第七章 模型评估

一、填空题

1 在当前训练集上经验误差为 0 的假设集合称为（　　）；符合参数化模型函数形式的所有映射关系的集合称为（　　）；

2 当两个假设在训练集上具有完全相同的经验误差时，应依据（　　）来进行假设选择。

3 函数$h(x;a,b)=\mathrm{sign}(\mathrm{a}x-b),\mathrm{a},\mathrm{b}\in\Re$支撑的假设空间为$\mathcal{H}$，则增长函数$\Pi_{\mathcal{H}}(3)=$（　　）

4 线性分类器$h(\boldsymbol{x})=w_1x_1+w_2x_2+w_0$的 VC 维是（　　）

5 使用 10-折交叉验证法处理 Iris 数据集，则每一次验证时训练集中的 setosa 类样本数量是（　　），测试集的样本数量是（　　）

二、判断题

6 一个高效 PAC 可学问题一定是 PAC 可学问题。（　　）

7 模型在独立同分布采样获得的训练集上的经验误差总是大于该模型在该分布上的泛化误差。（　　）

8 若存在假设$h\in\mathcal{H}$在数据集 D 上的经验误差为零，则有真实映射$f\in\mathcal{H}$。（　　）

三、选择题

9 当样本数为N，精度为ϵ，置信度为δ时，下述那种学习时间T对应的学习算法是 PAC 学习算法

A. $T=N!$　　B. $T=N\log N+2^{1/\epsilon}$

C. $T=N\log N+\left(\frac{1-\delta}{0.05}\right)^5$　　D. $T=\frac{2^N}{\delta^2}$

四、简答题

10 请简述“奥卡姆剃刀”原理，并说明其在模型评估中的应用。

11 请简述“没有免费午餐”定理，并说明其对于机器学习研究的意义。

12 请简要阐述经验误差与泛化误差之间的关系。

13 请简述 PAC 学习理论中 P，A 的概念及其对应于机器学习结果的哪些性能。

14 请简述不可知 PAC 学习与可知 PAC 学习的最主要差别

15 请结合本课程讲授的分类算法，说明三种具体的正则化方法。

五、计算（画图）题

16 考虑一个二分类问题，$x \in \Re^2, y \in \{+1, -1\}$. 给定一个假设空间如下：

$$\mathcal{H} = \{h(x; a, b) = \text{sign}[a(x^T x - b)] | a \in \{+1, -1\}, b \in [0, +\infty)\}$$

其中，函数$\text{sign}(x) = \begin{cases} -1, & x < 0 \\ +1, & x \geq 0 \end{cases}$。请通过画图法找出假设空间$\mathcal{H}$的 VC 维 d 的具体取值（提示：画出$\mathcal{H}$打散 d 个样本但无法打散 d+1 个样本的情况）

17 某二分类任务训练样本集为：$\boldsymbol{x}^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, x_3^{(i)}\right]^T, i = 1, \ldots, 100$，其中特征$x_1$有 3 个取值，特征$x_2$有 2 个取值，特征$x_3$有 2 个取值。某个集成分类器算法$\mathcal{L}_1$的假设空间$\mathcal{H}_1$包含该问题上所有可能的映射。

1) 请计算出假设空间$\mathcal{H}_1$的大小。
2) 请判断算法$\mathcal{L}_1$能否在当前数据集下以不低于 80%的概率获得泛化误差不超过 0.1 的识别模型，并给出计算过程。
3) 如果存在一个算法$\mathcal{L}_2$，可以生成 512 种不同的假设，且在当前数据机上的误差为 0.05。试计算$\mathcal{L}_2$在该问题上的泛化误差上界。

18 已知测试集真实标签 Y 和分类器分类结果 H 如下

样本序号	1	2	3	4	5	6	7	8	9	10
Y	P	N	N	P	P	P	N	P	N	P
H	P	N	P	P	N	P	P	P	N	N

请计算以下性能指标，并给出计算方法。

1) 真阳率
2) 假阳率
3) 查准率
4) 查全率
5) F1 度量