

第12讲 建立算法的管道模型

主讲教师：欧新宇

February 21, 2020

Outlines

- 管道模型的基本概念
- 使用管道模型的原因
- 管道模型案例——股票涨跌分析

1. 管道模型的概念

管道模型 (Pipeline)

把一系列算法打包在一起，让他们各司其职，形成一个流水线。就像组装汽车的工厂流水线。

管道机制在机器学习算法中得以应用的根源在于，**参数集**在训练过程中的重复使用。例如多个预处理算法需要测试，某个算法的参数有多个值需要测试。管道机制实现了对全部步骤的流式化封装和管理 (streaming workflows with pipelines) 。

值得注意的是：

管道机制更像是**编程技巧**的创新，而非**算法**的创新。

2. 使用管道模型的原因

一段标准的训练过程

1. 数据载入并划分为训练集和测试集
2. 数据预处理，使用训练集拟合预处理器scaler，并用其来预处理训练集和测试集
3. 使用基于MLP的网格搜索算法及交叉验证获取最优参数并输出交叉验证评分
4. 利用交叉验证获得的参数在测试集上进行预测

思考以上训练过程是否存在问题？

2. 使用管道模型的原因

1. 首先，我们在使用**交叉验证**的时候，会将输入到GridSearchCV中的训练集X_train_scaled进行拆分，划分为训练集Grid_train和验证集Grid_val。

注意：在训练模型的过程中，我们一直有一个原则：测试集永远不能参与训练中，只能用于最终的评估。

2. 其次，在进行最初的数据预处理的时候，我们使用整个训练集X_train_scaled去拟合预处理器**scaler**，这就意味着Grid_val被用来参与训练**scaler**。

3. 然后，我们又用基于Grid_val训练的预处理器**scaler**来拟合在交叉验证中作为验证集（或者称为测试集）的Grid_val，这就违背了之前的原则。

因此，前面的训练过程会导致交叉验证的结果出现偏差。

2. 使用管道模型的原因

How to fix it?

2. 使用管道模型的原因

一个容易理解的方式是：

1. 针对每一组**参数对**，我们都执行一次数据拆分，并在拆分数据的时候，直接将数据拆分成**三部分**，训练集`new_train`、验证集`new_val`、测试集`new_test`；
2. 然后用**最新**生成的训练集`new_train`来训练**预处理器**`scaler`，并用`scaler`来拟合证集`new_val`、测试集`new_test`；
3. 接下来再使用**被预处理过的**`new_train`和`new_val`来进行交叉验证和网格搜索。
4. 之后再将训练集`new_train`、验证集`new_val`合并成 `new_trainval`，并保持网格搜索的最优参数不变，训练新的模型；
5. 最后在用**新的模型**在新测试集`new_test`上输出最终评分。

是否还存在问题？

2. 使用管道模型的原因

看起来非常流畅，但是存在一个“问题”。

在第三步中，由于我们使用的是 **K 折交叉验证**，这就意味着在**每次**训练的过程中训练集`new_train`和验证集`new_val`**都是不同的**（实际上，不管是哪种交叉验证方法，都会有这样的问题，例如留一法），这样就会产生`new_train1,new_train2,...;new_val1,new_val2,...`。

很显然这样的操作没有“问题”，但是代码会变得**非常繁琐**。

2. 使用管道模型的原因

复杂的代码 还是 不可行的交叉验证评分？

管道模型Pipeline

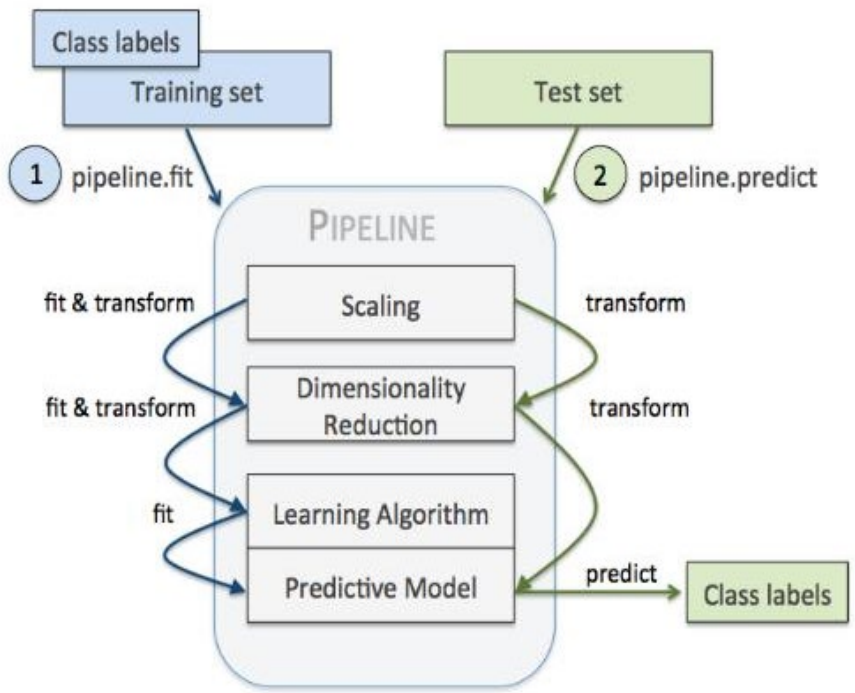
管道模型会在每次进行交叉验证的时候，都重新对trainval进行拆分，并分别对训练集pipe_train进行预处理，然后再用生成的scale对验证集pipe_val进行拟合。

使用管道模型，我们可以非常简单的代码实现刚刚的需求。

2. 使用管道模型的原因

值得注意的是：

管道模型的执行过程和我们刚刚分析的过程是一致的，只是它被更好地封装成了一个**类**，同时支持并行处理等优化算法，让程序员只需要调用Pipeline接口就可以实现复杂的循环操作。




Pipeline的基本流程：

1. scaler预处理模型
2. 降维模型
3. 分类回归模型（包含交叉验证和超参数搜索）

3. 管道模型案例——股票涨跌分析

代码基本设计思路

- 数据载入（观察数据、数据清洗、数据拆分）
 - 数据预处理
 - 使用交叉验证输出验证集评分
 - 使用管道模型进行交叉验证模型训练
 - 使用管道模型进行模型选择
 - 使用管道模型进行参数调优
 - 可视化输出
- 

欧老师的联系方式

读万卷书 行万里路 只为最好的修炼

QQ: 14777591 (宇宙骑士)

Email: ouxinyu@alumni.hust.edu.cn

Tel: 18687840023