

论文结构

1. Analysis of the Problem // Introduction

总结问题 (自己的话重描述) + 对问题的理解
和解决思路

2. Assumption & Nomenclature (Notion)

根据对问题理解的复杂度选择模型
并给出数学符号定义表

3. Data processing & Analysis

数据处理方法, 对处理后的数据进行描述.

因为这个任务可以额外增加数据, 没有想好这个部分
与假设、定义符号的前后关系

4. Model Construction

根据定义好的符号描述模型;
介绍该模型怎样作用于当前任务

5. Model Results & Analysis (根据需要有是否折分)

跟据模型编写好代码,

在处理好数据上进行测试



得出结果 (表格, 图象)

⇒ } 要什么类型?
 } 图象的种类?

6. Answers for required questions

回答题目给出的具体问题 1-3 (4)

7. Conclusion

8. Report for HK Broker 问题5

9. Appendice

提供的数据库 (xlsx 形式)

⇒

- ① monohulled sailboats
2346 个样本
- ② catamarans
1146 个样本

2 个类型 ⇒ 每个对应 excel 中的
一张 Tab

表的结构:

在包括以下数据
的前提下可继续
进行填充

Make	Variant	Length (ft)	Geographic Region	Country / Region / State	Listing Price (USD)	Year
str	str	float	str	str	int	str

目标: 解释 listing price

数据清洗什么?

① str 名称转化为数字形式 (才能用数学方法进行训练)

例如类型从 { 数字表示?
独热向量表示?

try

Year 的类型是 str 并不合适

② 归一化处理?

将数值区域限制在 0-1 内

~~③ 错误、不完整的数据. (没仔细看过是否存在这种问题)~~

数据扩充什么?

① 样本数量扩充?

继续找帆船

~~有无更多类型?~~

(当前仅有2大类)

~~能否满足已经xlsx要求?~~

~~下一步该怎么处理?~~

✓ 找相似物品

later

能否满足已经xlsx要求?

— 干脆的话怎么办;

什么样的物品更相似?

— 怎么度量相似度

最终牵涉到帆船上

② 扩充样本特征

↓

beam, draft, displacement, rigging, sail area, hull material, engine hours, sleeping capacity, headroom, electronics 船舶的特征 (物理)

影响性能、舒适性

区域地区的经济特征

关于世界各国的经济数据与人口数据

并不是越多越好, 我们应当为所应增加特征进行筛选

解决方法用什么?

个人认为是一个典型的机器学习问题

↓
回归问题

数据特征 \Rightarrow 离散 + 连续

地区
类型
型号

价格 长度
年份?

1. 线性回归模型 \Rightarrow 离散值进行 ^{one-hot} 哑变量 编码

2. 决策树回归

随机森林回归

4. 梯度提升回归

5. XG Boost 回归

6. 支持向量机 回归

帆船种类 (根据船体数量):

① Monohull

② Catamaran

⇒ 主要产品

还有 trimarans (3 hulls) 和 multi-hulls

当前问题:

1. 数据问题

题目要求研究“二手帆船的价格”

- 提供数据完全没有限制是否是二手的标记

- 二手价格的预测需要历史数据?

从哪获取数据? 获取什么数据?

2. 价格预测问题

• 数据的价格衡量单位 \Rightarrow 不同国家地区会受到
当时汇率的影响?
怎么消除这个问题?

• 过去数据预测结果再处理?

\Rightarrow 通货膨胀问题是否需要考虑?

• 住房的最终目的是对香港地区的价格进行预测,
比较强调 地区、船型号 的因素与 价格 的问题.
region variant listing price (USD)

3. 采用什么方法?

• 小词中有些要求去除意义. 统计意义的解释

• 不同问题要求不同

(可解释性要求)

• 方法是否容易实现或参考?

• 方法的建模表述是否能清晰导出?

• 我们的数据是否可以满足训练要求?