机器学习工程师纳米学位

毕业项目开题报告

宋振
2018 年 11 月 26 日

开题报告

领域背景

Rossmann 是欧洲的一家药品连锁店，在 7 个欧洲国家拥有 3,000 家药店。Kaggle 比赛其中有一个是“Rossmann Store Sales”，要求是预测 Rossmann 未来的销售额。通过了解，可以知道 Rossmann 有多种店铺类型，如果能够根据 Rossmann 药妆店的信息（比如促销，竞争对手，节假日）以及销售情况，实现对 Rossmann 不同类型药妆店的分类的预测，就有助于商家在各州、市的不同位置借助各类环境信息，合理判断应当设置的商店种类。

上面的问题属于机器学习领域的分类问题。采用的方法是监督学习，即把要预测的特征作为“标签”，利用其余特征训练学习器，用训练好的学习器对新的数据进行预测。这方面的案例有“基于 python 的商品购买能力预测”[1]等。

对客户等的分类问题是商业上经常遇到的问题，我希望通过这个项目，加深对监督学习的学习，学会利用机器学习这个工具完成这项工作。

问题描述

这里要解决的问题是，利用 Rossmann 积累的大量的销售数据和店铺信息数据，从中提取“店铺类型”作为标签，其余特征如“销售额”，“竞争对手”，“节假日”，“促销”等信息作为判断店铺类型的依据，最终通过监督学习的方法，根据不同额店铺信息数据进行店铺类型的分类。

数据集和输入

项目使用的数据集是在 Kaggle 上下载[2]的“train.csv”和“store.csv”，这里面分别包括了与 Rossmann 的销售额有关的信息，以及不同类型店铺的信息。虽然只有“store.csv”里面包含了店铺类型标签，但是为增加特征数量，提高预测的准确率，需要在预测时结合这两个文件的信息。

使用“train.csv”和“store.csv”合并后的数据集，提取“店铺类型”作为标签，其余字段作为特征集，在训练模型时，将标签和特征拆分为训练和测试集，用训练集训练模型，用测试集检验模型性能。

解决方案描述

对于判断商店类型的分类问题，这里选用的监督学习模型是决策树 (DecisionTree)。决策树模型通过从数据特征中学习决策规则来训练模型，运行速度较快，并且，针对本项目中使用的训练数据——含有缺失值，多类别字段，时间序列字段等，决策数模型还具有以下优势：

- 能够同时处理数据型和常规型属性；
- 在相对短的时间内能够对大型数据源做出可行且效果良好的结果；
- 对数据集中的缺失值不敏感；
- 可以处理不相关特征数据；
- 效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度；

在建立 DecisionTree 模型，还可以使用网格搜索算法优化模型参数。

基准模型

鉴于该分类问题是多分类问题（共有 4 种店铺类型），即对每一条要预测的数据，其属于任何一种店铺类型的概率为 1/4，即本项目采用 0.25 作为基准值，此为假设值，项目中使用的决策树模型准确率得分越高则模型表现约好。

评价指标

要衡量预测的效果，我们需要一个根据实际商店类型结果对预测进行打分的指标。因为我们对能够准确预测店铺类型感兴趣，因此我们使用准确率（accuracy_score）作为评价模型的标准是合适的。这里准确率计算的是对测试数据集中预测正确的店铺数量占测试集样本总数的比例，即：

$$\text{accuracy\_score} = \frac{\text{测试集中预测正确的店铺数量}}{\text{测试集中的样本总数}}$$

项目设计

- 数据熟悉

分别探索销售信息“train.csv”和店铺信息的“store.csv”，了解各字段含义，查看数据文件中的缺失值和异常值；

- 数据探索

清理数据集中的异常值，填充缺失值，从时间序列类型字段中分离出新的特征，只有将处理后两个数据集使用“pd.merge”语句合并数，将“店铺类型”与其他特征的关系，用 Seaborn 工具绘制诸如直方图和折线图等进行可视化；

- 数据预处理

对特征数据进行独热编码，准备用于学习器的训练；

- 训练和优化模型

利用 train_test_split 将标签和特征数据拆分为训练集和测试集，用于训练决策树模型；对于对决策树模型性能影响较大的超参数如“min_samples_split”等使用 GridSearchCV 进行优化；

- 模型评估

使用准确率得分，评估模型在测试集上的得分；

- 总结

总结在项目中的收获，分析不足和需要改进的方面；

备注

[1]基于 python 的商品购买能力预测

[2]Rossmann Store Sales 数据下载