机器学习理论导引
作业二

DZ1833019，欧先飞，ouxianfei@smail.nju.edu.cn

2019 年 4 月 22 日

1 [30pts] Generalization

机器学习中，我们总会通过先验知识对假设空间进行限制. 例如 SVM 中使用的典型超平面族 $\mathcal{H} = \left\{ \boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} : \|\boldsymbol{w}\| \leq \Lambda \right\}$（见讲义的定理 5.10）. 因为虽然大的假设空间更可能包含目标概念，但对应的学习难度即样本复杂度也随之增大，从而导致泛化性变差.

(1) **[10pts]** 试通过 VC 维的泛化误差界来解释对假设空间进行限制的合理性。

(2) **[10pts]** 试通过 Rademacher 的泛化误差界来解释对假设空间进行限制的合理性。

(3) **[10pts]** 二者的泛化误差界哪个更紧？为什么？

Proof.
(1) 因为 $\Pr[E(h) \leq \widehat{E}(h) + \left(\frac{em}{d}\right)^d + \sqrt{\frac{\ln 1/\delta}{2m}}] \geq 1 - \delta$，而当对假设空间进行限制时，相应的该假设空间的 VC 维也倾向于降低 (因为假设空间更小了，VC 维无论如何不会变大)，从该不等式获得的泛化误差的上界也会相应的变小，从而学习算法的泛化性能可以更好。
(2) 因为 $\Pr[E(h) \leq \widehat{E}(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln 1/\delta}{2m}}] \geq 1 - \delta$，当假设空间受限之后，$\mathfrak{R}_m(\mathcal{H}) = E_{D^m,\sigma}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)]$ 式中 $\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)$ 也将倾向于更小 (无论如何不会更大)，从而由该不等式获得的泛化误差的上界也会更小。
(3) 基于 *Rademacher* 的更紧，因为 $\mathfrak{R}_m(\mathcal{H}) \leq \left(\frac{em}{d}\right)^d$。

$\square$

2 [20pts] Stability

(1) **[10pts]** 为了应对未知的测试情况，实际机器学习算法在选择超参数取值时，通常通过交叉验证的方式来估计泛化能力。请讨论留一法交叉验证估计学习算法泛化能力的合理性（从稳定性的角度进行分析；留一法交叉验证参考周志华《机器学习》26 页）。

(2) **[10pts]** 假设讲义中定理 6.1 所需的条件均满足，如果算法非常稳定，即 $\beta \to 0$，是否可以通过同样的分析得到优于 $\mathcal{O}(1/\sqrt{m})$ 的泛化界？

Proof.
(1) 从稳定性角度来看，$\Pr[\ell(\mathfrak{L}, D) \leq \ell_{loo}(\mathfrak{L}, D) + \beta + (4m\beta + M)\sqrt{\frac{\ln 1/\delta}{2m}}] \geq 1 - \delta$，当使用留一法计算所得的损失较小时，整体的泛化损失在同等概率下也会更小，所以留一法用于评估模型的泛化能力是比较合理的。
(2) *slides* 中推导所得的泛化损失的界 $\Pr[\ell(\mathfrak{L}, D) \leq \widehat{\ell}(\mathfrak{L}, D) + \gamma + (2m\gamma + M)\sqrt{\frac{\ln 1/\delta}{2m}}] \geq 1 - \delta$，容易发现该上界由三个渐进项构成 $\mathcal{O}(\gamma)$、$\mathcal{O}(\gamma\sqrt{m})$ 和 $\mathcal{O}(\frac{1}{\sqrt{m}})$，所以无论 γ 取什么样的渐进函数，或者直接取 *0*，该不等式的泛化界都不会优于 $\mathcal{O}(\frac{1}{\sqrt{m}})$。

$\square$

3 [20pts] Optimality of Bayes Classifier

对任意定义在 $\mathcal{X} \times \{0,1\}$ 上的概率分布 $\mathcal{D}$，考虑所有分类器 $g: \mathcal{X} \mapsto \{0,1\}$，特定的，记 $f_{\mathcal{D}}$ 为 Bayes 分类器，其定义如下：

$$f_{\mathcal{D}} = \begin{cases} 1, & \text{if } \Pr[y=1|x] \geq 1/2 \\ 0, & \text{otherwise} \end{cases}$$

试证明，Bayes 分类器 $f_{\mathcal{D}}$ 是最优的，即对任何分类器 g，有 $R(f_{\mathcal{D}}) \leq R(g)$，其中 $R(g)$ 是分类器 g 在未知数据分布 $\mathcal{D}$ 的泛化误差，$R(g) = \Pr_{(x,y)\sim\mathcal{D}}[g(x) \neq y]$。

Proof. 首先证明 $\Pr[h(x)=k, y=k|x=x_0] = \Pr[h(x)=k|x=x_0]\Pr[y=k|x=x_0], k \in 0, 1$，由于当 x 给定时，$h(x)$ 只可能是确定的 *0* 或者 *1*，所以直接对 $h(x_0)$ 的取值进行讨论。
假设 $k=0$，如果 $h(x_0)=0$，那么 $\Pr[h(x)=k, y=k|x=x_0] = \Pr[h(x)=k|x=x_0]\Pr[y=k|x=x_0] = \Pr[y=0|x=x_0]$，如果 $h(x_0=1)$，那么 $\Pr[h(x)=k, y=k|x=x_0] = \Pr[h(x)=k|x=x_0]\Pr[y=k|x=x_0] = 0$，所以当 $k=0$ 时，$\Pr[h(x)=k, y=k|x=x_0] = \Pr[h(x)=k|x=x_0]\Pr[y=k|x=x_0]$，同理当 $k=1$ 时等式也成立。继而有：

$$\begin{aligned}
\Pr[h(x)=y|x=x_0] &= \Sigma_{k\in\mathcal{Y}} \Pr[h(x)=k|x=x_0]\Pr[y=k|x=x_0] \\
&= \Sigma_{k\in\mathcal{Y}} \mathbb{I}[h(x_0)=k]\Pr[y=k|x=x_0] \\
&= \mathbb{I}[h(x_0)=0]\Pr[y=0|x=x_0] + \mathbb{I}[h(x_0)=1]\Pr[y=1|x=x_0]
\end{aligned}$$

然后对 f_D 与 g 进行逐差，令 $\Delta = \Pr[f_D(x)=y|x=x_0] - \Pr[g(x)=y|x=x_0]$，则有：

$$\begin{aligned}
\Delta &= \Pr[y=0|x=x_0](\mathbb{I}[f_D(x_0)=0] - \mathbb{I}[g(x_0)=0]) \\
&\quad + \Pr[y=1|x=x_0](\mathbb{I}[f_D(x_0)=1] - \mathbb{I}[g(x_0)=1]) \\
&= (1-\Pr[y=1|x=x_0])(\mathbb{I}[g(x_0)=1] - \mathbb{I}[f_D(x_0)=1]) + \\
&\quad \Pr[y=1|x=x_0](\mathbb{I}[f_D(x_0)=1] - \mathbb{I}[g(x_0)=1]) \\
&= (2\Pr[y=1|x=x_0]-1)(\mathbb{I}[f_D(x_0)=1] - \mathbb{I}[g(x_0)=1])
\end{aligned}$$

当 $\Pr[y=1|x=x_0] \geq \frac{1}{2}$ 时，由 f_D 的定义有 $\mathbb{I}[f_D(x_0)=1]=1$，又因为 $\mathbb{I}[g(x_0)=1] \leq 1$，所以 $\Delta \geq 0$。当 $\Pr[y=1|x=x_0] < \frac{1}{2}$ 时，由 f_D 的定义有 $\mathbb{I}[f_D(x_0)=1]=0$，又因为 $\mathbb{I}[g(x_0)=1] \geq 0$，所以 $\Delta \geq 0$。综上，对于任意的 g，始终有 $\Delta = \Pr[f_D(x)=y|x=x_0] \geq \Pr[g(x)=y|x=x_0]$，其等价于 $\Delta = \Pr[f_D(x) \neq y|x=x_0] \leq \Pr[g(x) \neq y|x=x_0]$，也就是 $R(f_D) \leq R(g)$，所以 *Bayes* 分类器 f_D 是最优的。

□

4 [30pts] SVM with Squared Hinge Loss Function

在支持向量机（Support Vector Machine，SVM）的实际使用中，人们经常采用平方 hinge 损失函数（squared hinge loss function）。记损失函数为 $\ell : \mathcal{Y}' \times \mathcal{Y} \mapsto \mathbb{R}_+$，其中 $\mathcal{Y}' \subset \mathbb{R}$ 且 $\mathcal{Y} = \{-1, +1\}$，平方 hinge 损失函数的定义可写为

$$\ell(y', y) = ([1 - yy']_+)^2, \tag{4.1}$$

其中符号 $[x]_+$ 表示取 x 的非负部分，即 $[x]_+ = x$ 如果 $x \geq 0$；否则 $[x]_+ = 0$. 本题目中，我们采用第六讲中所讲授的稳定性工具对平方 hinge 损失 SVM 的泛化性进行分析。

(1) **[10pts]** 假设对于任意的分类器 $h \in \mathcal{H}$ 及样本 $x \in \mathcal{X}$，均有 $|h(x)| \leq M$，试证明平方 hinge 损失函数是有界的，并给出上界。

(2) **[20pts]** 试利用稳定性分析工具推导基于平方 hinge 损失 SVM 的泛化界。请给出严格的结论表述和具体的推导过程。

Proof.
(1) 首先容易证明 $[a+b]_+ \leq |a| + |b|$，所以有 $\ell(h(x), y) = ([1 - h(x)y]_+)^2 \leq (1 + |h(x)|)^2 \leq (1+M)^2$。
(2) 由于 $|\ell(h(a), y) - \ell(h(b), y)| = |([1 - h(a)y]_+)^2 - ([1 - h(b)y]_+)^2| \leq |(1 + |h(a)y|)^2 - (1 + |h(b)y|)^2| = |(1+|h(a)|)^2 - (1+|h(b)|)^2| = |(h(a)+h(b)+2)||h(a)-h(b)| \leq (2M+2)|h(a)-h(b)|$，所以损失函数 ℓ 对于假设空间 $\mathcal{H}$ 是 σ-可容许的，其中 $\sigma = 2M+2$。然后由 *slides* 上的命题 *6.1* 知 *SVMs* 具有 γ-稳定性，$\gamma < \frac{4(M+1)^2 r^2}{m\lambda}$，所以 $\ell(\mathcal{L}, \mathcal{D}) \leq \widehat{\ell}(\mathcal{L}, D) + \gamma + (2m\gamma + M)\sqrt{\frac{\ln 1/\delta}{2m}} < \widehat{\ell}(\mathcal{L}, D) + \frac{4(M+1)^2 r^2}{m\lambda} + (\frac{8(M+1)^2 r^2}{\lambda} + M)\sqrt{\ln 1/\delta} 2m$，以至少 $1-\delta$ 的概率成立。

□