Python机器学习基础教程

Stephen CUI

February 17, 2023

Chapter 1

数据表示与特征工程

到目前为止，我们一直假设数据是由浮点数组成的二维数组，其中每一列是描述数据点的连续特征（continuous feature）。对于许多应用而言，数据的收集方式并不是这样。一种特别常见的特征类型就是分类特征（categorical feature），也叫离散特征（discrete feature）。这种特征通常并不是数值。分类特征与连续特征之间的区别类似于分类和回归之间的区别，只是前者在输入端而不是输出端。

无论你的数据包含哪种类型的特征，数据表示方式都会对机器学习模型的性能产生巨大影响。额外的特征扩充（augment）数据也很有帮助，比如添加特征的交互项（乘积）或更一般的多项式。

对于某个特定应用来说，如何找到最佳数据表示，这个问题被称为特征工程（feature engineering），它是数据科学家和机器学习从业者在尝试解决现实世界问题时的主要任务之一。用正确的方式表示数据，对监督模型性能的影响比所选择的精确参数还要大。

1.1 分类变量

1.1.1 One-Hot编码（虚拟变量）

1.1.2 数字可以编码分类变量

1.2 分箱、离散化、线性模型与树

1.3 交互特征与多项式特征

1.4 单变量非线性变换

1.5 自动化特征选择

1.5.1 单变量统计

1.5.2 基于模型的特征选择

1.5.3 迭代特征选择

1.6 利用专家知识