在交叉验证中，留一法和普通交叉验证的区别？

作者：JH_Zhai 来源：CSDN

本文原地址：https://www.iikx.com/news/statistics/2478.html

在交叉验证中，留一法和普通交叉验证的区别？使用评估方法的动机：

通过实验测试来对学习器的泛化误差进行评估并进而做出选择。

评估方法 主要分三种：

留出法(分一次 互斥集)

交叉验证法(分多次 对k折形成多次互斥集)

自助法(有放回抽样)

留出法 只一次，随机性太大，说服力不强

而交叉验证 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。

那么交叉验证中 普通 和 留一 有什么区别呢?

下面以留一法为主体介绍优缺点：

优点：

1.我们用几乎所有的数据进行训练，然后用一个数据进行测试；2.确定性

确定性：

实验没有随机因素，整个过程是可重复的。

比如十折验证，你测两次，结果是不一样的

而你用留一法测多少次都是一样的

缺点：

1.计算时间很长；2.分层问题

分层问题：

让我们回到运动员分类的例子——判断女运动员参与的项目是篮球、体操、还是田径。

在训练分类器的时候，我们会试图让训练集包含全部三种类别。如果我们完全随机分配，训练集中有可能会不包含篮球运动员，在测试的时候就会影响结果。

比如说，我们来构建一个包含100个运动员的数据集：从女子NBA网站上获取33名篮球运动员的信息，到Wikipedia上获取33个参加过2012奥运会体操项目的运动员，以及34名田径运动员的信息。

现在我们来做十折交叉验证。我们按顺序将这些运动员放到10个桶中，所以前三个桶放的都是篮球运动员，第四个桶有篮球运动员也有体操运动员，以此类推。

这样一来，没有一个桶能真正代表这个数据集的全貌。最好的方法是将不同类别的运动员按比例分发到各个桶中，这样每个桶都会包含三分之一篮球运动员、三分之一体操运动员、以及三分之一田径运动员。

这种做法叫做分
层。而在留一法中，所有的测试
集都只包含一个数据。所以说，**留一法对小数据集是合适的**
，但大多数情况下我们会选择十折交叉验证。