机器学习第三次作业

2020年10月

1 第一题

定义函数族$\mathcal{G}$ (从输入集$\mathcal{Z}$映射到[0,1]) 在从分布D中抽样得到的容量为n的数据集$\mathcal{S}_n = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$上的经验Rademacher复杂度为:

$$\widehat{\mathcal{R}}_{\mathcal{S}_n}(\mathcal{G}) = E_{\boldsymbol{\sigma}} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g\left(\mathbf{z}_i\right) \tag{1}$$

其相应的Rademacher复杂度可以被定义为:

$$\mathcal{R}_n(\mathcal{G}) = E_{\mathcal{S}_n \sim D^n} \widehat{\mathcal{R}}_{\mathcal{S}_n}(\mathcal{G}) \tag{2}$$

变量σ_i是满足$P\left(\sigma_i = 1\right) = P\left(\sigma_i = -1\right) = \frac{1}{2}$的独立同分布的随机变量。请使用MicDiarmid不等式证明对于任意$\delta > 0$, 下列式子以不小于$1-\delta$的概率成立:

$$\mathcal{R}_n(\mathcal{G}) \leq \widehat{\mathcal{R}}_{\mathcal{S}_n}(\mathcal{G}) + O\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}}\right) \tag{3}$$

2 第二题

固定n≥1, 对于任意$\alpha \in R$ (R为实数集) 和任意两个从输入集$\mathcal{X}$映射到R的函数假设集(hypothesis sets) $\mathcal{H}$和$\mathcal{H}'$, 证明下列等式:

$$\begin{array}{l} \text{(a) } \mathcal{R}_n(\alpha \mathcal{H}) = |\alpha| \mathcal{R}_n(\mathcal{H}) \\ \text{(b) } \mathcal{R}_n\left(\mathcal{H} + \mathcal{H}'\right) = \mathcal{R}_n(\mathcal{H}) + \mathcal{R}_n\left(\mathcal{H}'\right) \end{array} \tag{4}$$

上式中$\mathcal{H} + \mathcal{H}' = \{h + h' | \forall h \in \mathcal{H}, \forall h' \in \mathcal{H}'\}$, $\alpha \mathcal{H} = \{\alpha h | \forall h \in \mathcal{H}\}$。

3 第三题

设$\mathcal{H}$是一个从$\mathcal{X}$映射到$\{0,1\}$的可数假设集, p是$\mathcal{H}$上的先验分布。请使用Hoeffding不等式证明对于任意$\delta > 0$, 下列式子以不小于$1-\delta$的概率成立:

$$\forall h \in \mathcal{H}, \epsilon(h) \leq \widehat{\epsilon}_{S_n}(h) + O\left(\sqrt{\frac{\log \frac{1}{p(h)} + \log \frac{1}{\delta}}{n}}\right) \tag{5}$$

请把以上误差上界和课件中有限假设集的上界进行比较，谈谈自己的理解。(提示：可以把$\delta' = p(h)\delta$代入Hoeffding不等式中。)