机器学习实战

Stephen CUI[1]

February 17, 2023

[1]cuixuanStephen@gmail.com

Contents

Part I

分类

前两部分主要探讨监督学习（supervised learning）。在监督学习的过程中，我们只需要给定输入样本集，机器就可以从中推演出指定目标变量的可能结果。

监督学习一般使用两种类型的目标变量：标称型和数值型。标称型目标变量的结果只在有限目标集中取值，如真与假、动物分类集合 { 爬行类、鱼类、哺乳类、两栖类 } ；数值型目标变量则可以从无限的数值集合中取值，如 0.100、42.001、1000.743 等。数值型目标变量主要用于回归分析。

Chapter 1

决策树

你是否玩过二十个问题的游戏，游戏的规则很简单：参与游戏的一方在脑海里想某个事物，其他参与者向他提问题，只允许提20个问题，问题的答案也只能用对或错回答。问问题的人通过推断分解，逐步缩小待猜测事物的范围。决策树的工作原理与20个问题类似，用户输入一系列数据，然后给出游戏的答案。我们经常使用决策树处理分类问题，近来的调查表明决策树也是最经常使用的数据挖掘算法[1]。

它之所以如此流行，一个很重要的原因就是不需要了解机器学习的知识，就能搞明白决策树是如何工作的。

k-近邻算法可以完成很多分类任务，但是它最大的缺点就是无法给出数据的内在含义，决策树的主要优势就在于数据形式非常容易理解。

决策树的一个重要任务是为了数据中所蕴含的知识信息，因此决策树可以使用不熟悉的数据集合，并从中提取出一系列规则，在这些机器根据数据集创建规则时，就是机器学习的过程。

1.1 决策树的构造

决策树

优点：计算复杂度不高，输出结果易于理解，对中间值的缺失不敏感，可以处理不相关特征数据。
缺点：可能会产生过度匹配问题。
适用数据类型：数值型和标称型。

首先我们讨论数学上如何使用**信息论**划分数据集，然后编写代码将理论应用到具体的数据集上，最后编写代码构建决策树。

在构造决策树时，我们需要解决的第一个问题就是，当前数据集上哪个特征在划分数据分类时起决定性作用。为了找到决定性的特征，划分出最好的结果，我们必须评估每个特征。完成测试之后，原始数据集就被划分为几个数据子集。这些数据子集会分布在第一个决策点的所有分支上。如果某个分支下的数据属于同一类型，则无需进一步对数据集进行分割。如果数据子集内的数据不属于同一类型，则需要重复划分数据子集的过程。如何划分数据子集的算法和划分原始数据集的方法相同，直到所有具有相同类型的数据均在一个数据子集内。

[1] Giovanni Seni and John Elder, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions, Synthesis Lectures on Data Mining and Knowledge Discovery (Morgan and Claypool, 2010), 28.

创建分支的伪代码函数createBranch()如下所示：

Algorithm 1: 决策树分支判断

```
if 数据集中的每个子项属于同一分类 then
    return 类标签
else
    寻找划分数据集的最好特征;
    划分数据集;
    创建分支节点;
    for 每个划分的子集 do
        调用函数createBranch并增加返回结果到分支节点中
    end
    return 分支节点(返回子树或者说决策树更好一点)
end
```