Python机器学习基础教程

Stephen CUI

February 17, 2023

Chapter 1

模型评估与改进

为了评估我们的监督模型，我们使用 `train_test_split` 函数将数据集划分为训练集和测试集，在训练集上调用 fit 方法来构建模型，并且在测试集上用 score 方法来评估这个模型——对于分类问题而言，就是计算正确分类的样本所占的比例。

请记住，之所以将数据划分为训练集和测试集，是因为我们想要度量模型对前所未见的新数据的泛化性能。我们对模型在训练集上的拟合效果不感兴趣，而是想知道模型对于训练过程中没有见过的数据的预测能力。

本章我们将从两个方面进行模型评估。我们首先介绍交叉验证，然后讨论评估分类和回归性能的方法，其中前者是一种更可靠的评估泛化性能的方法，后者是在默认度量（score方法给出的精度和 R^2）之外的方法。

我们还将讨论网格搜索，这是一种调节监督模型参数以获得最佳泛化性能的有效方法。

1.1 交叉验证

交叉验证（cross-validation）是一种评估泛化性能的统计学方法，它比单次划分训练集和测试集的方法更加稳定、全面。在交叉验证中，数据被多次划分，并且需要训练多个模型。最常用的交叉验证是 k 折交叉验证（k-fold cross-validation），其中 k 是由用户指定的数字，通常取 5 或 10。在执行 5 折交叉验证时，首先将数据划分为（大致）相等的 5 部分，每一部分叫作折（fold）。接下来训练一系列模型。使用第 1 折作为测试集、其他折（2 5）作为训练集来训练第一个模型。利用 2 5 折中的数据来构建模型，然后在 1 折上评估精度。之后构建另一个模型，这次使用 2 折作为测试集，1、3、4、5 折中的数据作为训练集。利用 3、4、5 折作为测试集继续重复这一过程。对于将数据划分为训练集和测试集的这 5 次划分，每一次都要计算精度。最后我们得到了 5 个精度值。

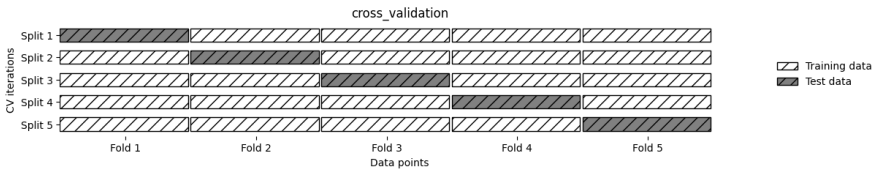

Figure 1.1: Data splitting in five-fold cross-validation

1.1.1 scikit-learn中的交叉验证

scikit-learn 是利用 `model_selection` 模块中的 `cross_val_score` 函数来实现交叉验证的。`cross_val_score` 函数的参数是我们想要评估的模型、训练数据与真实标签。默认情况下，`cross_val_score` 执行 5 折交叉验证，返回 5 个精度值。可以通过修改 cv 参数来改变折数。

总结交叉验证精度的一种常用方法是计算平均值

1.1.2 交叉验证的优点

1.1.3 分层k折交叉验证和其他策略

对交叉验证的更多控制

留一法交叉验证

打乱划分交叉验证

分组交叉验证

1.2 网格搜索

1.2.1 简单网格搜索

1.2.2 参数过拟合的风险与验证集

1.2.3 带交叉验证的网格搜索

分析交叉验证的结果

在非网格的空间中搜索

使用不同的交叉验证策略进行网格搜索

1.3 评估指标与评分

1.3.1 牢记最终目标

1.3.2 二分类指标

错误类型

不平衡数据集

混淆矩阵

考虑不确定性

准确率-召回率曲线

受试者工作特征（ROC）与AUC

1.3.3 多分类指标

1.3.4 回归指标

1.3.5 在模型选择中使用评估指标

1.4 小结