机器学习纳米学位

毕业项目 张立芹 优达学城

2018 年 06 月 04 日

项目背景：

- 在机器学习研究中，分为监督学习与非监督学习两大类。监督学习是通过历史的数据学习，该数据有特征信息（输入信息）和目标信息（预测信息），通过分析得出输入与输出之间的关系搭建模型，然后通过模型输入的特性信息进行预测。非监督学习是对特征信息进行聚类分析，但没有目标结果。
- 监督学习可以解决回归问题和分类问题，但其对数据比较敏感。目前在金融领域已经广泛运用，所以监督学习实用价值会越来越高。
- 以往人们都是根据积累的经验判断的出结果，这样很难一个量化标准去得出结果。而机器学习通过使用数学的方式，对大量的数据进行训练。从而金融、销售领域等在在大数据的基础上，能通过分析，训练数据，得出一定的预测结果。

问题描述：

Rossmann 是德国最大的日化日用品超市，拥有 3000 多家实体店和网店，并且很受中国游客、商务人士和留学生购买。每家商店的经理需要预测未来 6 周的日销售情况。每个经理根据自己的经验等依据预测销售情况但是结果的准确率不高。根据 Rossmann 商店的信息，比如促销、竞争对手、节假日、季节、气候、地域等因素，来预测 Rossmann 未来 6 周每天的销售额。提供可靠的预售销售额，可以有效管理商品，免受商品积压，带来的损失，同时也提供经理更加预售情况合理安排人员，所以提供给他们一个预测模型是非常有需要的。使用 1115 家门店的历史销售数据情况研究，分析后的数据进行创建模型，用于预测六周后的日销售情况。

对于有特征-标签模式的数据时，通过机器学习可以转化为简单数学问题，这类问题通过对大量数据的训练。而该项目提供了大量数据，所以可以根据该数据进行监督学习，得到预测结果。

输入数据：

- *输入数据集*
  train.csv 表示过去的销售数据为训练集。包含字段：
  Store,DayOfWeek,Date,Sales,Customers,Open,Promo,StateHoliday,SchoolHoliday
- test.csv 表示测试数据。包含字段：
  Id,Store,DayOfWeek,Date,Open,Promo,StateHoliday,SchoolHoliday
- sample_submission.csv 预测数据格式样本。包含字段：Id,Sales
- store.csv 表示门店信息。包含字段：
  Store,StoreType,Assortment,CompetitionDistance,CompetitionOpenSinceMonth，
  CompetitionOpenSinceYear,Promo2,Promo2SinceWeek,Promo2SinceYear,PromoInterval
- *数据集特征如下*
  Id - 测试集中表示一条记录的编号。
  Store - 每个商店的唯一编号。
  Sales - 任意一个给定日期的销售营业额。
  Customers - 给定那一天的消费者数。
  Open - 商店是否开门标志，0 为关，1 为开。
  StateHoliday - 表明影响商店关门的节假日，正常来说所有商店，除了极少数，都会在节假日关门，a=所有的节假日，b=复活节，c=圣诞节，所有学校都会在公共假日和周末关门。
  SchoolHoliday - 表明商店的时间是否受到公共学校放假影响。
  StoreType - 四种不同的商店类型 a，b，c 和 d。
  Assortment - 描述种类的程度，a = basic, b = extra, c = extended。
  CompetitionDistance - 最近的竞争对手的商店的距离。
  CompetitionOpenSince[Month/Year] - 最近的竞争者商店大概开业的年和月时间。
  Promo - 表明商店该天是否在进行促销。
  Promo2 - 指的是持续和连续的促销活动。: 0 = 商店没有参加, 1 = 商店正在参加。
  Promo2Since[Year/Week] - 表示参加连续促销开始的年份和周。
  PromoInterval - 描述持续促销间隔开始，促销的月份代表新的一轮，月份意味着每一轮的开始在哪几个月。
- 相关的初步统计分析如下
  test.csv 有 41088 条数据，特征 Open 有 11 条缺失值，但是根据 Date、StateHoliday 可以说明其实 1，所以使用 1 填充缺失值。其中缺失数据是在特征 Open 上缺失了 11 条，根据缺失的 Date。

以下是前五条输出样例

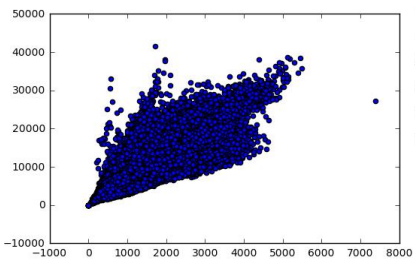
	Id	Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
0	1	1	4	2015-09-17	1.0	1	0	0
1	2	3	4	2015-09-17	1.0	1	0	0
2	3	7	4	2015-09-17	1.0	1	0	0
3	4	8	4	2015-09-17	1.0	1	0	0
4	5	9	4	2015-09-17	1.0	1	0	0

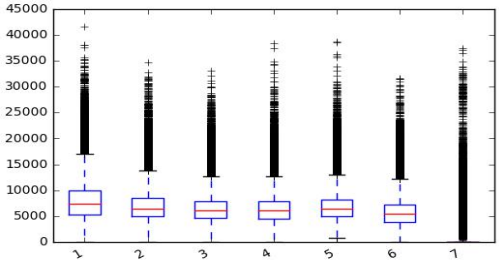
- store.csv 有 1115 条数据其中缺失数据特征是 CompetitionDistance，CompetitionOpenSinceMonth，CompetitionOpenSinceYear，Promo2SinceWeek，Promo2SinceYear 和 PromoInterval。CompetitionDistance 缺失 3 条，我推测这三家商店在有效的距离内没有竞争对手用一个特别大的值来处理，CompetitionOpenSinceMonth 和 CompetitionOpenSinceYear 缺失的情况一直，我推测就在很早之前或者说在 train 数据之前就存在这个竞争对手了我给一个默认的之前的时间 2010 年，Promo2SinceWeek，Promo2SinceYear 和 PromoInterval。这三个特征的确是情况也都一样，也就是没有参加 Promo2 的这三项均为空，那我就将时间设置为一个未来时间 2030 年，PromoInterval 用“0，0，0，0”来填补。
- train.csv 有 1017209 条数据，无数据缺失情况，所有提供的数据总体来看没有异常值情况。

  样本如下

	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

- 观察数据训练集数据分布情况：从下图可看出，客户数与销售额有关，且周末是不开门，所以根据提供的数据，可以有效的帮助我们预测分析。





解决办法：

- 查看缺失值，如果缺失值占比超过 90%，将移除该特征，其他的使用均值或中位数进行补充。根据单个特征分布情况，对异常值进行移除。
- 对特征值的相关度分析获取对模型有效变量。
- 使用 Train 和 Test，进行 XGBoost 模型训练与验证
- 通过不断的尝试和搭配 XGBoost 算法的高级用法来寻找到最适合的参数。
- 使用可视化结果展现

基准模型：

检验模型是否达标，需要建立基准模型，根据这个基准指标来衡量模型的合理性。通过参考抗过了比赛排名结果，个人初步定的基准为 0.25。根据项目要求 Leaderboard 的 10%作为基准，同理可得，测试集的评分可达到 0.11773。

评估指标：

一般情况下，会使用数据集中的一部分作为训练集，其中另一部分为测试集。对于二分类问题，可以使用准确率、对数损失函数、ks、AUC 等评分方法。但是对回归的数据预测问题，使用平方根误差（RMSE）或者均方根百分比误差指标衡量模型效果，同时也可以使用召回率。由于 RMSP 对数值的绝对值大小不敏感，更加适合作为评价模型的指标。

使用 RMSPE 来做为验证函数，该值越低代表差异性越小。它是指模型的预测值和实际观察值之间的差异的一种衡量方式。其公式如下：

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

设计大纲：

- 分别查看 train、test 和 store 数据情况。
- Store 与 train 和 test 进行合并

- 获取 train 的 70%为训练集，30%为测试集。Test 作为验证集。
- 对数据做清洗和整理。
- 对异常值进行移除处理。
- 对数据变量分布情况可视化分析。
- 进行模型训练。
- 配置 xgboost 参数并开始第一次的训练。
- 对第一次训练进行总结，再对特征做进一步的优化，对参数做进一步的优化
- 开始第二次训练，并通过配置 xgboost 参数的方式来寻找最佳超参，调参过程使用 gridsearch 方式。
- 记录训练的评估结果并保存模型。
- 用模型对 test 数据做 predict 并将结果按照 submission 的方式保存。
- 寻找最佳模型
- 记录训练的评估结果并保存模型。
- 用模型对 test 数据做 predict 并将结果按照 submission 的方式保存。

引用：

http://xgboost.readthedocs.io/en/latest/model.html

http://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model

https://www.kaggle.com/beiwenwu/xgboost-in-python-with-rmspe