Python机器学习基础教程

Stephen CUI

February 17, 2023

Chapter 1

处理文本数据

文本数据通常被表示为由字符组成的字符串。在上面给出的所有例子中，文本数据的长度都不相同。这个特征显然与前面讨论过的数值特征有很大不同，我们需要先处理数据，然后才能对其应用机器学习算法。

1.1 用字符串表示的数据类型

文本通常只是数据集中的字符串，但并非所有的字符串特征都应该被当作文本来处理。可能会遇到四种类型的字符串数据：

- 分类数据
- 可以在语义上映射为类别的自由字符串
- 结构化字符串数据
- 文本数据

分类数据（categorical data）是来自固定列表的数据。比如你通过调查人们最喜欢的颜色来收集数据，你向他们提供了一个下拉菜单，可以从“红色”“绿色”“蓝色”“黄色”“黑色”“白色”“紫色”和“粉色”中选择。这样会得到一个包含 8 个不同取值的数据集，这 8 个不同取值表示的显然是分类变量。你可以通过观察来判断你的数据是不是分类数据（如果你看到了许多不同的字符串，那么不太可能是分类变量），并通过计算数据集中的唯一值并绘制其出现次数的直方图来验证你的判断。

现在想象一下，你向用户提供的不是一个下拉菜单，而是一个文本框，让他们填写自己最喜欢的颜色。许多人的回答可能是像“黑色”或“蓝色”之类的颜色名称。其他人可能会出现笔误，使用不同的单词拼写（比如“gray”和“grey”），或使用更加形象的具体名称（比如“午夜蓝色”）。

从文本框中得到的回答属于上述列表中的第二类，可以在语义上映射为类别的自由字符串（free strings that can be semantically mapped to categories）。可能最好将这种数据编码为分类变量，你可以利用最常见的条目来选择类别，也可以自定义类别，使用户回答对应用有意义。这样你可能会有一些标准颜色的类别，可能还有一个“多色”类别（对于像“绿色与红色条纹”之类的回答）和“其他”类别（对于无法归类的回答）。这种字符串预处理过程可能需要大量的人力，并且不容易自动化。如果你能够改变数据的收集方式，那么我们强烈建议，对于分类变量能够更好表示的概念，不要使用手动输入值。

通常来说，手动输入值不与固定的类别对应，但仍有一些内在的结构（structure），比如地址、人名或地名、日期、电话号码或其他标识符。这种类型的字符串通常难以解析，其处理方法也强烈依赖

于上下文和具体领域。

最后一类字符串数据是自由格式的文本数据（text data），由短语或句子组成。例子包括推文、聊天记录和酒店评论，还包括莎士比亚文集、维基百科的内容或古腾堡计划收集的 50 000 本电子书。在文本分析的语境中，数据集通常被称为语料库（corpus），每个由单个文本表示的数据点被称为文档（document）。这些术语来自于信息检索（information retrieval，IR）和自然语言处理（natural language processing，NLP）的社区，它们主要针对文本数据。

1.2 示例应用：电影评论的情感分析

1.3 将文本数据表示为词袋

1.3.1 将词袋应用于测试数据集

1.3.2 将词袋应用于电影评论

1.4 停用词

1.5 用tf-idf缩放数据

1.6 研究模型系数

1.7 多个单词的词袋（n元分词）

1.8 高级分词、词干提取与词形还原

1.9 主题建模与文档聚类

1.9.1 隐含狄利克雷分布