

Reprehensible
side

2147483647

私信

归档

RSS

搜索

怎样快糙猛的开始搞Kaggle比赛

本文写给想开始搞Kaggle比赛又害怕无从下手的小朋友们。

最近比较多人问我怎么快速成为数据科学家可以挣钱多干活少整天猎头跳跳槽涨一倍。我一般的答案是，没有这好事，但是搞搞Kaggle的比赛有助于快速成为数据科学家，之后挣钱多少看个人。关于Kaggle比赛是什么，限于篇幅关系，请自行谷歌。


我不是专业机器学习的人，但是我见的太多了。对于有一定数理基础的人来说，快速起步搞起来个Kaggle比赛并且获得不错的名次，难度并非难于成为王思聪的官方老婆。这里有三个部分的知识需要强化：

1. 数理基础。基本上高考数学不错的理工科学生，学过了大一大二的数学基础课程（包括微积分、数理统计、数理方程、集合论等），不存在任何问题。如果想测试一下自己，那就看看这个题目：

如果一个妹子喜欢我可能因为我帅或者我有钱，因为我既帅又有钱的概率是0.1，只是因为我有钱的概率是0.5，问，如果妹子喜欢我只是因为我就是帅的概率是多少？

如果能不费力气（心算更好）的解答这个问题，基本上这部分知识是足够了。

2. 机器学习。Kaggle比赛多依靠机器来自动处理，机器学习几乎是必须要的技能。开始搞Kaggle需要的机器学习技能并不深入，只是需要对于机器学习的常见几个方法有基本了解即可，比如说对于一个问题，你可以认识到它是个classification的问题啊还是regression的问题啊，为什么机器可以根据你输入的一个矩阵来算出来分类结果啊。推荐Coursera上Andrew Ng的机器学习课程 <https://www.coursera.org/course/ml> 一个捷径就是，如果你时间紧的话，只要知道什么叫做Supervised learning并且会自己实现一个Logistic Regression，差不多就够了。



Reprehensible s...
phunters.lofter.com

+ 关 注

注册LOFTER

下载LOFTER App



Reprehensible side

2147483647

私信

归档

RSS

搜索

分享

关注

注册LOFTER



Reprehensible s...
phunters.lofter.com

+ 关注

注册LOFTER

下载LOFTER App

3. Coding。限于篇幅只介绍Python。我可没有说什么钦定Python，你支持不支持，我用python我当然支持。基本的python编程得熟练，如果不熟练可以先学习 Learn Python the Hard Way。会了python之后，把scikit-learn的基本教程的classification的部分练练，你会发现在Andrew Ng课上的知识，在python里面实际跑跑简单数据，能对课上的知识深刻的理解。同时，如果有富余时间的话，可以顺道看看numpy和pandas的一些基础操作，这些是用来数据处理好工具。

上面三点对一个数理基础不错的人来说，差不多几周的空余时间就可以了，如果是在校学生可能更快。

开始搞Kaggle的时候，建议选个入门容易的比赛。如何选择，简单来说就选个参赛人多的就好了，基本上认真搞搞结果还不会差呢。如果一个比赛还有自带Tutorial 就更好了。比如我们可以选泰坦尼克号的比赛，根据乘客的信息来判断他是不是可能在沉船中遇难。地址是 <https://www.kaggle.com/c/titanic-gettingStarted>

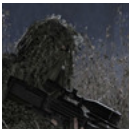
这个比赛有个很好的tutorial，第一次参加比赛的，可以在比赛过程里遇到但是不限于如下的问题：

1. 数据怎么读取
2. 有missing value怎么办
3. 一等舱二等舱之类的feature为什么得当作categorical feature
4.

等等等等之类的问题。这些问题都是在数据科学领域的实际工作每天都能遇到的。最好的学习方法就是针对这个问题，你看Discuss Forum和Tutorial里面教你怎么解决，自己google一下看别人写好的代码怎么解决这些问题。这阶段我建议靠自己的力量搜索答案而不是去论坛上问一些伸手党类的SB问题，即使问了也没关系有人会替你解答的但是这慢嘛。

然后你会开始训练你的模型，又会遇到但是不限于以下的问题

1. 啥叫random forest，咋用，为什么我调了这几个参数不灵呢



Reprehensible side

2147483647

- 私信
- 归档
- RSS
- 搜索

2. 怎么我本地结果很好，但是提交名次掉成狗
3. 原来我要Cross Validation啊（ Andrew Ng的课里说到的那些看起来很无聊的曲线现在知道是为什么了吧 ）
4. 。 。 。 。 。 。

等等之类的。这些问题也是实际工作每天都能遇到的。你就看人家怎么调你就跟着模仿，然后体会思考一下不同调法对结果有什么区别。这比在@七月问答 上面问“如果某某情况我的随机森林的参数该怎么调才能避免这个情况”之类，对问题领悟的更深刻。折磨过几波模型调参，你就差不多知道这些模型的套路是什么了。

然后你开始刷名次，又会遇到但是不限于以下的问题：

1. 怎么CV的结果挺好但是上去还是比不过呢
2. 那谁说用Vowel Wabbit对每个分类做优化怎么搞啊
3. 组合模型这概念我知道，但是实际怎么组合呢
4. 。 。 。 。 。 。

经过这些，你差不多就知道解决一个实际的机器学习问题需要做什么事情了。对的，这就是数据科学家几乎每天的工作，各部分比重不一样，但是理解问题、数据清理、模型调参、评估结果这些循环反复的动作，基本上就是数据科学家需要做的。

在这个摸索挨打的过程中，你可以快速学会数据科学的常用工具（numpy scipy pandas scikit等等），也会在别人的带动下发现新工具（比如@陈天奇怪的xgboost，vowel wabbit之类的），也会学会新技能（比如深度学习以及如何用深度学习去解决实际的问题）。这个学习速度远超过于看看blog，在挨打的过程里，回想一下从可可老师那里看到的每天十条数据科学经验，会不会觉得理解的更深入了呢？

在有实际工业界工作经验之前，搞搞Kaggle比赛几乎是最有效的跨过”数据科学家”门槛的方法。有了实际工作经验，搞搞Kaggle比赛也能扩大视野，也能把前沿研究的第一手结果用到实际问题里。大家加油，跳槽就翻倍的



Reprehensible s...
phunters.lofter.com

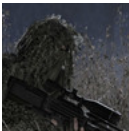
+ 关 注

注册LOFTER

下载LOFTER App

高薪工作指日可待（我没有保证能高薪啊，不要到时候把我拉出来批判番）

分享 关注 注册LOFTER



Reprehensible side

2147483647

2015/04/01 68 7 #kaggle

私信

归档

RSS

搜索

也谈谈新浪微博可能感兴趣 一些找工业界工作的经验总

本文是看过@王小科科科的《聊聊可能感兴趣的人》 个人背景：高能物理PhD，10年多（泪）编程经验，

Reprehensible s... phunters.lofter.com

+ 关注

注册LOFTER

下载LOFTER App

上一篇 下一篇

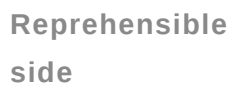
评论(7)

登录/注册LOFTER 开始评论 | 其他方式登录 新浪微博 腾讯QQ

- 932974672：666，笑死了
- 天涯 回复了 劉戀1990-2099：0.6吧。
- 劉戀1990-2099：这道题目的答案是0.2么
- 劉戀1990-2099：如果一个妹子喜欢我可能因为我帅或者我有钱，因为我既帅又有钱的概率是0.1，只是因为我有钱的概率是0.5，问，如果妹子喜欢我只是因为我就是帅的概率是多少？
- n0thing233：to be a data scientist in the near future.Thank you for the useful materials.
- The harmonious side 回复了 阿波_ZJU：刚刚才看到留言，top 5%很厉害了，我来看看最近有么有时间再搞一下比赛，你懂的已婚人士时间不是很多啊，得陪媳妇
- 阿波_ZJU：我刚开始玩kaggle，有没有兴趣交流。最近的成绩是在预测bike出租数量的比赛中排top5%，不过结果还没出来，最终成绩还可能会变化。
<https://www.kaggle.com/c/bike-sharing-demand/leaderboard>，搜haibo wu。

热度(68)

- jlusdjava 推荐了此文字
- 能见度 I 很喜欢此文字
- m18751206665 从 Reprehensible side 转载了此文字
- m18751206665 很喜欢此文字

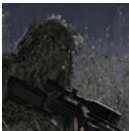


搜索

注册LOFTER

 sl_yuliliao 很喜欢此文字

下载LOFTER App



Reprehensible side

2147483647

私信

归档

RSS

搜索

 me_selfish 很喜欢此文字

 小土豆 很喜欢此文字

 ~~莹~~ 很喜欢此文字

 wing 很喜欢此文字

 Tearbaby 很喜欢此文字

 yes、i do· 很喜欢此文字

 z.]x- 很喜欢此文字

 Minzc 从 Reprehensible side 转载了此文字

 Minzc 很喜欢此文字

 angela297 很喜欢此文字

 jiajiavsjaxi 很喜欢此文字

 臭鱼 很喜欢此文字

 、Chi 很喜欢此文字

 仲夏夜之星 很喜欢此文字

 hmxxlsy 很喜欢此文字

分享

关注

注册LOFTER



Reprehensible s...

phunters.lofter.com

+ 关 注

注册LOFTER

下载LOFTER App

查看更多