项目学习总结

贾开程 | 2019 年 12 月

目录：

Predict Future Sales 项目表

Week	EDA	特征工程	训练结果	课堂反馈
3	train： 整体销量呈下滑趋势，商店有新开与关闭，商品有新上与下架 test： 存在训练集中没有的商品-商店组合，分成 3 类占比 shop： 商店名称有重复，且名称可拆为城市、种类、名称 category： 名称可拆为子类型 1，子类型 2	增加 4 个 feature *(city & type)* *(type & subtype)*	暂未训练	模型会考虑 outlier，不需要删除，另外可以尝试训练模型
4	沿用 Week3	沿用 Week3	linear regression：/ random forest：/ lightgbm： ranking 57%， ame = 0.406 xgboost： ame = 0.453	依据 common sense 尝试挖掘新特征：features interactions、周期性特征等 对于参数仅做个别适当调整即可
5	沿用 Week3	增加个 19 个时间序列 feature 以及 15 个其他特征	lightgbm： ranking 36%+， rmse = 0.95 lightgbm： ranking 20%+， rmse = 0.90	可以尝试模型 ensemble
6	答辩	答辩	答辩	1、针对 one-hot code 的弊端，可以采用 catboost 2、产品生命周期特征可以尝试加入 3、模型 ensemble 有两种

Final Week：

模型融合、catboost 学习资料
https://zhuanlan.zhihu.com/p/83426796?utm_source=wechat_session&utm_medium=social&utm_oi=975869084436553728

Week 4 & Week 5：

以下内容除了 gplearn 外均已学习，另外，两个模型的原理的理解略难...似懂非懂

特征工程内容（比较全）：
https://zhuanlan.zhihu.com/p/26444240?utm_source=weibo&utm_medium=social

label encoding 与 one hot encoding 的区别：
https://zhuanlan.zhihu.com/p/36804348 from 知乎
https://www.cnblogs.com/king-lps/p/7846414.html from 博客园

gplearn 遗传算法：（创建新特征的一种方式，没看懂）
https://gplearn.readthedocs.io/en/stable/index.html from 官网
https://zhuanlan.zhihu.com/p/31185882 from 知乎解析
https://bigquant.com/community/t/topic/120709 from 案例解析

lightgbm 原理+代码实现+参数
https://blog.csdn.net/huacha__/article/details/81057150
https://blog.csdn.net/lomodays207/article/details/88045852
https://www.jianshu.com/p/3f114699c6ed

xgboost 原理+代码实现+参数
https://blog.csdn.net/qq_19446965/article/details/82079486
https://blog.csdn.net/qq_29831163/article/details/90486802
https://blog.csdn.net/iyuanshuo/article/details/80142730

Lecture 3：

Random forest

1）random forest 的特点：

- 随机性：子数据集抽取的随机性、每个 tree 的问题选择的随机性
- 使用范围：continuous、catagory 问题都能使用
- 优点：能够避免 overfitting
- 缺点：不可解释性（casual random forest 这篇文章解释了），以及 hyperparameter 需要人为调试参数

2）random forest 的 R 实现：

- randomForest 包
- set.seed（ ）：保证多次重复运行结果一致
- importance = T： 衡量变量的重要性
- importance type1（代表重要性） type2（代表熵减）
- ntree 默认 500 个，ntree 代表树数，mtry 代表问题数
- 用双 for loop 测试最优结果

Neural Network

1）neural network 的特点：

- 优点：能够反应看不到的关系
- 缺点：运行时间太长，且结果不可解释

2）neural network 的 R 实现：

- neuralnet 包
- hidden 代表层数，node 代表个数
- act.fuc
- Linear.output 代表是否输出线性结果
- step.max 代表迭代次数（防止 crash）
- compute 类似 predict

Sentiment analysis

1）sentiment analysis 的特点：

- scalability、real-time analysis、consistant criteria
- 应用领域：如搜集新产品的 insight 等
- Polarity Analysis 两极
- Valence Shifters：包含转折、否定、加强等
- Emotion Analysis：可检测出 16 种不同的情感关键词

2）sentiment analysis 的 R 实现：

- Sentimentr 包
- sentiment（ ）
- element_id：单个元素
- sentence_id：单个句子
- sentiment_by（ ）：按元素不按句子
- emotion（ ）
- emotion_by（ ）

Lecture 2：

Machine Learning

1） **Supervised Learning：结果已知**

Linear regression

Logistics regression

Probit regression

Support vector machines

Random forest

Neural networks

2） **Unsupervised Learning：结果未知**

K-Means clustering

Hidden markov models

3） Reinforcement Learning

Prediction methods

1） **Linear regression**

假设误差符合正态分布，可以被解释，属于连续型变量。也可以把因变量转化为 category variables 以解决 classification problem，但不太好。

Loss function（对称）

- Quadratic loss function (Continuous and discrete)

$$L(y) = \sum_i (\hat{y}_i - y_i)^2$$

Asymmetric binary loss（不对称）

$$L(y,t) = \begin{cases} a & y = 0 \text{ and } t = 1 \\ b & y = 1 \text{ and } t = 0 \\ 0 & y = t \end{cases}$$

2） **Logistics regression**

因变量只有两个 category，假设误差符合 type1 extreme value distribution。取 0 代表 baseline，取 1 代表非 baseline。仍属于 linear，缺陷是界限单一。

3） **Decision tree**

Entropy：判断有序无序

提问是一个熵减的过程，若小于 0.1 可能就会出现 overfitting（问得太细不具备普遍可预测性了），以及没有前瞻性属于 local optimization

4） Random forest

5） Neural network

R 演练语句

Training set、Validation set、Test set

predict（model，验证集）：pred 输出的结果是β*x 的乘积之和

mean（（预测值 – 实际值）^2）：mean square error

mean（abs（预测值 –实际值））：mean absolute error

quantile（，0.5）：四分位中位

glm（，family=binary）：run logistics regression

评判 logistics 优劣以 null deviance 和 residual deviance 的差值，越大越好

rpart 包

rpart（，method=，cp=）：如果是 binary category，method 为“class “，如果是连续型为”anova “，cp 值代表熵减的最小值

rpart.plot 包

attach(mtcars)

par(mfrow=c(1,2)) ：把两张图并列放在一起

prune（，cp=）：修剪决策树

Exercise1

1）数据集排序

[order（ ）,]：排序函数，前面加 - 号代表相反

2）Plot 画图相关

ylim=c（ ， ），xlim=c（ ， ）：横纵轴的刻度

points（ ）：在图中新增点

legend（ ）：添加图例，x=，y=确定位置，pch 代表形状，cex 代表大小，intercept 代表图例之间的距离

Axis（side=，at=c（ ， ， ， ），lables=c（ ， ， ， ）），其中las=2，0垂直或平行

3）数据结果处理

用stargazer的时候，有很多variable时，可以选择性omit=c（ ）一些变量

4）数据集优化、选取

gsub（ ）：替代函数，比如 gsub（“,” ，” ”，数据) 用空白代替逗号

as.numeric（ ）：转化数字

grepl（ ）：判断字符是否存在，（ “判断关键词” ，数据集）返回 TRUE/FALSE

ignore.case = T：忽略大小写

根据判断选取特定的行或列，可以用>,<,=，&，|，及其他判断条件。

str_replace_all（数据， “[^[:alnum:]]” ， “” ）：仅保留数字和字母

5）新增变量

$ + 变量名称 = 赋值

6）查找特定字符的位置

str_locate_all（pattern=， ）

Lecture 1：数据抓取

抓取目的：MOTIVATION
抓取逻辑：单页名称、url > 多页名称、url > 每个子页信息 > 输出数据集

Rvest：网页数据抓取包 — html_nodes()\html_text()
stringr：字符串工具集 — str_sub()
read_html（ ）：读取网页
%>%：分布操作
html_nodes（ ）：读取标签
html_text（ ）：提取文本
str_sub（ ）：选取 string 的一部分
html_attr(href)：读取 URL
paste0(,) ：把多个粘在一起
identical（ ）：判断是否相等
write.csv（ ）：输出数据集
for loop
1、定义空向量 c()
2、for loop
3、抓取数据
4、存入空向量
5、检查向量长度 length（ ）

exercise0
1）install.packages()
2）library()
3）R square、Adjusted R square ：越大越好
4）significance level：显著程度
5）P 值：越小越相关
6）Estimate：越大越相关
7）lm 函数取 log 的意义：规范数据
8）stringAsFactors = F，读取数据集时文字不用转换成 factor
9）head（ ）：显示数据集的前几行
10）创建矩阵：matrix（值，行，列）

11）设置目录：除了 setwd()还有菜单栏的 session

Lecture 0

1）file.choose（导入数据
2）read.table（False 属无标题，直接开数
3）read.csv（True 属有标题，排除首行往下数，修改的话用 header = F
4）summary（输出 6 个结果：min、1/4 位、median、mean、3/4 位、max
5）command + shift + c（变文本为注释
6）names（更改元素名称
7）colnames、rownames（更改列名称，更改行名称
8）$（筛选元素
9）<-（赋值
10）dim（查看变量的维数
11）mean、median、max、min 函数
12）which 函数（找出符合输入条件的数据
13）which.max（找出最大数的位置
14）sl[which.max(sl)] (找出最大数的具体值，这里的 sl 是定义的条件范围
15）iris[行,列] （留空取所有，1:10 代表第 1 至第 10，-10 代表排除第 10，-c(1:10)代表排除第 1 至第 10 也可直接输入行、列的名称，如 iris[, "sepal_width"] 也可嵌入 which 函数，如 iris[which(sl==5,1),]
16）list 函数 list（ ， ， ， ， ）输出的时候用[1]、[2]、[3]、[4]

- exp() 自然对数的幂
- ^ 幂
- -> 与-< 反向赋值
- vec 赋值 1：10，c（1，2，3），rep，seq
- rep rep（数字，重复次数）
- seq seq（from=，to=，by=）或 seq（ ，，length.out=）
- vec 取数 [ ] 中括号，间隔数据需要用 c 连接，意为取"第 n 个数"
- vec 反向取数 如 vec [9:1]

- all.equal　　all.equal(,)检验是否相等
- vec 比较大小　　返回值为单个向量的比较结果
- vec 的 list 函数　　如 list (vec1, vec2, vec3), 取 List [[ ]] 为单个 vec
- 向量内积 %*%、crossprod(,)：a1b1+a2b2+a3b3
- 向量外积 %o%、tcrossprod(,)、outer(,)
- cbind、rbind：按列合并矩阵，按行合并矩阵

- matrix：创建矩阵 matrix (x : y, nrow=, ncol=, dimnames = list (c("行 1","行 2"),c("列 1","列 2"))。也可以直接用 (x: y, nrow=)

- plot 函数：plot(x=x 轴数据,y=y 轴数据,main="标题",sub="子标题"，type="线型",xlab="x 轴名称"，ylab="y 轴名称"，xlim = c(x 轴范围，x 轴范围),ylim = c(y 轴范围,y 轴范围))

- plot 函数的 type　："#"点状　　"l"线状　　"b" 点+线(分离)　　"o"点+线(覆盖)　"c" 线 -点　"h"直方线状　　"s/S"阶梯状　　"n"无形状
- 点样式 pch：