机器学习纳米学位

毕业项目开题报告

赵鹏举

2017-03-16

ROSSMANN药店销售额预测

项目背景

销售预测对于每一个企业都非常重要，机器学习的方法在其中得到了非常广泛的应用，掌握进行预测的常用方法和工作流程对于以后从事数据分析工作有着巨大的现实意义。

本项目来自Kaggle比赛Rossman Store Sales. 截至2015年，Rossmann在欧洲7个国家运行着超过3000家连锁药店，这些药店的营收会受到促销、竞争者、国家和学校假期、季节、地域等因素的影响。Rossmann希望参加比赛项目的选手，可以准确地地预测出位于德国1115家药店在六周内每天的销售情况；进而利用可靠的销售预测情况帮助药店经理制定更加高效的工作安排。

比赛过程中，第一名Gert在原有数据集基础上，构造出临近信息、时间信息、趋势信息等特征，并采用XGBoost方法训练模型；第三名Cheng Guo 将原本主要用于自然语言处理的深度学习entity embedding模型应用到类别特征中，取得第三名。

问题描述

本问题为监督学习中的回归问题：已知1115家药店的信息以及每家药店在2年多时间内每天的销售情况，需要对接下来6周内每家药店的销售状况进行预测。回归问题的常见机器学习方法有K近邻学习、线性回归、决策树、随机森林、XGBoost、神经网络等；而实际中，为了训练出效果较好的模型，一般需要根据数据集的特点，进行特征工程，构造出有用的新特征，并对特征进行选择，同时注意防止过拟合。

在本问题提供的数据集中，销售数据作为标记值，其他属性作为特征，对选择的模型进行训练；模型的预测销售数据与标记销售数据之间的差异可以用来对模型进行评估；训练好的模型，对测试数据的预测是可以再现的。

输入数据

输入数据包含 train.csv和store.csv：

- train.csv是历史销售数据，每条信息包含了药店编号、日期、星期几、是否营业、是否节假日、是否促销、当日销售额以及客户数量；
- store.csv是药店补充数据，每条信息包含了点药店编号、药店类型、商品组合、最近竞争者距离及开店时间、促销有无、促销间隔和开始时间。

输入的数据对于药店销售预测是非常有用的：日期和星期几等可以提供销售额周期性的时间标定；是否节假日和促销，以及每家药店的信息和竞争者的信息，对于销售额也会有一定影响。

解决办法

本项目将尝试三种方法进行最终预测：

- 方法1：采用XGBoost方法；XGBoost模型是一种有监督的集成学习方法，可以直观理解为对决策树的集成，是非常有效的解决非结构化数据的方法，在Kaggle比赛中得到广泛的应用；
- 方法2：采用Entity Embedding方法；Entity Embedding属于深度学习中处理自然语言的重要方法，用来表示不同单词之间的关系，本文将用来研究不同特征之间的关系。
- 方法3：前两种方法的集成；

本项目将基于python进行实现，将会主要用到numpy、pandas、matplotlib、seaborn、sklearn、XGBoost、TensorFlow、Keras等函数库。

基准模型

本项目将采用两个基准模型：

- 基准一：具有相同特征参数数据子集的中位数；采用的特征包含药店编号、星期几、是否促销、药店类型、商品组合、是否节假日等；
- 基准二：Kaggle Private/Public Leader Board；Kaggle Leader Borad记录了参赛者的预测结果和最终名次，所以可用来衡量本项目能够达到怎样的最终预测结果；Public Leader Board使用了33%测试数据的预测结果，而Private Leader Board使用了67%测试数据的预测结果。

为了避免对测试数据的过拟合而导致的效果提升，本项目在完成过程中，将主要根据输入数据中划分出来的验证集对模型进行评估和优化，用Public Leader Board作为训练过程中过拟合的验证手段，仅在最终用Private Leader Board对结果进行评估。

评估指标

本项目采用Kaggle比赛的评估指标：RMSPE（误差百分比的均方差），可表示为

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y_i}}{y_i})^2}$$

其中，任何当天销售额为0的数据在评估时将被忽略； y_i 表示某药店在某天的实际销售额，而$\hat{y_i}$ 表示该药店在对应这一天的预测销售额。

设计大纲

工作流程：

1. 数据读入：将train.csv、store.csv、test.csv读入
2. 数据可视化：针对不同的特征，对数据进行可视化；
    1. 展示不同分类特征（如星期几、节假日、促销、药店类型、商品组合等）对于销售额的影响
    2. 展示时间序列对于对于销售额的影响，比如指定店铺随着时间的销售额变化、不同因素（促销、装修、竞争等）产生的销售额变化；
    3. 观察是否有异常数据
    4. 展示不同特征之间的相关性
3. 数据整理和特征工程
    1. 对数据中缺失值、异常值进行处理
    2. 分析数据的相关性，并进行数据降维处理
    3. 进行数据聚类分析
    4. 根据需要，对数据进行清理（如归一化、采用对数处理改善分布状况）、聚合
    5. 构造更多的特征（如更详细时间信息、趋势信息等）
4. 训练基准模型
    1. 划分训练集、验证集、测试集
    2. 训练并评估基准模型
5. 训练XGBoost模型
    1. 进行特征选择，找出效果较好的特征组合；
    2. 训练不同的模型
    3. 对较好的模型进行集成
    4. 评估模型
6. 训练、评估、优化entity embedding模型
    1. 选择合适的特征，搭建entity embedding模型
    2. 搭建神经网络模型
    3. 训练、评估、优化模型及特征选择
7. 对XGBoost模型和entity embedding模型进行集成，并评估
8. 结果分析并完成报告