 令上述定理中 $\Lambda = 1/\rho$, 则典型超平面的 VC 维上界可以重新写为 r^2/ρ^2.

定理 5.10. 令 $D \subset \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq r\}$ 是大小为 m 的数据集, 则典型超平面族 $\mathcal{H} = \{\boldsymbol{x} \mapsto \boldsymbol{w}\cdot\boldsymbol{x} : \|\boldsymbol{w}\| \leq \Lambda\}$ 的经验 *Rademacher* 复杂度满足下列不等式

$$\hat{R}_D(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{m}}.$$

证明. 该定理的证明使用了和上一定理证明中相似的方法.

$$\begin{aligned}
\hat{R}_D(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i=1}^{m}\sigma_i \boldsymbol{w}\cdot\boldsymbol{x}_i\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\boldsymbol{w}\cdot\sum_{i=1}^{m}\sigma_i\boldsymbol{x}_i\right] \\
&\leq \frac{\Lambda}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|\sum_{i=1}^{m}\sigma_i\boldsymbol{x}_i\|\right] \\
&\leq \frac{\Lambda}{m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\|\sum_{i=1}^{m}\sigma_i\boldsymbol{x}_i\|^2\right]\right]^{1/2} \\
&= \frac{\Lambda}{m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{m}\sigma_i\sigma_j(\boldsymbol{x}_i\boldsymbol{x}_j)\right]\right]^{1/2} \\
&\leq \frac{\Lambda}{m}\left[\sum_{i=1}^{m}\|\boldsymbol{x}_i\|^2\right]^{1/2} \\
&\leq \sqrt{\frac{r^2\Lambda^2}{m}},
\end{aligned}$$

定理得证. ■

为了推导出主要的基于间隔的泛化误差界, 还需要介绍一种间隔损失函数, 因为上述定理中并没有要求数据集是可分的, 此时 $\rho > 0$ 应被看做是想要获得的间隔大小, 而不非实际得到的间隔.

定义 5.2. *对于任意的 $\rho > 0$, ρ-间隔损失是一个定义在 $y, y' \in \mathbb{R}$ 上的函数 $L_\rho : \mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$, $L_\rho(y, y') = \Phi_\rho(yy')$, 其中*

$$\Phi_\rho(x) = \begin{cases} 0 & \rho \leq x \\ 1 - x/\rho & 0 \leq x \leq \rho \\ 1 & x \leq 0 \end{cases}$$

定义 5.3. *给定数据集 $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ 与假设 h, 经验间隔损失定义为*

$$\hat{E}_\rho(h) = \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(y_i h(\boldsymbol{x}_i)).$$

考虑到 $\Phi_\rho(y_i h(\boldsymbol{x}_i)) \leq \mathbb{I}_{y_i h(\boldsymbol{x}_i)\leq\rho}$, 则经验间隔损失可由下列不等式控制住

$$\hat{E}_\rho(h) \leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}_{y_i h(\boldsymbol{x}_i)\leq\rho}.$$

在下面的结果中, 经验间隔损失将被它的上界替代. 根据经验间隔损失的定义可以知道 Φ_ρ 最多是 $1/\rho$-Lipschitz 的, 因为它的导数最大为 $1/\rho$. 接下来的引理说明一个 Lipschitz 函数和假设空间 $\mathcal{H}$ 复合后的经验 Rademacher 复杂度可以被假设空间的经验 Rademacher 复杂度控制住.

引理 5.2. *令 $\Phi:\mathbb{R}\to\mathbb{R}$ 是一个 l-Lipschitz 函数, 则对于任意的值域为实值的假设空间 $\mathcal{H}$ 有如下不等式成立:*

$$\hat{R}_D(\Phi\circ\mathcal{H})\le l\hat{R}_D(\mathcal{H}).$$

下面的定理给出了一般的基于间隔的二分类问题的泛化误差界.

定理 5.11. *令 $\mathcal{H}$ 为一族实值函数, 给定 $\rho>0$, 则对于任意 $\delta>0$ 和 $h\in\mathcal{H}$ 下式至少以 $1-\delta$ 的概率成立.*

$$E(h)\le\hat{E}_\rho(h)+\frac{2}{\rho}R_m(\mathcal{H})+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}},$$

$$E(h)\le\hat{E}_\rho(h)+\frac{2}{\rho}\hat{R}_D(\mathcal{H})+3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$

证明. 令 $\widetilde{\mathcal{H}}=\{\boldsymbol{z}=(\boldsymbol{x},y)\mapsto yh(\boldsymbol{x}):h\in\mathcal{H}\}$, 考虑值域为 $[0,1]$ 的函数族 $\mathcal{F}=\{\Phi_\rho\circ f:f\in\widetilde{\mathcal{H}}\}$, 根据前一定理可知对于所有 $g\in\mathcal{F}$, 下式至少以 $1-\delta$ 的概率成立

$$\mathbb{E}[g(\boldsymbol{z})]\le\frac{1}{m}\sum_{i=1}^m g(\boldsymbol{z}_i)+2R_m(\mathcal{F})+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

因此对任意的 $h\in\mathcal{H}$ 有

$$\mathbb{E}[\Phi_\rho(yh(\boldsymbol{x}))]\le\hat{E}_\rho(h)+2R_m(\Phi_\rho\circ\widetilde{\mathcal{H}})+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

由于 $\mathbb{I}_{u\le0}\le\Phi_\rho(u)$, $\forall u\in\mathbb{R}$ 有 $E(h)=\mathbb{E}[\mathbb{I}_{yh(\boldsymbol{x})\le0}]\le\mathbb{E}[\Phi_\rho(yh(\boldsymbol{x}))]$, 代入上式可得

$$E(h)\le\hat{E}_\rho(h)+2R_m(\Phi_\rho\circ\widetilde{\mathcal{H}})+\sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

由于 Φ_ρ 为 $1/\rho$-Lipschitz, 根据前面的引理可知

$$R_m(\Phi_\rho\circ\widetilde{\mathcal{H}})\le\frac{1}{\rho}R_m(\widetilde{\mathcal{H}}).$$

另外 $R_m(\widetilde{\mathcal{H}})$ 可以重写为

$$R_m(\widetilde{\mathcal{H}})=\frac{1}{m}\mathbb{E}_{D,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m\sigma_i y_i h(\boldsymbol{x}_i)\right]=\frac{1}{m}\mathbb{E}_{D,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m\sigma_i h(\boldsymbol{x}_i)\right]=R_m(\mathcal{H}),$$

整理可得定理中的第一个不等式. 第二不等式可以通过

$$\mathbb{E}[g(\boldsymbol{z})]\le\frac{1}{m}\sum_{i=1}^m g(\boldsymbol{z}_i)+2\hat{R}_D(\mathcal{F})+3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}$$

以相似的方法得到. ∎

在上述定理中, ρ 是固定的, 下面的定理则对于任意 $\rho\in(0,1)$ 均成立.

定理 5.12. *令 $\mathcal{H}$ 为实值的假设空间, 则对于任意 $\delta > 0$, 任意 $h \in \mathcal{H}$ 以及 $\rho \in (0,1)$, 下式以 $1-\delta$ 的概率成立*

$$E(h) \leq \hat{E}_\rho(h) + \frac{4}{\rho} R_m(\mathcal{H}) + \sqrt{\frac{\ln \log_2 \frac{2}{\rho}}{m}} + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}},$$

$$E(h) \leq \hat{E}_\rho(h) + \frac{4}{\rho} \hat{R}_D(\mathcal{H}) + \sqrt{\frac{\ln \log_2 \frac{2}{\rho}}{m}} + 3\sqrt{\frac{\ln \frac{4}{\delta}}{2m}}.$$

上面提过

$$\hat{R}_D(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{m}},$$

两边取期望可得 $R_m(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{m}}$, 再结合前面两个定理可得如下推论.

推论 5.3. *令 $\mathcal{H} = \{\boldsymbol{x} \mapsto \boldsymbol{w} \cdot \boldsymbol{x} : \|\boldsymbol{w}\| \leq \Lambda\}$ 并且 $X \subset \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq r\}$, 对于固定的 $\rho > 0$, 下式对于任意 $h \in \mathcal{H}$ 和 $\delta > 0$ 以 $1-\delta$ 的概率成立*

$$E(h) \leq \hat{E}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

5.4.2 Adaboost 的泛化误差界

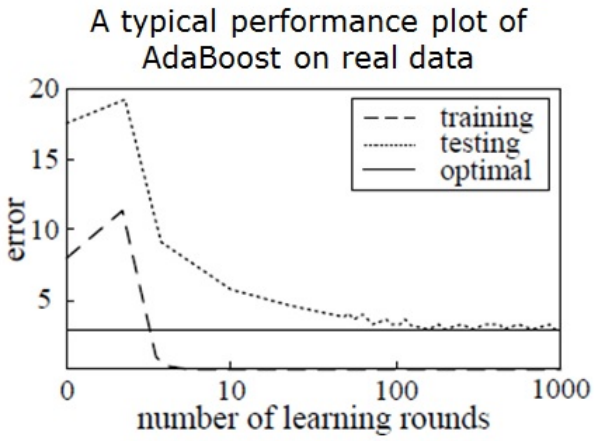

图 4: *AdaBoost* 算法不易陷入过拟合现象

图 4表明: AdaBoost 算法即使训练轮数增加, 仍不易陷入过拟合, 甚至当训练错误率为零时还可以降低泛化错误率. 这一现象显然违背了奥卡姆剃刀原理: 在与经验数据一致的模型中, 模型越简单, 泛化性越好. 为解释 AdaBoost 的这种现象, Schapire 等人于 1998 年提出了著名的间隔理论.

给定基学习器的函数空间 $\mathcal{H}$, 定义函数空间 $\mathcal{H}$ 的凸包 $\mathcal{C}(\mathcal{H})$ 为

$$\mathcal{C}(\mathcal{H}) = \left\{ H = \sum_i \alpha_i h_i \colon h_i \in \mathcal{H}, \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1 \right\}.$$

下面给出 AdaBoost 的间隔的定义:

定义 5.4. 给定一个样本 (x,y) 和 *AdaBoost* 算法返回的学习器 $H(x)=\sum_{t=1}^{T}\alpha_t h_t(x)\in\mathcal{C}(\mathcal{H})$, 间隔定义为 $yH(x)$, 即

$$yH(x)=\sum_{i:\ y=h_i(x)}\alpha_i-\sum_{i:\ y\neq h_i(x)}\alpha_i.$$

间隔计算正确/错误分类样本 (x,y) 的基分类器权重之差, 在一定程度上反映了分类器的可信度.

根据间隔定义, 下面给出著名的间隔理论:

定理 5.13. 假设 $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ 是一个大小为 m 的训练集. 对任意 δ 和 $\theta>0$, 每个分类器 $H\in\mathcal{C}(\mathcal{H})$ 至少以 $1-\delta$ 的概率满足:

$$P_{\mathcal{D}}(yH(x)<0)\leq P_S(yH(x)\leq\theta)+O\left(\frac{1}{\sqrt{m}}\left(\frac{\ln m\ln|\mathcal{H}|}{\theta^2}+\ln\frac{1}{\delta}\right)^{1/2}\right),$$

其中 $|\mathcal{H}|$ 表示空间 $\mathcal{H}$ 所包含函数的个数 (这里讨论有限情况), $P_D(yH(x)<0)$ 表示学习器 H 在真实分布 D 上的泛化错误率, $P_S(yH(x)\leq\theta)=\sum_{(x_i,y_i)}I(y_if(x_i)\leq\theta)/m$ 表示在训练集 S 中间隔小于 θ 的样本所占比例.

该定理表明 AdaBoost 的泛化性与训练集的大小, 基学习器以及间隔 θ 相关, 但与算法的迭代轮数无关. AdaBoost 算法当训练错误率为零时不停止训练, 可以进一步增大间隔 θ, 从而提高学习方法的泛化性而没有过拟合, 间隔理论成功解释了 AdaBoost 不易过拟合现象.

证明. 关键证明技术: Chernoff 不等式和全概率公式 $P(A)=P(A\cap B)+P(A\cap\bar{B})$.

对每个函数 $H=\sum_{i=1}^{T}\alpha_i h_i\in\mathcal{C}(\mathcal{H})$, 其系数 α_i 可以看作选择函数 h_i 的概率. 从而每个函数 H 与函数空间 $\mathcal{H}$ 的分布建立联系, 记为 $\mathcal{Q}(H)$, 即 $P_{h\sim\mathcal{Q}(H)}(h=h_i)=\alpha_i$. 给定整数 $N>0$, 假设函数 $g=\sum_{j=1}^{N}h_j/N$ 其中 $h_j\sim\mathcal{Q}(H)$, 即函数 g 是根据分布 $\mathcal{Q}(f)$ 从 $\mathcal{Q}(H)$ 中随机选择 N 个函数的平均. 于是有

$$E(g)=E_{h_j\sim\mathcal{Q}(f)}\left(\sum_{j=1}^{N}\frac{h_j}{N}\right)=E_{h_1\sim\mathcal{Q}(f)}(h_1)=\sum_{i=1}^{T}\alpha_i h_i\in\mathcal{C}(\mathcal{H})=H.$$

主要证明分为下面三步:

1. (I:) $P_{\mathcal{D}}(yH(x)\leq 0)\Longleftrightarrow P_{\mathcal{D}}(yg(x)\leq\theta/2)$
2. (II:)$P_{\mathcal{D}}(yg(x)\leq\theta/2)\Longleftrightarrow P_S(yg(x)\leq\theta/2)$
3. (III:)$P_S(yg(x)\leq\theta/2)\Longleftrightarrow P_S(yH(x)\leq\theta)$

对于第 I 步, 由全概率公式可得

$$\begin{aligned}&P_{\mathcal{D}}(yH(x)\leq 0)\\ &=\ P_{\mathcal{D}}(yH(x)\leq 0,yg(x)\leq\theta/2)+P_{\mathcal{D}}(yH(x)\leq 0,yg(x)>\theta/2)\\ &\leq\ P_{\mathcal{D}}(yg(x)\leq\theta/2)+P_{\mathcal{D}}(yH(x)\leq 0,yg(x)>\theta/2).\end{aligned}$$

由 $P(A\cap B)=P(A)P(A|B)\le P(A|B)$, 以及 Chernoff 不等式可知

$$P_{\mathcal{D}}(yH(x)\le 0, yg(x)>\theta/2)\le P_{\mathcal{D}}(yg(x)>\theta/2|yH(x)\le 0)\le \exp(-N\theta^2/8).$$

于是得到

$$P_{\mathcal{D}}(yH(x)\le 0)\le P_S(yg(x)\le\theta/2)+\exp(-N\theta^2/8). \tag{27}$$

对于第 II 步, 首先发现

$$\begin{aligned} P_S(yg(x)\le\theta/2) &= \frac{1}{m}\sum_{i=1}^{m} I(y_i g(x_i)\le\theta/2) \\ P_{\mathcal{D}}(yg(x)\le\theta/2) &= E_S(P_S(yg(x)\le\theta/2)). \end{aligned}$$

利用 Union 不等式和 Chernoff 不等式可得到

$$P_{\mathcal{S}\sim\mathcal{D}^m}\left(P_{\mathcal{D}}(yg(x)\le\theta/2)\ge P_{\mathcal{S}}(yg(x)\le\theta/2)+\epsilon\right)\le (N+1)|\mathcal{H}|^N\exp(-2m\epsilon^2).$$

令 $\delta=(N+1)|\mathcal{H}|^N\exp(-2m\epsilon^2)$, 则下面表达式至少以 $1-\delta$ 的概率成立:

$$P_{\mathcal{D}}(yg(x)\le\theta/2)<P_{\mathcal{S}}(yg(x)\le\theta/2)+\sqrt{(1/2m)\ln((N+1)|\mathcal{H}|^N/\delta)}. \tag{28}$$

对于第 III 步, 再次利用全概率公式有

$$\begin{aligned} &P_S(yg(x)\le\theta/2) \\ &= P_S(yg(x)\le\theta/2, yH(x)\le\theta)+P_S(yg(x)\le\theta/2, yH(x)>\theta) \\ &\le P_S(yH(x)\le\theta)+P_S(yg(x)\le\theta/2, yH(x)>\theta). \end{aligned}$$

由 Chernoff 不等式可知

$$P_S(yg(x)\le\theta/2, yH(x)>\theta)\le P_S(yg(x)\le\theta/2|yH(x)>\theta)\le\exp(-N\theta^2/8).$$

于是有

$$P_S(yg(x)\le\theta/2)\le P_S(yH(x)\le\theta)+\exp(-N\theta^2/8). \tag{29}$$

根据式(27)-(29), 下式至少以 $1-\delta$ 的概率成立

$$P_{\mathcal{D}}(yf(x)\le 0)\le P_{\mathcal{S}}(yf(x)\le\theta)+2\exp(-N\theta^2/8)+\sqrt{(1/2m)\ln((N+1)|\mathcal{H}|^N/\delta)}$$

通过令 $N=8\ln m/\theta^2$ 以及整理计算, 定理得证. ■

习题 5.1. 假设 *AdaBoost* 算法每轮迭代过程中基学习器产生的错误率分别为 $\epsilon_1,\epsilon_2,\ldots,\epsilon_T$. 对任何 $\theta>0$, 最终产生的学习器 $H(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$ 满足

$$\frac{1}{m}\sum_{i=1}^{m} I(y_i H(x_i)\le\theta)\le 2^T\prod_{t=1}^{T}\sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{\theta}}.$$

接下来将利用 Rademacher 复杂度来研究 AdaBoost 算法的间隔泛化性, 经验 Rademacher 复杂度和分布 Rademacher 复杂度相关定义如下:

定义 5.5. *给定样本数据 $Z=\{z_1,\ldots,z_m\}$, 函数空间 $\mathcal{H}$ 关于 Z 的经验 Rademacher 复杂度为*

$$\hat{R}_Z(\mathcal{H})=\mathbb{E}_\sigma\Big(\sup_{f\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\Big).$$

函数空间 $\mathcal{F}$ 关于 $\mathcal{Z}$ 在分布 $\mathcal{D}$ 上的 Rademacher 复杂度为

$$R_m(\mathcal{F})=\mathbb{E}_{Z\subset\mathcal{Z}:|Z|=m}\Big(\hat{R}_Z(\mathcal{F})\Big).$$

基于上述定义有

定理 5.14. *对实值函数空间 $\mathcal{H}:\mathcal{Z}\to(0,1)$, 根据分布 $\mathcal{D}$ 从 $\mathcal{Z}$ 独立同分布采样得到示例集 $Z=\{z_1,z_2,\ldots,z_m\}$,$z_i\in\mathcal{Z}$, $0<\delta<1$, 对任意 $f\in\mathcal{F}$, 下列不等式以至少 $1-\delta$ 的概率成立:*

$$\begin{aligned}E(f(z)) &\le \frac{1}{m}\sum_{i=1}^{m}f(z_i)+2R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{2m}},\\ E(f(z)) &\le \frac{1}{m}\sum_{i=1}^{m}f(z_i)+2\hat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}.\end{aligned}$$

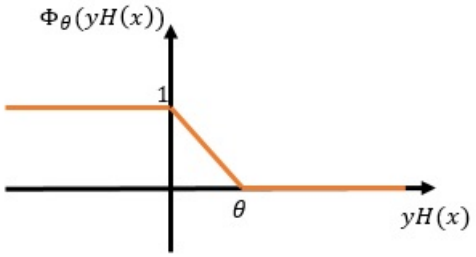

图 5: θ-间隔损失函数 Φ_θ

下面研究基于 Rademacher 复杂度的 Boosting 间隔分布理论, 首先引入一个新的函数.

定义 5.6. *给定任何 $\theta>0$, 定义 θ-间隔损失函数 Φ_θ 为*

$$\Phi_\theta(t)=\begin{cases}1 & t<0\\ 1-t/\theta & t\in(0,\theta)\\ 0 & t>\theta\end{cases}$$

图 5 给出了 θ-间隔损失函数 Φ_θ 的线段, 进一步得到如下性质:

引理 5.3. *对任意给定样本 (x,y) 和函数 H, 有如下关系成立*

$$I(yH(x)<0)\le\Phi_\theta(yH(x))\le I(yH(x)<\theta).$$

引理 5.4. 给定实值函数空间 $\mathcal{H}$, 假设函数 $\phi\colon \mathbb{R}\to\mathbb{R}$ 满足李普西兹条件且其系数为 L, 有

$$R_S(\phi\circ\mathcal{H})\le LR_S(\mathcal{H}) \quad 以及 \quad R(\phi\circ\mathcal{H})\le LR(\mathcal{H}).$$

引理 5.5. 对任意给定的函数空间 $\mathcal{H}$, 以及样本数据 $Z=\{z_1,z_2,\ldots,z_m\}$, 有

$$R_S(\mathcal{C}(\mathcal{H}))=R_S(\mathcal{H})$$

成立, 其中 $\mathcal{C}(\mathcal{H})=\{H=\sum_{t=1}^T\alpha_t h_t\colon \alpha_t\ge 0,\sum_t\alpha_t=1,h_t\in\mathcal{H}\}$.

证明. 容易得到

$$\begin{aligned}
R_S(\mathcal{C}(\mathcal{H})) &= \frac{1}{m}E_\sigma\left(\sup_{H\in\mathcal{C}(\mathcal{H})}\sum_{i=1}^m\sigma_i H(z_i)\right)\\
&= \frac{1}{m}E_\sigma\left(\sup_{h_t\in\mathcal{H},\sum\alpha_t=1}\sum_{t=1}^T\alpha_t\sum_{i=1}^m\sigma_i h_t(z_i)\right)\\
&= \frac{1}{m}E_\sigma\left(\sup_{h_t\in\mathcal{H}}\max_{t\in(T)}\sum_{i=1}^m\sigma_i h_t(z_i)\right)\\
&= \frac{1}{m}E_\sigma\left(\sup_{h_t\in\mathcal{H}}\sum_{i=1}^m\sigma_i h_t(z_i)\right)=R_S(\mathcal{H}).
\end{aligned}$$

引理得证. ∎

下面给出基于 Rademacher 复杂度的间隔理论:

定理 5.15. 假设 $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ 是一个大小为 m 的训练集. 对任意 δ 和 $\theta>0$, 每个分类器 $H\in\mathcal{C}(\mathcal{H})$ 至少以 $1-\delta$ 的概率满足:

$$P_{\mathcal{D}}(yH(x)<0) \le P_S(yH(x)\le\theta)+\frac{2}{\theta}R_m(\mathcal{H})+\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{30}$$

$$P_{\mathcal{D}}(yH(x)<0) \le P_S(yH(x)\le\theta)+\frac{2}{\theta}R_S(\mathcal{H})+3\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{31}$$

习题 5.2. 证明上述定理中式(31).

证明. 假设函数空间 $\hat{\mathcal{H}}=\{z=(x,y)\to yH(x)\colon H\in\mathcal{C}(\mathcal{H})\}$, 考虑一个新的函数集合

$$\bar{\mathcal{H}}=\{\Phi_\theta\circ f\colon f\in\hat{\mathcal{H}}\}.$$

首先根据引理 5.3可知

$$P_{\mathcal{D}}(yH(x)<0)\le P_{\mathcal{D}}(\Phi_\theta(yH(x))).$$

再根据定理 5.14得到

$$P_{\mathcal{D}}(\Phi_\theta(yH(x)))\le P_S(\Phi_\theta(yH(x)))+2R_m(\bar{\mathcal{H}})+\sqrt{\frac{\ln(1/\delta)}{2m}}.$$

由于函数 Φ_θ 满足李普西兹条件, 且其李普西兹系数为 $1/\theta$, 根据引理 5.4可得

$$\begin{aligned} R_m(\bar{\mathcal{H}}) = R_m(\Phi_\theta \circ \hat{\mathcal{H}}) &= \frac{1}{\theta} R_m(\hat{\mathcal{H}}) \\ &= \frac{1}{\theta m} \sum_{i=1}^{m} E_\sigma \left(\sup_{H \in \mathcal{C}(\mathcal{H})} \sum_{i=1}^{m} \sigma_i y_i H(x_i) \right) \\ &= \frac{1}{\theta m} \sum_{i=1}^{m} E_\sigma \left(\sup_{H \in \mathcal{C}(\mathcal{H})} \sum_{i=1}^{m} \sigma_i H(x_i) \right) \\ &= \frac{1}{\theta} R_m(\mathcal{C}(\mathcal{H})) = \frac{1}{\theta} R_m(\mathcal{H}), \end{aligned}$$

其中最后一个等式成立是因为引理 5.5. 式(30)得证. ■

注: 本章 SVM 的间隔理论以及 SVM 基于间隔理论的泛化误差界参考了 [Mohri et al., 2012] 一书第 4 章的内容.

参考文献

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320.

Abernethy, J., Bartlett, P. L., Rakhlin, A., and Tewari, A. (2008a). Optimal stragies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 415–423.

Abernethy, J., Hazan, E., and Rakhlin, A. (2008b). Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 263–274.

Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 28–40.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Dani, V., Hayes, T. P., Kakade, S. M., Rakhlin, A., and Tewari, A. (2008). High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 335–341.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Cesa-bianchi, N., Conconi, A., and Gentile, C. (2002). On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 359–366.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008a). The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*, pages 345–352.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008b). Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning*, pages 355–366.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. (2015). Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1405–1411.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations, 3rd Edition*. Johns Hopkins University Press.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

Hazan, E. and Kale, S. (2011). Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436.

Hazan, E. and Kale, S. (2014). Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512.

Hazan, E. and Seshadhri, C. (2007). Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity*, 88.

Hazan, E. and Seshadhri, C. (2009). Efficient learning algorithms for changing environments. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 393–400.

Hou, B.-J., Zhang, L., and Zhou, Z.-H. (2017). Learning with feature evolvable streams. In *Advances in Neural Information Processing Systems 30*.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pages 99–109.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pages 440–447.

Kakade, S. M. and Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, pages 801–808.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition.

Mahdavi, M., Zhang, L., and Jin, R. (2015). Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Annual Conference on Learning Theory*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A., editors (2012). *Foundations of Machine Learning*. MIT Press, Cambridge, MA.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization*. Kluwer Academic Publishers.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Core discussion papers.

Nesterov, Y. (2011). Random gradient-free minimization of convex functions. Core discussion papers.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 636–642.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory - Series A*, 13(1):145–147.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009a). Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Shalev-Shwartz, S., Shamir, O., Sridharan, K., and Srebro, N. (2009b). Learnability and stability in the general learning setting.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007). Pegasos: primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, pages 807–814.

Shelah, S. (1972). A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261.

Srebro, N., Sridharan, K., and Tewari, A. (2010). Optimistic rates for learning with a smooth loss. *ArXiv e-prints*, arXiv:1009.3896.

Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009). Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552.

Tseng, P. (2008). On acclerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Vapnik, V. N. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280.

Yang, T., Zhang, L., Jin, R., and Yi, J. (2016). Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 449–457.

Zhang, L., Yang, T., and Jin, R. (2017). Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds. In *Proceedings of the 30th Annual Conference on Learning Theory*, pages 1954–1979.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-H. (2016). Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*.

Zhang, L., Yang, T., Jin, R., and Zhou, Z.-H. (2018). Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936.