机器学习纳米学位——预测Rossmann商店销售额

刘晟西

2018年5月31日

开题报告

项目背景

Rossmann是欧洲的一家连锁药店。 在这个源自Kaggle比赛**Rossmann Store Sales**中，我们需要根据Rossmann药妆店的信息（比如促销，竞争对手，节假日）以及在过去的销售情况，来预测Rossmann未来的销售额。

销售预测在零售行业对于运营有重大的指导意义，直接关系到公司的资源调配，人员的安排等等，进而影响公司成本和效率，著名快时尚公司zara正是得力于有力高效的极致供应链，压缩仓储和运营成本，取得了成功。

在没有电子化的过去，商业实践中多采用一些指标式的商业模型来进行预测。在计算机时代到来之后，业界开始采用一些数据挖掘手段来进行销售预测，主要是一些统计分析方法，比如时间序列分析、线性回归模型分析、非线形回归模型分析等等。但是，由于影响销售是由多种因素综合决定的，而神经网络作为一种非线性、自适应动力学系统，逐步被用来解决销售预测问题。这当中最引人注目的就是在时尚销售领域的算法研究，先后有极限学机（Extreme Learning Machine, ELM），扩展的极限学机（Extended ELM, EELM）算法等等算法被提出。[1][2]

问题描述

项目需要根据已知的各个商店的多个特征域的值以及过往销售数据来对未来的销售数据进行预测，本质是回归问题。由于有历史数据，所以是一个监督学习的问题。对于此类问题，目前kaggle上比较流行

采用xgboost模型进行训练，此外，近期极限学机在销售预测方面也不少论文发表。但考虑到极限学机目前争议性还比较大，因此考虑采用xgboost模型来解决问题。该问题的各个特征值可用数值或可以进行独热编码，预测目标销售可以用数值表示，因此是可以量化的。同时，可以预计各个店面的销售与其某些特征相关，具有类似特征的店面应当体现出类似的销售表现，因此问题是可复现的。

输入数据

该项目的数据可以在其**kaggle页面**下载。

其数据包含4个文件：

- sample_submission.csv：正确格式的提交文件的样本
- store.csv：关于商店的补充信息
- test.csv：含销售数据的历史数据
- train.csv：不含销售数据的历史数据

在store.csv文件中给出了1115家商店的信息，除了商店的编号之外，还包含了9项额外的信息，这些信息中包含5项离散分类信息（不含编号）和4项连续数值信息，在数据预处理时，可将离散分类信息进行独热编码，构成完整的特征向量，用于训练。

在train.csv中给出了这些商店从2013年1月1日到2015年7月1日的销售数据，相关信息主要是与特定日期相关的特征数据，共6个维度（不含编号和ID）。

所以总体来看就是提供了2种信息，一种是商店的特征，另一种是特定时间特征，从常识可以推测，类似种类的商店在类似的日期会有相似的销售额。

由于数据本身的特征维度较多，在具体使用中，可能要考虑使用PCA进行处理。

解决方法

拟采用流行的xgboost工具包来训练。xgboost被证明在数据挖掘中效果很好的工具包，有大量的kaggle队伍采用该工具包，在2015年29支Kaggle冠军队伍中就有17队采用xgboost工具包。xgboost是对GBDT（Gradient Boost Decision Tree）的一种改进。这里面涉及boosting的概念。Boosting方法是指对一份数据，建立多个模型，这些模型都比较简单，称为弱分类器，然后每次分类都将上一次分错的数据权重提高再进行分类，最终得到的分类器即可在测试数据与训练数据上都取得比较好的成绩。[3][recite3]而Gradient Boosting是一种Boosting的方法，它的主要思想是，每一次建立模型是在之前建立模型损失函数的梯度下降方向。[3][recite3]而拟使用的XGBoost则是对GBDT在多个方面进行了改进。关于XGBoost为什么在竞赛中表现优越，挪威科技大学的Didrik Nielsen的硕士论文**《使用XGBoost的树提升：为什么XGBoost能赢得“每一场”机器学习竞赛》**尝试做出解答，作者认为，一方面，提升树模型很好的在高纬度问题中“打败了”维度的诅咒，另一方面，XGBoost在每一轮的训练中使用更高阶的近似，因此可以更好的学习。此外，XGBoost在自适应决定每个树的终端节点数上面，以及调整树的权重以更好地减小方差上面采用更聪明的策略。综合的结果就是，XGBoost成为了竞赛中最受欢迎的算法。

就该项目而言，其本身是一个高纬度的时间序列回归问题，因此可以采用XGBoost来量化和解决。

基准模型

如解决方法所言，基准模型将采用基于XGBoost算法的决策树模型。关于模型的各个参数需要在解决过程中进行优化尝试。

评价指标

评价指标沿用Kaggle上的评价指标均方根百分比误差RMSPE：

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

设计大纲

数据预处理

- 从kaggle下载相关数据
- 对数据进行预处理，包括归一化（特征标准化）以及PCA分析等
- 经过预处理后形成完整干净的训练数据

模型搭建

- 调用xgboost进行模型的搭建

模型训练/调参

- 使用训练数据对模型进行训练
- 根据实际训练的成果对模型参数进行调整，尝试不同的弱分类器

模型评估

- 使用RMSPE进行模型评估，上传kaggle进行最终评估

参考文献

[1]:刘卫校，基于离散灰色预测模型与人工神经网络混合智能模型的时尚销售预测．计算机应用 2016, Vol. 36 Issue (12): 3378-3384 DOI: 10.11772/j.issn. 1001-9081.2016.12.3378

[2]:MBA百科

[3]:[机器学习——Gradient Boost Decision Tree(&Treelink)][https:// blog.csdn.net/yangtrees/article/details/7506052]