Python机器学习基础教程

Stephen CUI[1]

February 17, 2023

[1]cuixuanStephen@gmail.com

Chapter 1

无监督学习与预处理

无监督学习包括没有已知输出、没有老师指导学习算法的各种机器学习。在无监督学习中，学习算法只有输入数据，并需要从这些数据中提取知识。

1.1 无监督学习的类型

这里将研究两种类型的无监督学习：数据集变换与聚类。

数据集的**无监督变换**（unsupervised transformation）是创建数据新的表示的算法，与数据的原始表示相比，新的表示可能更容易被人或其他机器学习算法所理解。无监督变换的一个常见应用是降维（dimensionality reduction），它接受包含许多特征的数据的高维表示，并找到表示该数据的一种新方法，用较少的特征就可以概括其重要特性。降维的一个常见应用是为了可视化将数据降为二维。

无监督变换的另一个应用是找到“构成”数据的各个组成部分。这方面的一个例子就是对文本文档集合进行主题提取。

聚类算法（clustering algorithm）将数据划分成不同的组，每组包含相似的物项。

1.2 无监督学习的挑战

无监督学习的一个主要挑战就是评估算法是否学到了有用的东西。无监督学习算法一般用于不包含任何标签信息的数据，所以我们不知道正确的输出应该是什么。因此很难判断一个模型是否“表现很好”。通常来说，评估无监督算法结果的唯一方法就是人工检查。

因此，如果数据科学家想要更好地理解数据，那么无监督算法通常可用于探索性的目的，而不是作为大型自动化系统的一部分。无监督算法的另一个常见应用是作为监督算法的预处理步骤。学习数据的一种新表示，有时可以提高监督算法的精度，或者可以减少内存占用和时间开销。

虽然预处理和缩放通常与监督学习算法一起使用，但缩放方法并没有用到与“监督”有关的信息，所以它是无监督的。

1.3 预处理与缩放

一些算法（如神经网络和 SVM）对数据缩放非常敏感。因此，通常的做法是对特征进行调节，使数据表示更适合于这些算法。通常来说，这是对数据的一种简单的按特征的缩放和移动。Figure 1.1 给出了一个简单的例子：

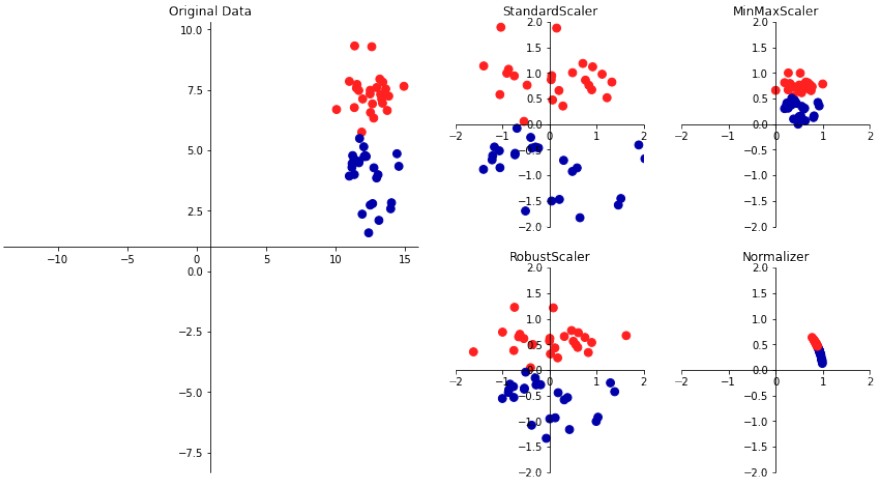

Figure 1.1: Different ways to rescale and preprocess a dataset

1.3.1 不同类型的预处理

1.3.2 应用数据变换

1.3.3 对训练数据和测试数据进行相同的缩放

1.3.4 预处理对监督学习的作用

1.4 降维、特征提取与流形学习

1.4.1 主成分分析

1.4.2 非负矩阵分解

1.4.3 用t-SNE进行流形学习

1.5 聚类

1.5.1 k均值聚类

1.5.2 凝聚聚类

1.5.3 DBSCAN

1.5.4 聚类算法的对比与评估

1.5.5 聚类方法小结

1.6 小结与展望