

Studies ML

Jacob Xie

2023-03-11

2 模型评估与选择

2.1 误差与拟合

误差与拟合		
名称	英文	描述
错误率	error rate	如果在 m 个样本中有 a 个样本分类错误, 则错误率 $E = a/m$
精度	accuracy	$1 - a/m$
误差	error	学习器的实际预输出与样本的真实输出之间的差异
训练误差	training error	学习器在训练集上的误差
经验误差	empirical error	
泛化误差	generalization error	在新样本上的误差
过拟合	over fitting	学习器把训练样本自身的一些特点当做了所有潜在样本都会具有的一般性质, 导致泛化性能下降
欠拟合	under fitting	与过拟合相对应
模型选择	model selection	

2.2 评估方法

评估方法		
测试集	testing set	
测试误差	testing error	
留出法	hold-out	直接将数据集 D 划分为两个互斥的集合, 其中一个集合作为训练集 S , 另一个作为测试集 T , 即 $D = S \cup T, S \cap T = \emptyset$ 。在 S 上训练出模型后, 用 T 来评估其测试误差, 作为对泛化误差的估计。
采样	sampling	
分层采样	stratified sampling	保留类别比例的采样方式
保真性	fidelity	
交叉验证法	cross validation	将数据集 D 划分为 k 个大小相似的互斥子集, 即 $D = D_1 \cup D_2 \cup \dots \cup D_k, D_i \cap D_j = \emptyset (i \neq j)$ 。每个子集 D_i 都尽可能保持数据分布的一致性, 即从 D 中通过分层采样的到。然后每次用 $k - 1$ 个子集的并集作为训练集, 余下的那个子集作为测试集; 这样就可获得 k 组训练/测试集, 从而可进行 k 次训练和测试, 最终返回的是这 k 个测试结果的均值。
k 折交叉验证	k-fold cross validation	
留一法	leave-one-out	
自助法	bootstrapping	
自助采样法	bootstrap sampling	
包外估计	out-of-bag estimate	
参数	parameter	
调参	parameter tuning	
验证集	validation set	

2.3 性能度量

2.3.1 错误率与精度

性能度量 (performance measure): 衡量模型泛化能力的评价标准。

均方误差 (mean squared error):

$$E(f; D) = \frac{1}{m} \sum_{i=1}^m (f(\mathbf{x}_i) - y_i)^2 \quad (2.2)$$

对于数据分布 \mathcal{D} 和概率密度函数 $p(\cdot)$, 均方误差可描述为:

$$E(f; \mathcal{D}) = \int_{\mathbf{x} \in \mathcal{D}} (f(\mathbf{x}) - y)^2 p(\mathbf{x}) d\mathbf{x} \quad (2.3)$$

错误率是分类错误的样本数占样本总数的比例:

$$E(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(\mathbf{x}_i) \neq y_i) \quad (2.4)$$

精度则是分类正确的样本数占样本总数的比例:

$$\begin{aligned} acc(f; D) &= \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(\mathbf{x}_i) = y_i) \\ &= 1 - E(f; D) \end{aligned} \quad (2.5)$$

对于数据分布 \mathcal{D} 和概率密度函数 $p(\cdot)$, 错误率与精度可分别描述为

$$E(f; \mathcal{D}) = \int_{\mathbf{x} \in \mathcal{D}} \mathbb{I}(f(\mathbf{x}) \neq y) p(\mathbf{x}) d\mathbf{x} \quad (2.6)$$

$$\begin{aligned} acc(f; \mathcal{D}) &= \int_{\mathbf{x} \in \mathcal{D}} \mathbb{I}(f(\mathbf{x}) = y) p(\mathbf{x}) d\mathbf{x} \\ &= 1 - E(f; \mathcal{D}) \end{aligned} \quad (2.7)$$

2.5 偏差与方差

偏差-方差分解 (bias-variance decomposition): 对学习算法的期望泛化错误率进行拆解。

偏差-方差窘境 (bias-variance dilemma)

3 线性模型

4 决策树

5 神经网络

6 支持向量机

7 贝叶斯分类器

8 集成学习

9 聚类

10 降维与度量学习

11 特征选择与稀疏学习

12 计算学习理论

13 半监督学习

14 概率图模型

15 规则学习

16 强化学习