交叉验证(Cross-validation)

作者：writer 来源：爱科学

本文原地址：https://www.iikx.com/news/statistics/2479.html

交叉验证
是一种用来评价一个统计分析的结果是否可以推广到一个独立的数据集上的技术。主要用于预测，即，想要估计一个预测模型的实际应用中的准确度。它是一种统计学上将数据样本切割成较小子集的实用方法。于是可以先在一个子集上做分析，
而其它子集则用来做后续对此分析的确认及验证。

交叉验证的理论是由Seymour Geisser所开始的。 它对于防范testing hypotheses suggested by the data是非常重要的， 特别是当后续的样本是危险、成本过高或不可能(uncomfortable science)去搜集。

一个交叉验证将样本数据集分成两个互补的子集，一个子集用于训练(分类器或模型)称为训练集(training set);另一个子集用于验证(分类器或模型的)分析的有效性称为测试集(testing set)。利用测试集来测试训练得到的分类器或模型，以此作为分类器或模型的性能指标。得到高度预测精确度和低的预测误差，是研究的期望。为了减少交叉验证结果的可变性，对一个样本数据集进行多次不同的划分，得到不同的互补子集，进行多次交叉验证。取多次验证的平均值作为验证结果。

在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，称为PRESS(predicted Error Sum of Squares)。

目的

用交叉验证的目的是为了**得到可靠稳定的模型**。在建立PCR 或PLS
模型时，一个很重要的因素是取多少个主成分的问题?用cross validation
校验每个主成分下的PRESS值，选择PRESS值小的主成分数。或PRESS值不在变小时的主成分数

交叉验证的目的
：假设分类器或模型有一个或多个未知的参数，并且设这个训练器(模型)与已有样本数据集(训练数据集)匹配。训练的过程是指优化模型的参数，以使得分类器或模型能够尽可能的与训练数据集匹配。我们在同一数据集总体中，取一个独立的测试数据集。

常见类型的交叉验证：

1、**重复随机子抽样验证**
。将数据集随机的划分为训练集和测试集。对每一个划分，用训练集训练分类器或模型，用测试集评估预测的精确度。进行多次划分，用均值来表示效能。

优点：与k倍交叉验证相比，这种方法的与k无关。

缺点：有些数据可能从未做过训练或测试数据;而有些数据不止一次选为训练或测试数据。

2、**K倍交叉验证**
(K>=2)。将样本数据集随机划分为K个子集(一般是均分)，将一个子集数据作为测试集，其余的K-1组子集作为训练集;将K个子集轮流作为测试集，重复上述过程，这样得到了K个分类器或模型，并利用测试集得到了K个分类器或模型的分类准确率。用K个分类准确率的平均值作为分类器或模型的性能指标。10-倍交叉证实是比较常用的。

优点：每一个样本数据都即被用作训练数据，也被用作测试数据。避免的过度学习和欠学习状态的发生，得到的结果比较具有说服力。

3、**留一法交叉验证**
。假设样本数据集中有N个样本数据。将每个样本单独作为测试集，其余N-1个样本作为训练集，这样得到了N个分类器或模型，用这N个分类器或模型的分类准确率的平均数作为此分类器的性能指标。

优点：每一个分类器或模型都是用几乎所有的样本来训练模型，最接近样本，这样评估所得的结果比较可靠。实验没有随机因素，整个过程是可重复的。

缺点：计算成本高，当N非常大时，计算耗时。

训练集和测试集的选取：

1、训练集中样本数量要足够多，一般至少大于总样本数的50%。

2、训练集和测试集必须从完整的数据集中均匀取样。均匀取样的目的是希望减少训练集、测试集与原数据集之间的偏差。当样本数量足够多时，通过随机取样，便可以实现均匀取样的效果。(随机取样，可重复性差)