

Python机器学习基础教程

Stephen CUI¹

February 17, 2023

¹cuixuanStephen@gmail.com

Chapter 1

监督学习

1.1 分类与回归

监督机器学习问题主要有两种，分别叫作分类（classification）与回归（regression）。

分类问题的目标是预测类别标签（class label），这些标签来自预定义的可选列表。分类问题有时可分为二分类（binary classification，在两个类别之间进行区分的一种特殊情况）和多分类（multiclass classification，在两个以上的类别之间进行区分）。在二分类问题中，我们通常将其中一个类别称为正类（positive class），另一个类别称为反类（negative class）。这里的“正”并不代表好的方面或正数，而是代表研究对象。

回归任务的目标是预测一个连续值，编程术语叫作浮点数（floating-point number），数学术语叫作实数（real number）。

区分分类任务和回归任务有一个简单方法，就是问一个问题：输出是否具有某种连续性。如果在可能的结果之间具有连续性，那么它就是一个回归问题。

1.2 泛化、过拟合与欠拟合

在监督学习中，我们想要在训练数据上构建模型，然后能够对没见过的新数据（这些新数据与训练集具有相同的特性）做出准确预测。如果一个模型能够对没见过数据做出准确预测，我们就说它能够从训练集泛化（generalize）到测试集。我们想要构建一个泛化精度尽可能高的模型。

判断一个算法在新数据上表现好坏的唯一度量，就是在测试集上的评估。然而从直觉上看¹，我们认为简单的模型对新数据的泛化能力更好。因此，我们总想找到最简单的模型。构建一个对现有信息量来说过于复杂的模型，这被称为过拟合（overfitting）。如果你在拟合模型时过分关注训练集的细节，得到了一个在训练集上表现很好、但不能泛化到新数据上的模型，那么就存在过拟合。与之相反，如果你的模型过于简单，那么你可能无法抓住数据的全部内容以及数据中的变化，你的模型甚至在训练集上的表现就很差。选择过于简单的模型被称为欠拟合（underfitting）。

模型复杂度与数据集大小的关系

需要注意，模型复杂度与训练数据集中输入的变化密切相关：数据集中包含的数据点的变化范围越大，在不发生过拟合的前提下你可以使用的模型就越复杂。通常来说，收集更多的数据点可以有更大的

¹在数学上也可以证明这一点。

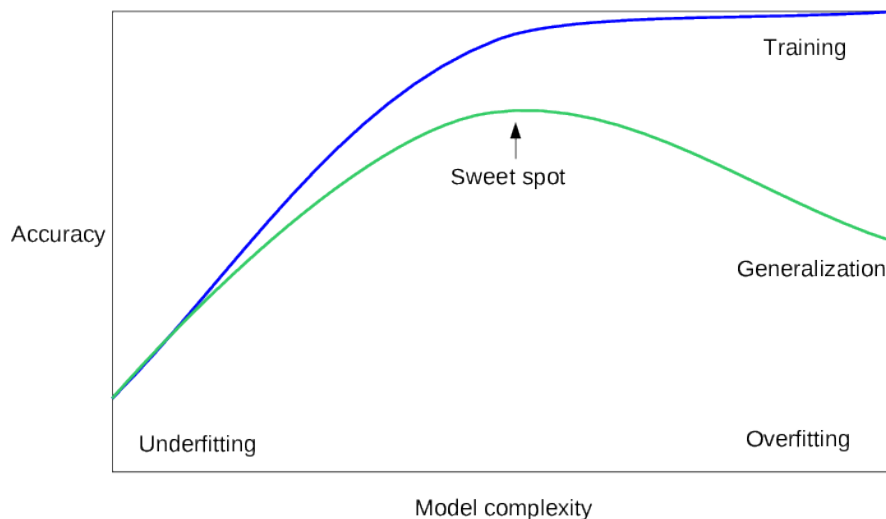


Figure 1.1: Trade-off of model complexity against training and test accuracy

Table 1.1: 一些样本数据集

| 名称 | 来源 | 特征量 | 用途 |
|--------|----|-----------------|---------|
| forge | 模拟 | 26×2 | 二分类数据集 |
| wave | 模拟 | 2 | 回归算法数据集 |
| cancer | 真实 | 569×30 | 未指明 |
| boston | 真实 | 506×13 | 未指明 |

变化范围，所以更大的数据集可以用来构建更复杂的模型。但是，仅复制相同的数据点或收集非常相似的数据是无济于事的。

收集更多数据，适当构建更复杂的模型，对监督学习任务往往特别有用。本书主要关注固定大小的数据集。在现实世界中，你往往能够决定收集多少数据，这可能比模型调参更为有效。永远不要低估更多数据的力量！

1.3 监督学习算法

现在开始介绍最常用的机器学习算法，并解释这些算法如何从数据中学习以及如何预测。我们还会讨论每个模型的复杂度如何变化，并概述每个算法如何构建模型。我们将说明每个算法的优点和缺点，以及它们最应用于哪类数据。此外还会解释最重要的参数和选项的含义，许多算法都有分类和回归两种形式。

1.3.1 一些样本数据集

我们将使用一些数据集来说明不同的算法。其中一些数据集很小，而且是模拟的，其目的是强调算法的某个特定方面。其他数据集都是现实世界的大型数据集。