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

机器学习基础

机器学习能让我们自数据集中受到启发，换句话说，我们会利用计算机来彰显数据背后的真实含义，这才是机器学习的真实含义。它既不是只会徒然模仿的机器人，也不是具有人类感情的仿生人。

1.1 何谓机器学习

机器学习就是把无序的数据转换成有用的信息。（看似无序的）

为什么需要统计学？现实世界中存在着很多例子，我们无法为之建立精确的数学模型，而为了解决这类问题，我们就需要统计学工具。（世界是概率的）

1.1.1 机器学习非常重要

> “我不断地告诉大家，未来十年最热门的职业是统计学家。很多人认为我是开玩笑，谁又能想到计算机工程师会是20世纪90年代最诱人的职业呢？如何解释数据、处理数据、从中抽取价值、展示和交流数据结果，在未来十年将是最重要的职业技能，甚至是大学，中学，小学的学生也必需具备的技能，因为我们每时每刻都在接触大量的免费信息，如何理解数据、从中抽取有价值的信息才是其中的关键。这里统计学家只是其中的一个关键环节，我们还需要合理的展示数据、交流和利用数据。我确实认为，能够从数据分析中领悟到有价值信息是非常重要的。职业经理人尤其需要能够合理使用和理解自己部门产生的数据。”
>
> ——McKinsey Quarterly，2009年1月

1.2 关键术语

在开始研究机器学习算法之前，必须掌握一些基本的术语。通过构建下面的鸟类分类系统，我们将接触机器学习涉及的常用术语。这类系统非常有趣，通常与机器学习中的专家系统有关。

Table 1.1是我们用于区分不同鸟类需要使用的四个不同的属性值，我们选用体重、翼展、有无脚蹼以及后背颜色作为评测基准。下面测量的这四种值称之为特征，也可以称作属性，但这里一律将其称为特征。Table 1.1中的每一行都是一个具有相关特征的实例。

Table 1.1的前两种特征是数值型，可以使用十进制数字；第三种特征（是否有脚蹼）是二值型，只可以取0或1；第四种特征（后背颜色）是基于自定义调色板的枚举类型，这里仅选择一些常用色彩。

假定我们可以得到所需的全部特征信息，那该如何判断鸟是不是象牙喙啄木鸟呢？这个任务就是分类，有很多机器学习算法非常善于分类。

Table 1.1: 基于四种特征的鸟物种分类表

	体重（克）	翼展（厘米）	脚蹼	后背颜色	种属
1	1000.1	125.0	无	棕色	红尾鵟
2	3000.7	200.0	无	灰色	鹭鹰
3	3300.0	220.3	无	灰色	鹭鹰
4	4100.0	136.0	有	黑色	普通潜鸟
5	3.0	11.0	无	绿色	瑰丽蜂鸟
6	570.0	75.0	无	黑色	象牙喙啄木鸟

最终我们决定使用某个机器学习算法进行分类，首先需要做的是算法训练，即学习如何分类。通常我们为算法输入大量已分类数据作为算法的训练集。训练集是用于训练机器学习算法的数据样本集合，Table 1.1是包含六个训练样本的训练集，每个训练样本有4种特征、一个目标变量。目标变量是机器学习算法的预测结果，在分类算法中目标变量的类型通常是标称型的，而在回归算法中通常是连续型的。我们通常将分类问题中的目标变量称为类别，并假定分类问题只存在有限个数的类别。

> 特征或者属性通常是训练样本集的列，它们是独立测量得到的结果，多个特征联系在一起共同组成一个训练样本。

为了测试机器学习算法的效果，通常使用两套独立的样本集：训练数据和测试数据。当机器学习程序开始运行时，使用训练样本集作为算法的输入，训练完成之后输入测试样本。输入测试样本时并不提供测试样本的目标变量，由程序决定样本属于哪个类别。比较测试样本预测的目标变量值与实际样本类别之间的差别，就可以得出算法的实际精确度。

假定这个鸟类分类程序，经过测试满足精确度要求，是否我们就可以看到机器已经学会了如何区分不同的鸟类了呢？这部分工作称之为知识表示，某些算法可以产生很容易理解的知识表示，而某些算法的知识表示也许只能为计算机所理解。知识表示可以采用规则集的形式，也可以采用概率分布的形式，甚至可以是训练样本集中的一个实例。在某些场合中，人们可能并不想建立一个专家系统，而仅仅对机器学习算法获取的信息感兴趣。此时，采用何种方式表示知识就显得非常重要了。

1.3 机器学习的主要任务

已经介绍了机器学习如何解决分类问题，它的主要任务是将实例数据划分到合适的分类中。机器学习的另一项任务是回归，它主要用于预测数值型数据。分类和回归属于监督学习，之所以称之为监督学习，是因为这类算法必须知道预测什么，即目标变量的信息。

与监督学习相对应的是无监督学习，此时数据没有类别信息，也不会给定目标值。在无监督学习中，将数据集合分成由类似的对象组成的多个类的过程被称为聚类；将寻找描述数据统计值的过程称之为密度估计。此外，无监督学习还可以减少数据特征的维度，以便我们可以使用二维或三维图形更加直观地展示数据信息。

- 监督学习：k-近邻算法、线性回归、朴素贝叶斯算法、局部加权线性回归、支持向量机、Ridge 回归、决策树、Lasso 最小回归系数估计
- 无监督学习：K-均值、最大期望算法、DBSCAN、Parzen窗设计