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（版本空间）；符合参数化模型函数形式的所有映射关系的集合称为（假设空间）。

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（归纳偏好）来进行假设选择。

3 函数$h(x;a,b)=\text{sign}(ax-b),a,b\in\Re$支撑的假设空间为$\mathcal{H}$，则增长函数$\Pi_{\mathcal{H}}(3)=$(4)

4 线性分类器$h(\boldsymbol{x})=w_1x_1+w_2x_2+w_0$的 VC 维是（3）

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（45）。验证集的样本数量是（15）

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小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：（C）

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。

答：

奥卡姆剃刀原理为：“如非必要，勿增实体”，在模型评估中，如果两个假设有相同的经验误差，此时根据奥卡姆剃刀原理应该选择相对简单的假设。

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。

答：

对于基于迭代的最优化算法，不存在某种算法对于所有有限空间中的搜索问题均有效；如果一个算法对某些问题有效，那么它一定在另外一些问题上比纯粹随机搜索算法更差。

在机器学习领域，NFL 定理说明任意两种机器学习算法$\mathcal{L}_A$和$\mathcal{L}_B$，在所有任务上的平均性能是相同的，不存在某种机器学习算法适用于所有任务。

12 请简要阐述经验误差与泛化误差之间的关系。

答：

1) 当训练样本数量 N 趋近于无穷大时，泛化误差等于经验误差；
2) 经验误差的数学期望等于泛化误差；
3) 利用经验误差与模型参数可以描述泛化误差的边界。

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。

答：

P 表示 Probably，是指模型泛化误差小于某个给定阈值的概率，反映了模型评估结果的置信度；A 表示 Approximately，是指在不低于某个给定的概率下模型的泛化误差，反映了模型评估结果的精度。

14 请简述不可知 PAC 学习与可知 PAC 学习的最主要差别

答：

可知 PAC 学习与不可知 PAC 学习最大的差别在于真实映射 f 是否属于学习算法$\mathcal{L}$所定义的假设空间$\mathcal{H}$.

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。

答：

(1) 线性回归，在损失函数中加入正则化项$\|\boldsymbol{w}\|^2$,得到岭回归封闭解
(2) SVM，在线性 SVM 基础上加入松弛变量，构建软间隔 SVM，形成结构风险项；
(3) 决策树，加入剪枝操作，降低决策树的复杂度。

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：

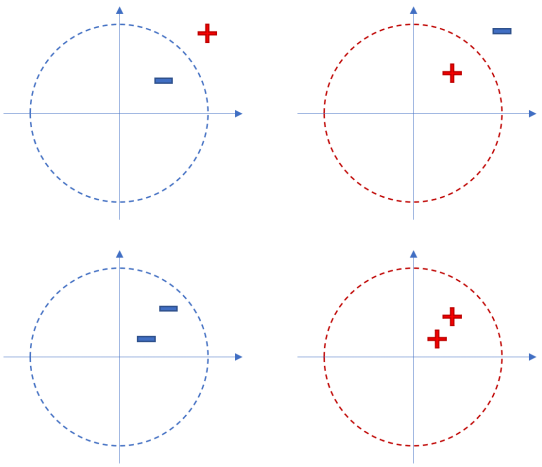
$$\mathcal{H} = \{h(x;a,b) = \text{sign}[a(x^T x - b)] | a \in \{+1,-1\}, b \in [0,+\infty)\}$$

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）

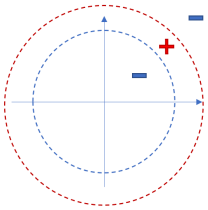
答：

映射函数$\text{sign}[a(x^T x - b)]$的判别效果相当于在利用一个以圆点为中心，半径为$\sqrt{b}$的圆将样本空间$\mathcal{X}$分为两个区域，并根据 a 的数值决定区域的类别标签。

当 d=2 时，有四种情况如下，（红圈表示圈内为正类，蓝圈表示圈内为负类）显然能被$\mathcal{H}$打散



当 d=3 时，有四种情况如下，以下样本分布无法被任意$h \in \mathcal{H}$正确分类，因此，d=3 时样本集无法被假设空间$\mathcal{H}$打散



结论：假设空间$\mathcal{H}$的 VC 维 d=2。

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。

答：根据样本$\boldsymbol{x}$的取值描述，样本空间大小为$|\mathcal{X}| = 3 * 2 * 2 = 12$。每一

个样本有两种可能的类别标签，因此假设空间大小$|\mathcal{H}_1| = 2^{12} = 4096$

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。

解：由于$\mathcal{H}_1$包含所有可能的二分类映射，因此属于有限假设条件下的一致情况。根据 PAC 理论，其误差上界满足：

$$E \leq \frac{1}{N}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

设置信度为$\delta = 1 - 0.8 = 0.2$，

假设空间大小$|\mathcal{H}| = |\mathcal{H}_1| = 4096$，

训练样本数量$N = 100$

代入上式有：

$$E \leq \frac{1}{100}\left(\ln(4096) + \ln\frac{1}{0.2}\right) \approx 0.099$$

因此，泛化误差上界不超过 0.1，可以满足 PAC 学习条件。

3) 如果存在一个算法$\mathcal{L}_2$，可以生成 512 种不同的假设，且在当前数据集上的误差为 0.05。试计算$\mathcal{L}_2$在该问题上的泛化误差上界。

解：算法$\mathcal{L}_2$对应的假设空间大小为$|\mathcal{H}_2| = 512$，且由于在当前训练集上的误差不为 0，因此属于有限假设条件下的不一致情况。根据 PAC 理论有，此时的泛化误差上界为：

$$= \hat{E}(h) + \sqrt{\frac{1}{2N}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)}$$

$$= 0.05 + \sqrt{\frac{1}{200}\left(\ln(512) + \ln\frac{1}{0.2}\right)}$$

$$\approx 0.248$$

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

解：首先统计 TP，FP，TN，FN 的样本数量，如下表：

	T	F
P	4	2
N	2	2

结论：$TP=4, FP=2, TN=2, FN=2$

真阳率：$TPR=\frac{TP}{TP+FP}=\frac{4}{4+2}\approx 0.67$

假阳率：$FPR=\frac{FP}{FP+TN}=\frac{2}{2+2}\approx 0.5$

查准率：Precision$=\frac{TP}{TP+FP}=\frac{4}{4+2}\approx 0.67$

查全率：Recall$=TPR\approx 0.67$

F1 度量：$F_1=\frac{2*Precision*Recall}{Precision+Recall}=\frac{2*\frac{2}{3}*\frac{2}{3}}{\frac{2}{3}+\frac{2}{3}}=\frac{2}{3}\approx 0.67$