机器学习工程师纳米学位

报告人/时间

何龙

2018-09-28

开题报告

领域背景

Rossmann 在全欧洲有超过 6000 家药店，预测销售额一直是他们商店经理的工作，他们根据直觉来预测，准确率有很大变化，现在我们要帮助构建一个销售额预测模型，针对位于德国的 1115 家店进行 6 周的销售额预测，对于销售额的预测可以帮助经理们更合理的安排员工上班时间表、销售活动等。

使用机器学习去预测能够达到人不能做到的准确率，因为模型考虑了整个分布在德国的商店的数据，并结合这些数据得到结论，相比较，商店经理更多的只是依赖自己管理的商店的数据，我们直到在数据量少的时候结果是无法泛化的，准确率也就无法保证，这是一个典型的机器学习有监督的回归问题，相关算法模型已经非常成熟，同时拥有 Rossmann 提供的大量数据保证了这一问题是可以解决的。

这是一个非常普遍的机器学习应用场景，相信在现实中也有无数公司在利用机器学习进行自己的销售额预测，解决一个现实的问题会让我更加对机器学习有自信，相信它是未来社会不可或缺的工具，并且跟人们的生活息息相关，解决起来体会会更深。

问题描述

针对位于德国的 1115 家店进行 6 周的销售额预测，销售额是一个数值型数据，所需数据 kaggle 已经提供，其中包含训练数据、商店数据、测试数据，根据评价指标，也就是预测值与实际值的差值越小越好。

数据集和输入

本项目数据集来源于 kaggle，数据本身是针对该项目收集的数据，所以可以使用他们来进行模型的训练，对于额外的 store.csv 数据，它是针对每个商店的具体信息，既然是分析商店的销售额，它肯定是非常有用的，因此要想办法结合进 train.csv 中一起处理，同时对于类别数据要多关注，时间序列相关的字段也要特别处理。

解决方案

首先对于数据的预处理中要关注 store.csv 以及类别字段、时间序列字段等，由于处理后维度很高，此处需要进行主成分分析，模型上使用集成学习 Adaboost，根据 kaggle 指定的评价指标 RMSPE 进行模型评估以及调参。

基准模型

基准模型选择恒预测为 mean 值，使用评价指标对于基准模型和所选模型的预测分别进行计算，并比较二者差别。

评价指标

`RMSE(Root Mean Square Error):sqrt((sum(X_obs_i-X_model_i)^2)/n)`，均方根误差用于描述预测值与实际值之间的偏差，应用于本项目即为预测销售额与实际销售额之前的偏差，可以很好的评估模型的性能。

项目设计

1. 工作流程

    0. train.csv、store.csv 数据读取。

    1. 根据 store Id 链接成一个大表。

    2. One-Hot 处理类别字段。

    3. 暂时不处理处理时间序列。

    4. PCA。

    5. 划分为训练数据和验证数据。

    6. 应用模型训练。

    7. 对比模型与基准模型在评价指标上的得分。

    8. 持续优化模型。

    9. 上传到 kaggle。

2. 采用什么策略

    PCA+Adaboost。

3. 需要对数据进行哪些前期分析

    0. train 与 store 的链接。

    1. 分类字段的 One-Hot 编码处理。

    2. 时间序列字段的针对性处理。

    3. PCA 处理。

4. 算法是否有更详细的讨论

    1. train 与 store 的链接，以及链接后的字段保留与否。

    2. 时间序列如何使用。

    3. Adaboost 算法细节。

5. 考虑包含小的可视化、伪代码、图表等信息来更好的描述项目设计

1. 不使用 store、使用 store 不处理时间序列、使用 store 且处理时间序列，三种不同数据处理下的模型预测结果对比。

2. 最优的数据处理下的模型与基准模型的结果对比。

3. Adaboost 对比其他模型的结果对比。