经验误差、过拟合

2020年9月15日

15:12

一、

1. ***错误率***：错误样本占比——通常，我们把分类错误的样本占样本总数的比例称为‘错误率’。
2. ***精度***：1 - 错误率。
3. ***误差***：样本真实输出与预测输出之间的差异
   1. ***训练误差/经验误差***：在**训练集**上的误差
   2. ***测试误差***：在**测试集**上的误差
   3. ***泛化误差***：在新样本上的误差。

二、***过拟合***

1. 概念：学习器把训练样本学习的‘太好’，训练样本本身特点当作所有样本的一般性质，导致泛化能力下降——死记硬背。
2. 原因：‘学习能力’太好
3. 改善策略：
   1. 优化目标加正则项
   2. early stop

三、***欠拟合***

1. 概念：对训练样本一般性质尚未学好
2. 原因：学习能力不足。
3. 解决：
   1. 决策树学习中，扩展分支
   2. 神经网络，增加训练轮数

四、模型选择：

1. 现实任务中，多种学习方法可供选择，一种方法对应一个模型
2. 一类模型，通过不同的参数调整，又会产生不同的模型
3. 选择哪一个模型，模型中用哪些参数，就是模型选择需要解决的问题——***模型选择问题***
4. 理想的解决方案：对候选模型的泛化误差进行评估。but，a.无法直接获得泛化误差；b.现有的‘训练集’因为‘过拟合’现象，不适合作为标准。
5. 评估方法：在【4】的各种情况下，现实中，对***模型进行评估和选择***。

（见下一节）