房价预测项目总结报告

1 引言

1.1 背景

- 略

1.2 目的

- 学习常见机器学习算法原理、模型
  - 线性回归
- 机器学习的一般流程（pipeline）
  - 包括数据的预处理，填充，探索性数据集分析，特征工程，归一化，用不同的模型训练，调参，找到最优参数，模型融合。

2 理论基础

2.1 数学基础

- 梯度下降
- 牛顿法

2.2 模型简介

2.2.1 线性回归类模型

- 1 朴素线性回归
- 2 基于L1的线性回归
- 3 基于L2的线性回归
- 4 ElasticNet(弹性网络L1与L2)

2.2.2 树回归–CART

- 1 CART
- 2 RF
- 3 AdaBoost
- 4 GBDT—-XGBoost—-lightGBM

2.2.3 SVM类型

- 1 SVR

2.2.4 神经网络

- 1 FC神经网络

2.2.5 集成学习

- 1 Stacking Ensemble

3 项目流程与进展

3.1 项目流程

3.1.1 算法的pipeline

数据采集–>数据清洗–>数据仓库–>数据挖掘–>数据标注–>数据集市(训练集/评测集)–>训练–>评测–>模型工程(int8定点化)–>模型的集成和应用

3.1.2 算法的baseline

训练集的制作–>训练–>评测
在形成一个BaseLine之后，模型迭代的过程进步的标志是评测指标的提升。

3.1.3 实际流程

- 初步认识数据，理解数据，理解业务
- 数据的相关性分析，空值填充
- 数据集的准备，划分
- 各种模型训练，包括Lasso,RidgR,ElasticNet,Xgboost,FC。各种模型调参。
- 选择表现好的模型进行stacking。
- stacking之后就是不断的重复上面的过程，其中特征工程花大量时间，需要尝试不同特征之间的组合，相加、相乘等。
- 经过大量的尝试，大量重复上面步骤，最后得到评测指标较优的模型。

3.2 项目进展

- 基本按照3.1.3中进行。作了如下一些尝试：
  - 去掉一些相关性低的特征列
  - 尝试组合了一些特征
  - 进行过pca降维度
  - robosaler
  - y值以万为单位，以及log后的w。y归一化后训练，反归一化测评。
  - grid与rand search
  - 在相关性分析中，发现疑似噪声点，删除后模型在测试集上表现稍好。
  - 用全连接神经网络训练，发现效果并不是很好。
  - 老师课上的代码下来基本手敲过，包括首先线性模型和树模型。

4 经验和不足

4.1 经验

- 了解机器学习项目的流程。
- 掌握常见的回归模型的原理，了解一些调参技巧。
- 特征工程很重要，有时候做了一些组合特征之后分数有所提高。
- 初步掌握numpy,pandas以及一些机器学习库的使用。

4.2 不足

- 有时候很盲目，不知道怎么去做特征工程才能提高分数
- 对一些知识点掌握得不好，很多知识点能够听懂，大概也知道怎么回事，但是要我自己讲出来，还不行。
- 时间上投入还不够。
- 代码能力不足，老师的代码很漂亮，我要是没有参考，不copy肯定写不出来。
- 还有一些遗漏的不足，暂时没想起来。

5 成果与展望

5.1 成果

- 我认为几乎每个知识点都能写一篇blog或总结，但是投入真的不够，效率也不够高。不过后续会逐渐把各个知识点都写上，抠脚的博客和GitHub请老师批评指正。
- blog:
  - https://www.cnblogs.com/zingp/p/10375691.html
  - https://www.cnblogs.com/zingp/p/10278223.html
  - https://www.cnblogs.com/zingp/p/10511176.html
- github:
  - https://github.com/zingp/kaggle/tree/master/LosAngelesHousePricesForecast

5.2 展望

- 未来希望能完成各个知识点的总结，包括算法的公式推导。
- 希望在特征工程上多做尝试。

6 建议

- 老师讲课很有激情，很nice。技术和讲解都6得飞起！很喜欢黄老师。
- 如果可以，带我们完整刷一个kaggle项目，感受下特征工程中的痛苦和一点点提分那种快感。至少刷到top5%，这样我觉得理解会更深一些，也方便面试。
- 希望公司给老师配一个手写板，提高老师书写公式的效率，相信同学们上课体验会好很多。
- 最后，希望老师画画重点，出点面试题目给我们练练。

参考文献

- 老师讲课的PPT、《统计学习方法》、《机器学习与应用》、《百面机器学习》、《DeepLearning》、还有一些博客，知乎等就不一一列出了。