

实训总结

14331226 彭瑞


关于此次实训的一些技术上方法和技巧，我已在用于展示的文档里给出。在这篇实训总结中，我想谈谈我在数据挖掘上的一些摸索。

在寒假我参加了数学建模的美赛，这学期我有一门数据挖掘的课程，除此以外，我在 Coursera 上完成了 Machine Learning 课程的学习并满分通过。在对有关理论有了一些了解和思考后，我开始了数据挖掘的实战。数据挖掘课程的课内竞赛(Kaggle in Class)是非常好的锻炼机会，经过大量的摸索和尝试，我在两次课内竞赛中均取得了第一名，并积累了大量的数据挖掘实战经验，包括尝试线性模型、神经网络、支持向量机、集成学习、XGBoost 等多种方法。尤其是 XGBoost 的调参，对于不同的数据集，最适宜的参数也可能不同。把若干个结果以适当的方式融合起来是使自己成绩脱颖而出的关键。每天的提交次数限制也十分让人恼火，因此最好要在本地跑出不错的结果，对自己结果有足够信心之后，再去尝试提交。但有时候可能也需要去测试一种方法或模型在榜单上的分数如何，使用一次提交机会也在所难免。


Large-scale classification-S...
a month to go - Top 1%
1st
of 271

Linear Regression-SYSU-20..
8 days to go - Top 1%
1st
of 280


对我而言，这两次课内竞赛毕竟只是进入数据挖掘领域的敲门砖，真实世界的数据很可能比这些经过预处理的数据复杂，因此我开始转向 Kaggle 上的数据挖掘竞赛，从入门级别的竞赛开始做起，仔细研究这些竞赛的 Tutorial 部分的每一个 Notebook 的技巧和原理，在这个过程中，我熟悉了 Python 的一些使用方法，也更加了解了数据挖掘竞赛的“解题思路”。



Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started - 3 years to go - Entered
7,233 teams



House Prices: Advanced Regression Techniques
Predict sales prices and practice feature engineering, RFs, and gradient boosting
Getting Started - 3 years to go - Entered
2,021 teams



Digit Recognizer
Learn computer vision fundamentals with the famous MNIST data
Getting Started - 3 years to go
1,917 teams

这些学习的过程和这次预测房价的比赛是穿插着进行的。在这次比赛中，我使用了一个非常简单的 XGB 模型作为尝试，交上去的结果已经能够达到 0.316 左右了。特征工程和模型融合是非常重要的环节，因为倘若有了现成的数据，很多事都可以借助 XGBoost 来完成，但是特征工程和模型融合不行。Kernels 和 Discussion 部分的代码和讨论给了我非常大的启发。经过不断地学习与改良，截至目前（6 月 27 日）我的最好结果已经达到了 0.310 左右，但是由于越往上提升越困难，我目前仍然处于一个瓶颈阶段，希望能够在接下来的过程中得到更多的提升和启发。

在数据挖掘实践的过程中，我遇到了很多问题，有的问题很快得到了解决，有的问题困扰我的时间较长，在这里我列出数据挖掘竞赛中遇到的主要问题与体会：

1. 我本来倾向于用 MATLAB 进行数据处理，而对 Python 的了解十分有限，但是出于实用、内存等原因，我不得不从实践中逐渐学习 Python 的使用。

2. XGBoost 的安装遇到了很多麻烦，通过查看官方文档并没有有效解决问题，最终我查阅了大量资料，完成了安装。其中 MINGW-W64 的版本问题，使用 pip 安装的问题令我印象深刻。Anaconda 是 Windows 下 Python 开发的不错环境。

3. 即使是同样的代码，在本地运行和 Kernel 运行的结果仍有较大差异，这可能是由于本地与 Kernel 在环境配置上的差异，而 XGBoost 对某些参数差异“异常敏感”。

4. XGBoost 的参数、融合的参数，看起来相当“武断”，因为很多是根据提交后的结果来调整的，这可能造成在 Public Leaderboard 上的过拟合，但是根据往常的经验来看，当 Private Leaderboard 公布后，成绩也许会“变糟”，但是排名一般不会有较大变化。

我希望自己在数据挖掘领域的浓厚兴趣能够转化为对数据的“洞察力”和能够给出切实有效的数据科学解决方案的能力。我希望今后能够继续深化在数据科学领域的学习与研究，这是我理想的未来研究方向。