

机器学习基础作业 1

原梓轩

2025 年 2 月 19 日

问题 1. 证明: $E_{T' \sim D^{N'}} [\hat{R}_{\text{test}(h_T)}] = R(h_T)$.

证明.

$$\begin{aligned} E_{T' \sim D^{N'}} [\hat{R}_{\text{test}(h_T)}] &= E_{T' \sim D^{N'}} \left[\frac{1}{N'} \sum_{i=1}^{N'} L(h_T(x_i), y_i) \right] \\ &= \frac{1}{N'} \sum_{i=1}^{N'} E_{T' \sim D^{N'}} [L(h_T(x_i), y_i)] \\ &= \frac{1}{N'} \sum_{i=1}^{N'} E_{x_i \sim D} [L(h_T(x_i), y_i)] \\ &= \frac{1}{N'} \sum_{i=1}^{N'} E_{x_i \sim D} [L(h_T(x_i), y_i)] \\ &= \frac{1}{N'} \sum_{i=1}^{N'} R(h_T) \\ &= R(h_T) \end{aligned}$$

□

问题 2. 比较交叉验证过法和自助法的异同.

证明. 交叉验证法和自助法都是用来估计模型的泛化误差的方法, 并且可以重复利用数据集.

交叉验证法将数据集分为 k 个大小相等的子集, 每次取其中一个子集作为验证集, 其余的子集作为训练集, 这样可以得到 k 个模型, 最后将这 k 个模型的泛化误差的平均值作为最终的泛化误差. 自助法是通过数据集进行有放回的采样, 从而得到一个大小为 N 的训练集, 然后用这个训练集训练模型, 将其余的数据作为测试集, 从而得到泛化误差.

交叉验证法可以更多次平均地利用数据集, 保证数据训练的稳定性, 但是需要训练 k 个模型, 计算量较大. 自助法引入了更多样本随机性, 降低了过拟合的可能. 但是由于是有放回的采样, 会有一部分数据被多次采样或没有被采样到, 从而选取的数据分布未必与原先一致, 因而如果模型对数据分布敏感则不适用.

□