McNemar检验

一、前言

对于只能执行一次的算法，McNemar检验是唯一具有可接受的第一类错误的检验，在比较两个二元分类算法时，检验是这两个模型是否存在相同的分歧的说明。它不会说明一个模型是否比另一个模型更准确或更容易出错。

二、列联表

二分类问题A、B算法分类差别		算法A	
		正确	错误
B算法	正确	e00	e01
	错误	e10	e11

三、检验实例的计数

检验统计量的这种计算假定计算中使用的列联表中的每个单元具有至少25个计数。检验统计量具有1自由度的卡方分布

$\chi^2 = \frac{(e01-e10)^2}{e01+e10}$

如果通过卡方分布进行评估，则通过将上表的e01与e10两个频率中较小的一个加上0.5、较大的一个减去0.5来进行连续性校正。这种纠正在统一降低差异的绝对值e01-e10中具有明显的效果。那么通过连续性校正，我们可以重写公式：

$\chi^2 = \frac{(|e01-e10|-1)^2}{e01+e10}$

给定显着性水平的选择，通过检验计算的p值可以解释如下：

- p>alpha：未能拒绝H0，在分歧上没有差异。
- p<=alpha：拒绝H0，分歧的显着差异。

四、使用McNemar检验解释分类器

H0:两学习器性能相同（分类器在测试集上具有相似的错误比例）
那么，我们可以总结如下：

- 不拒绝零假设：分类器在测试集上具有相似的错误比例 。
- 拒绝零假设：分类器在测试集上具有不同的错误比例 。

给定显著度 α，当以上变量恒小于临界值功时，不能拒绝假设，即认为两学习器的性能没有显著差别;否则拒绝假设，即认为两者性能有显著差别，且平均错误率较小的那个学习 器性能较优.自由度为1的χ2检验的临界值当α=0.05时为3.8415，α=0.1时为2.7055.

五 、两个重要限制

1.没有训练集或模型可变性的度量

通常，模型行为基于用来拟合模型的特定训练数据而变化 。
这是由于模型与特定训练实例的交互作用以及在学习过程中使用的随机性 。将模型拟合到多个不同的训练数据集并评估技能（如重采样方法所做的那样），提供了一种度量模型差异的方法 。

结论：如果可变性的来源很小，则该检验是合适的

2.较少模型直接比较

两个分类器在一个测试集上进行评估，并且测试集应该小于训练集 。这里更多的是使用重采样方法的假设检验不同，因为在评估期间，数据集可用作测试集 。它要求测试集有适当地代表性，这通常意味着测试数据集很大