评估方法——评价模型的方案

2020年9月15日

15:21

pre：

①为解决现实问题需要一定的方法进行‘评估’。

②需要一个‘测试集’来测试学习器的‘泛化能力’

③在‘测试集’上得到的‘**测试误差**’作为‘**泛化误差**’的近似。

问题：

①现实问题：数据集需要分两个作用——①训练用；②测试评估用。

②对于有限的数据集，既要训练，又要测试，故，需要对D进行处理，一下介绍几种方法：

一、***留出法***——直接将数据集划分为‘数据集’‘测试集’

1. 
2. 分层采样法：就是，将S、T中，每种‘label’分类，而S、T每一类对应的数量比例都是相同的。比如，（0，1）分类中，S中正反比例为1：2、T中正反比例亦为1：2。（在采样上下功夫）
3. 在‘比例相同’的情况下，又根据集合中的正反例的顺序关系不同分成不同的训练集。（采样后，在顺序上下功夫）在产生n个测试集后，取n个测试集的测试误差平均数作为性能度量——***若干次随机划分，重复实验***。
4. 注意：***比例窘境***——训练集S太多，性能接近D，得到的误差结果不准确；测试集T太多，不接近D，有较大差异。（目前无解）——可能的解决方案：取2/3~4/5的数据用作训练集S。

二、***交叉验证***——①数据集划分为k个子集，每次用k-1个进行训练，用最后一个进行训练；②多次重复①实验，得到组合的值；③最后所有组合取平均值。

【注！！使用同一个数据集合，进行多次测试，每一个数据都有可能成为测试集】

1. “***留一法***”：极端情况，每一个集合中只有一个样本。
   1. 好处：绝大情况下，评估结果往往被认为比较准确。
   2. 坏处：在数据量大的情况下，计算开销是无法忍受的（100W的数据需要计算100W次）。
   3. ps ——也符合‘没有免费午餐’定理，留一法的估计结果也未必永远比其他评估方法准确。

三、自助法——从m个样本的D数据集中随机放回抽取一个，到D'（下次再取还能取到相同的）。取m次后，大概有0.368的数据没被取到。 → 使用取出来的D'作为训练集，剩下没取到的36.8%的数据作为测试集。