Machine Learning Engineer Nanodegree

Rossman 销售预测开题报告

战柏瑞
2018 年

主要背景

该项目是 Kaggle 上 Rossmann 公司举办的一个竞赛项目.Rossmann 是德国第二大药品销售链,在欧洲 7 个国家拥有近 3600 家药店.公司由 Dirk Rossmann 建立于 1972 年[i].在 2015 年 Rossmann 的管理层被指派预测近 6 周的日销售额.销售额会被很多因素印象例如, 促销,竞争对手,学校以及州节假日,季节性,和地域性.[ii]

问题称述

在分类中，我们想了解模型隔多久正确或不正确地识别新样本一次。而在回归中，我们可能更关注模型的预测值与真正值之间差多少。[iii] 通过对 Rossmann 数据分割,得到 35 天的销售数据,然后使用 xgboost 进行回归预测,因为此次项目更关注预测值和真实值之间的差,所以是一个回归问题.

数据集与输入

使用 Kaggle website (https://www.kaggle.com/c/rossmann-store-sales)的数据,将 Train.csv 进行分割,从而获得最近的 35 天数据集作为测试集.

```
# 显示train特征
df_train.head()
```

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	1	5	2015-07-31	5263	555	1	1	0	1
1	2	5	2015-07-31	6064	625	1	1	0	1
2	3	5	2015-07-31	8314	821	1	1	0	1
3	4	5	2015-07-31	13995	1498	1	1	0	1
4	5	5	2015-07-31	4822	559	1	1	0	1

训练集具有 9 个特征,其中 Sales 作为结果应从特征中分离出来. Date 特征中需要从字符串处理为年月日三个不同的特征.对于 StateHoliday 进行 one-hot 编码.最后一共有 13 个特征值.

```
# 显示store特征
df_store.head()
```

	Store	StoreType	Assortment	Competition Distance	Competition Open SinceMonth	Competition Open SinceYear	Promo2	Promo2 SinceWeek	Promo2 SinceYear	Promo Interval
0	1	c	a	1270.0	9.0	2008.0	0	NaN	NaN	NaN
1	2	a	a	570.0	11.0	2007.0	1	13.0	2010.0	Jan,Apr,Jul,Oct
2	3	a	a	14130.0	12.0	2006.0	1	14.0	2011.0	Jan,Apr,Jul,Oct
3	4	c	c	620.0	9.0	2009.0	0	NaN	NaN	NaN
4	5	a	a	29910.0	4.0	2015.0	0	NaN	NaN	NaN

Store 数据集是训练集的衍生部分,通过共有 10 个特征值,对其进行 one-hot 编码,最后有 17 个特征值.对于缺省值,competition distance 可以用平均值 meadian 代替,别的缺省值由 xgboost 自行处理(默认处理 0)[iv]. 然后对 store 数据集和训练集以 Store Id 对应的方式进行合并填充

解决方案

此次项目通过 XGboost 来完成. XGBoost 实现的是一种通用的 Tree Boosting 算法，此算法的一个代表为梯度提升决策树（Gradient Boosting Decision Tree, GBDT），又名 MART(Multiple Additive Regression Tree)。GBDT 的原理是，首

先使用训练集和样本真值（即标准答案）训练一棵树，然后使用这棵树预测训练集，得到每个样本的预测值，由于预测值与真值存在偏差，所以二者相减可以得到“残差”。接下来训练第二棵树，此时不再使用真值，而是使用残差作为标准答案。两棵树训练完成后，可以再次得到每个样本的残差，然后进一步训练第三棵树，以此类推。树的总棵数可以人为指定，也可以监控某些指标（例如验证集上的误差）来停止训练。[v]

评估标准

通过对测试集的 Root Mean Square Percentage Error (RMSPE)对已知模型进行评价.

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$ [vi]

Y_i 为单个门店单天的销售 $Y_i$hat 是对此进行的预测

基准模型

设定基准阈值为 kaggle 排行榜前 10%（330/3303），也就是在 Private Leaderboard 上的分数要低于 0.11737。[vii]

项目设计

- 数据预处理

  此阶段是对数据进行分割,以产生本地测试集以及训练集,对 Store 数据集进行合并,以及简单的时间转化成年月日,对其他特征进行 one-hot 编码.

- 模型搭建

此阶段对特征的重要性进行排序,去掉不必要的特征,还有重复特征.

- 模型训练

  训练模型,对训练集进行分割,从而训练 N 次,每次对模型进行修正.

- 模型调参

  通过参照 XGboost API[viii]进行调参

  例如,max_depth 尝试不同深度对模型的影响,从而选定合理的数值.

  Learning_rate,调整学习率,太大的学习率可能无法收敛.

  n_estimators:调整有几个 boosted tree.

  以及其他[ix]

- 模型评估

  同评估标准

- 可视化
- 对必要以及可观的数据进行可视化,从而更好的了解数据

参考文献

[i] https://en.wikipedia.org/wiki/Rossmann_(company)
[ii] https://www.kaggle.com/c/rossmann-store-sales#description
[iii] Udacity,机器学习(进阶)-课程 12 评估指标-分类指标与回归指标
[iv] http://xgboost.readthedocs.io/en/latest/faq.html?highlight=missing
[v] http://www.a-site.cn/article/714295.html
[vi] https://www.kaggle.com/c/rossmann-store-sales#evaluation
[vii] https://www.kaggle.com/c/rossmann-store-sales/leaderboard

[viii] http://xgboost.readthedocs.io/en/latest/python/python_api.html
[ix] http://blog.csdn.net/sb19931201/article/details/52557382