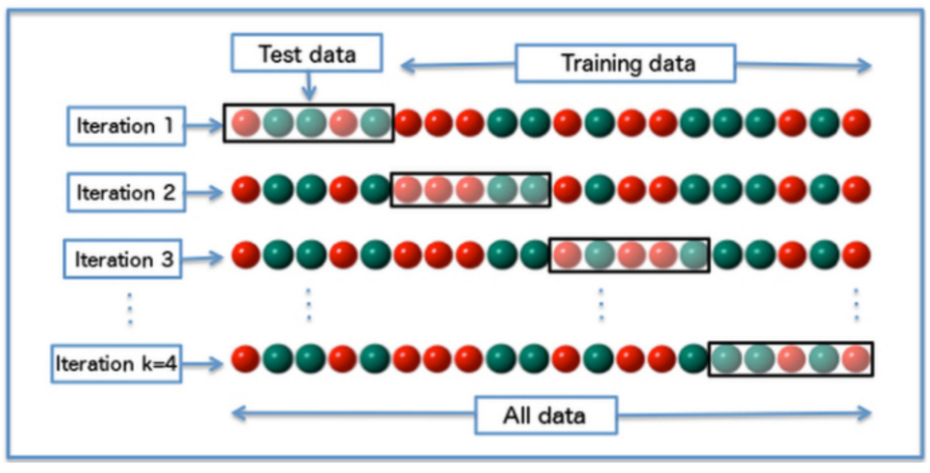

机器学习中testing和hold-out的区别【为什么要分出一个hold-out】

云中帆
数据科学工程博士

【声明：有的英语我实在脑子卡克不知道怎么翻译，又懒得去找正确翻译，所以，就将就看吧。。。】

在机器学习中，训练模型的时候大家理所当然会想说要把已有数据分成training set（训练集）和test set（测试集）。用训练集去训练，然后用测试集去测试模型。但是有时候会看到有人不是像这样分成两份的，而是三份：

- training set
- testing set
- hold-out set

这个holdout set 是什么鬼。其实他就是用来做最终验证的。可是，明明有testing set了为啥还要分出一个hold-out set。且待我娓娓道来原因：

假设你的数据集中有100个例子（每个例子=数据+相应的label）：

情况1：用所有的100个例子作为训练集训练模型，再用训练好的模型来测试这个100个例子。调整参数使得最终的预测准确率最高。

问题：Overfitting(过拟合)。因为这个模型不能准确预测除了训练数据以外的其他数据。

情况2：用75%的例子，也就是100个中的75个例子作为训练集来训练模型， 剩下25个用来调整模型参数使得最终的测试正确率最高。这样就避免了过拟合。

问题：这75%的例子你是咋选出来的？我们能不能更有效率地利用好手头上已有的100个例子呢？你只用75%的例子来训练模型，那剩下的25%岂不是浪费了。

情况3：我们在情况2的基础上改进一下。我们把100个例子等分成4份，每份25个，然后做4次模型训练。每次都用到不同的25% 的测试集（剩下的75%则作为训练集）。这样，在训练过程中，所有数据都被用上了。【该方法也称作k-fold cross validation,本例中，k=4】。最终，在四次训练的基础上，优化参数，取平均值。

问题： 有人会抬杠来跟你说：不管怎样，你丫的结果还不是从同样例子里训练出来的。你凭啥说它能很好地工作在全新的例子上呢？

情况4：这次你从100个例子里硬生生抽出20个例子把他们晾在一边，完完全全不参与模型训练。这20个例子我们称之为hold-out set。然后剩下80个例子，等分成4份，按照情况3的办法来训练一个模型，等模型训练好后，用这个hold-out set来证明给别人看：我的模型是可以很好地工作在没有参与模型训练的例子上滴！

最后，没人可以反对你在测试集的基础上优化参数这件事，而你自己也很确定你的模型的确可以工作的很好。

编辑于 2018-06-03

机器学习 统计学习 深度学习（Deep Learning）

文章被以下专栏收录

从零学AI
零基础学AI
进入专栏

推荐阅读

元学习[1]：深度阐述元学习的理论模型
ALme 发表于贝叶斯与元...

机器学习-模型选择与评价
HyeRi

如何搭建一个简单的机器学习流水线?
阅读此分步教程，学会通过导入scikit-learn包来搭建一个简单的机器学习流水线(pipeline) 一个机器学习模型中，有很多可移动的组件需要被组合在一起，模型才能被执行并成功的得到结果。把机...
集智学园

机器学习为什么要划分训练集、测试集和验证集？这3个样本...
圆派314 发表于机器学习+...

2 条评论 切换为时间排序

写下你的评论...

杨天琪 2018-06-05
涨知识了，多谢
赞

知乎用户 2019-01-21
秒懂!
赞