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

作为例子，我们将使用美国成年人收入的数据集，该数据集是从 1994 年的普查数据库中导出的。

1.1.1 One-Hot编码（虚拟变量）

到目前为止，表示分类变量最常用的方法就是使用 one-hot 编码（one-hot-encoding）或 N 取一编码（one-out-of-N encoding），也叫虚拟变量（dummy variable）。虚拟变量背后的思想是将一个分类变量替换为一个或多个新特征，新特征取值为 0 和 1。对于线性二分类（以及 scikit-learn 中其他所有模型）的公式而言，0 和 1 这两个值是有意义的，我们可以像这样对每个类别引入一个新特征，从而表示任意数量的类别。

> 我们使用的 one-hot 编码与统计学中使用的虚拟编码（dummy encoding）非常相似，但并不完全相同。为简单起见，我们将每个类别编码为不同的二元特征。在统计学中，通常将具有 k 个可能取值的分类特征编码为 k-1 个特征（都等于零表示最后一个可能取值）。这么做是为了简化分析（更专业的说法是，这可以避免使数据矩阵秩亏）。

将数据转换为分类变量的 one-hot 编码有两种方法：一种是使用 pandas，一种是使用 scikit-learn。

检查字符串编码的分类数据

读取完这样的数据集之后，最好先检查每一列是否包含有意义的分类数据。在处理人工（比如网站用户）输入的数据时，可能没有固定的类别，拼写和大小写也存在差异，因此可能需要预处理。举个例

Table 1.1: The first few entries in the adult dataset

	age	workclass	education	gender	hours-per-week	occupation	income
0	39	State-gov	Bachelors	Male	40	Adm-clerical	<=50K
1	50	Self-emp-not-inc	Bachelors	Male	13	Exec-managerial	<=50K
2	38	Private	HS-grad	Male	40	Handlers-cleaners	<=50K
3	53	Private	11th	Male	40	Handlers-cleaners	<=50K
4	28	Private	Bachelors	Female	40	Prof-specialty	<=50K
5	37	Private	Masters	Female	40	Exec-managerial	<=50K
6	49	Private	9th	Female	16	Other-service	<=50K
7	52	Self-emp-not-inc	HS-grad	Male	45	Exec-managerial	>50K
8	31	Private	Masters	Female	50	Prof-specialty	>50K
9	42	Private	Bachelors	Male	40	Exec-managerial	>50K
10	37	Private	Some-college	Male	80	Exec-managerial	>50K

子，有人可能将性别填为“male”（男性），有人可能填为“man” （男人），而我们希望能用同一个类别来表示这两种输入。检查列的内容有一个好方法，就是使用 pandas Series（Series 是 DataFrame 中单列对应的数据类型）的 `value_counts` 函数，以显示唯一值及其出现次数。

在实际的应用中，你应该查看并检查所有列的值。

用 pandas 编码数据有一种非常简单的方法，就是使用 `get_dummies` 函数。`get_dummies` 函数自动变换所有具有对象类型（比如字符串）的列或所有分类的列（这是 pandas 中的一个特殊概念）。

将输出变量或输出变量的一些导出属性包含在特征表示中，这是构建监督机器学习模型时一个非常常见的错误。

注意：pandas 中的列索引包括范围的结尾，因此 `'age':'occupation_ Transport-moving'` 中包括 `occupation_ Transport-moving`。这与 NumPy 数组的切片不同，后者不包括范围的结尾，例如 `np.arange(11)[0:10]` 不包括索引编号为 10 的元素。

警告

在这个例子中，我们对同时包含训练数据和测试数据的数据框调用 `get_dummies`。这一点很重要，可以确保训练集和测试集中分类变量的表示方式相同。
如果训练集和测试集的数据字段不同，或者字段的未知排列不同，将导致严重的错误!!!

1.1.2 数字可以编码分类变量

在 adult 数据集的例子中，分类变量被编码为字符串。一方面，可能会有拼写错误；但另一方面，它明确地将一个变量标记为分类变量。无论是为了便于存储还是因为数据的收集方式，分类变量通常被编码为整数。例如，假设 adult 数据集中的人口普查数据是利用问卷收集的，workclass 的回答被记录为 0（在第一个框打勾）、1（在第二个框打勾）、2（在第三个框打勾），等等。现在该列包含数字 0 到 8，而不是像 ”Private” 这样的字符串。如果有人观察表示数据集的表格，很难一眼看出这个变量应该被视为连续变量还是分类变量。但是，如果知道这些数字表示的是就业状况，那么很明显它们是不同的状态，不

应该用单个连续变量来建模。

> 分类特征通常用整数进行编码。它们是数字并不意味着它们必须被视为连续特征。一个整数特征应该被视为连续的还是离散的（one-hot 编码的），有时并不明确。如果在被编码的语义之间没有顺序关系（比如 workclass 的例子），那么特征必须被视为离散特征。对于其他情况（比如五星评分），哪种编码更好取决于具体的任务和数据，以及使用哪种机器学习算法。

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