如何使用交叉验证(Cross Validation)?

作者：Maggie张张 来源：CSDN

本文原地址：https://www.iikx.com/news/statistics/2480.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

如何使用交叉验证(Cross Validation)?在机器学习的相关研究中，如果是有监督的算法，需要将原始数据集分为训练集和测试集两个集合。训练集中的数据带有标签，用这些数据来训练出一个模型，告诉机器什么样的数据可以分成哪一类，然后用这个模型来预测测试集中数据的标签。然后用预测得到的标签跟真实的标签作比对，就可以得到这个模型的预测准确率，其实是考察这个模型的generalization ability(泛化能力)，即，从训练集中总结出来的规律能不能用到别的数据上去。

那么，怎样分训练集和测试集呢?需要考虑两个问题：

1. 训练集中的数据要足够多，一般要大于原始数据集的一般，否则总结出来的规律太小众

2. 两组集合必须是原始集合的均匀取样，否则比如说，训练集选择都是1类数据，测试集都是2类数据，训练之后模型知道1类数据的特点，用它来分别2类数据，这好难。。。

于是，cross validation的目的就是：科学地统计训练模型的泛化能力。

cross validation可以分成三种： double CV， k-fold CV, leave-one-out.

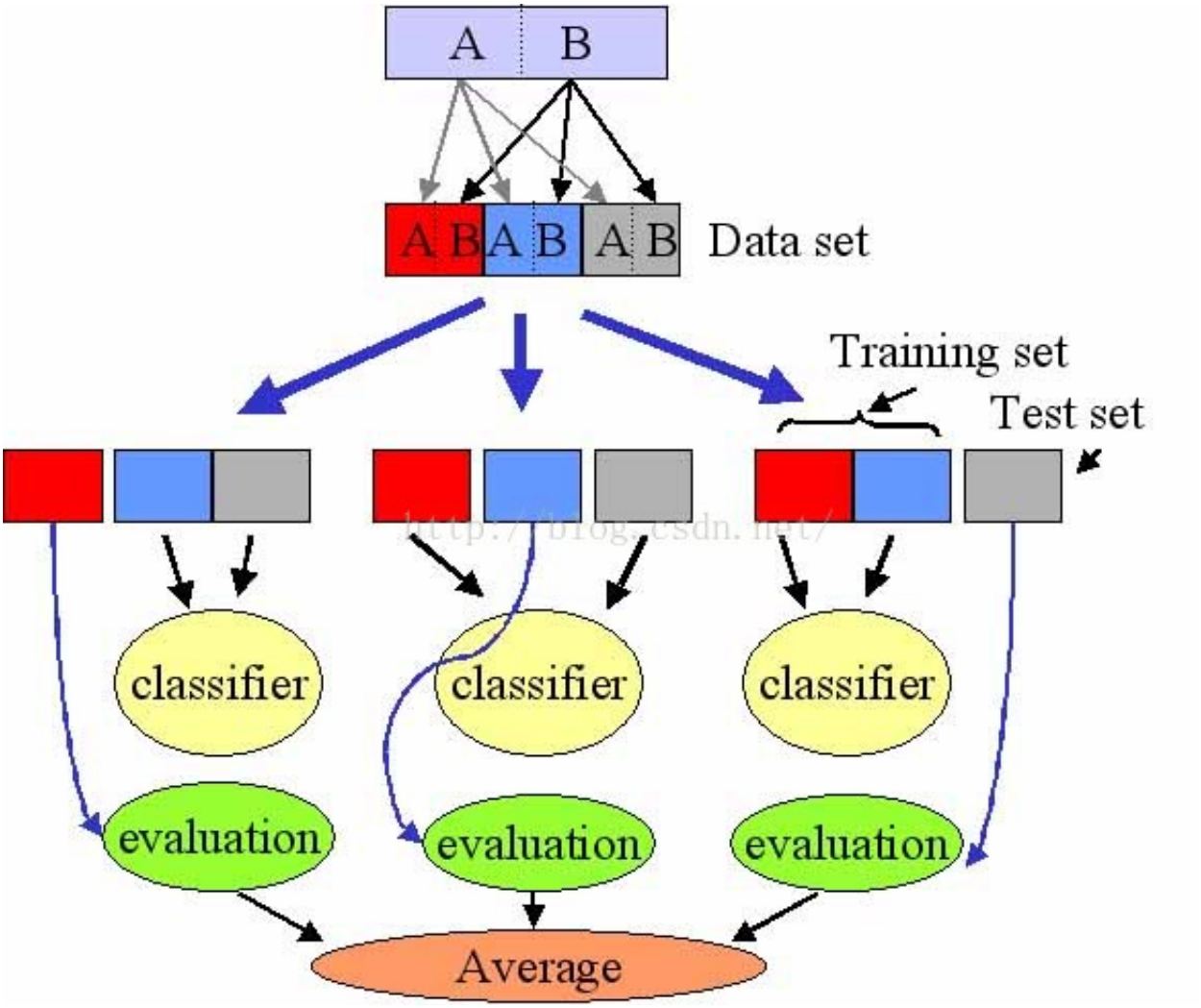
(1) double CV

原理：将每一类别下的数据平均分成两份，从每一类中各取一份组合起来做training，剩下的做testing。然后再把training和testing交换。这样就可以训练两次，得到两个模型，也得到两个accuracy，把这两个值平均一下，就是整个模型的分类准确率。

(2) k-fold CV

原理：其实是double
CV的一般化。把每一类别下的数据平均分成k份。从每一类中各取一份组合起来testing, 剩下的做training。然后再换一个每一类别下的子集，一共有k种组合方式，就可以得到k个模型，也就可以得到k个accuracy， 把这k个值做平均，就是整个模型的分类准确率。

下图是3-fold CV的示意图



一般来说，10折就已经足够多了，这样就可以保证训练集所用的数据是足够多的。

(3)leave-one-out

原理：每一个样本单独做一次test set，剩下的都拿来做train set，这样的话原始数据集有N个样本，就可以训练出N和模型，得到N和准确率，再做平均。

优点比较明显：

1. 每一回合几乎所有的样本都用来作训练，训练样本呢非常接近与原始集合，估测得到的generalization ability比较可靠

2.
实验过程中没有随机因素会影响实验数据(前两种方法中把样本平均切割成k份，是有随机性的)

同时缺点也是显而易见的：计算代价高，如果N比较大，那得算所长时间(如果算一次都比较久的话。。)

更多 统计方法 请访问 https://www.iikx.com/news/statistics/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发