开通VIP

聊聊买房那些事儿?

snowsmile1211 | 微信 | 设置 | 消息 | 提醒 | 签到领奖! | 新手上路 | 退出

开始答题 积分: 156 | 用户组: 中级农民-加分请看右边栏-多参与|分享|记录|反馈

请输入搜索内容 帖子 热搜: 美国找工作 定位评估 申请总结 绿卡移民

论坛 学位+学习 机器学习 机器学习`侠`练成记录 Becoming a Machine Learning Pra ...

最近看过此主题的会员

buaalsy2 春山寒 yufanllll

中国数据智能独角兽企业 坐标杭州 | 个推诚聘 数据/算法/分析/研发等岗位

码农求职神器Triplebyte 不用海投 内推多家公司面试

科技公司如何 用数据分析驱动产品开发 $366 off coupon code: best

深入浅出AB Test 从入门到精通 $366 off coupon code: best

E轮2.5亿美元融资 一起作业诚聘 机器学习/数据/教育等职位

导读 VIP瞬间解锁 网课 应用中心 留学 加入我们 免米搜索 关于

返回列表 1 2 3 1 / 3 页 下一页

查看: 2883 | 回复: 27

机器学习`侠`练成记录 Becoming a Machine Learning Practitioner

[复制链接] | 试试Instant~

K姐 发表于 2018-12-3 06:34:59 | 只看该作者

垄头 电梯直达

本楼: 【顶】 100% (7) 0% (0) 【踩】
全局: 顶 95% (4787) 4% (236) 踩

我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

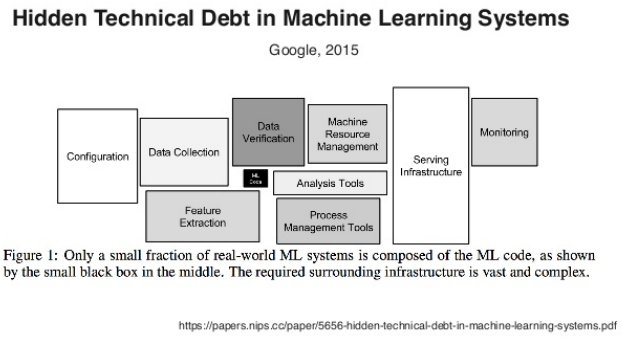
一匹黑马

发消息

分享帖子到朋友圈

Becoming a Machine Learning Practitioner

为什么叫机器学习`侠`，是`调包侠`，`调参侠`的梗



上图来自Google 2015的论文，图片里面很小的黑色矩形是ML算法，其他部分是围绕算法的很多其他成分。

业界做ML，当然有算法的成分，有些公司也有很高深的算法（比如Google），但是工程的成分其实经常会远远，远远的多于算法，在一个激灵想出来的算法能真正落地给业务产生impact之前，有大量的工程方面的设计和考虑。好的大公司也在纷纷构建自家的机器学习平台让ML的投产变得更容易。

但是这些需求在学校ML课程里面很少被提及，给人的印象就是，做ML就是"调包，调参"。但是实际工作里面调包和调参的时间比例可能也就10-20%...

为什么发这篇帖子，也是来自于之前一直在DS领域总结学习，希望持续分享自己学习的历程，求讨论

Why this post

It's an opinionated list of core skills that I found useful in the daily work of a machine learning practitioner, in the tech industry. I am not a researcher and am not interested in becoming one, so the list does not go in depth into any active research domain.

需要强调的是，我不是做ML科研的，也不想做科研，所以本文就是于业界做ML的大家交流讨论，并不打算深入任何科研领域，请做科研的大牛轻拍。

I'll be maintaining this list just as I maintained the learning path for Data Science in the past here https://www.1point3acres.com/bbs/forum.php?mod=viewthread&tid=76429&extra=&page=1.
It's not meant to be exhaustive, and we probably don't need to know them all.
另外这篇内容也并不会多完备。机器学习领域大牛很多，领域很广，应用更是广阔到难以尽数，所以不求尽善尽美，只是抛砖引玉，求交流学习。
Suggestions/discussions welcome.

本文适合什么人

0. Who is this for

This is a practitioner's approach. Researchers: This post is not for you. 不是给ML方向做研究的人；牛人路过就好，不喜勿进。

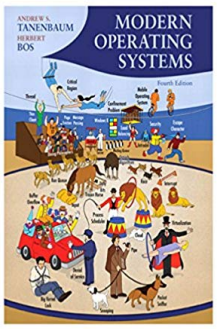
- **Data analyst**: you don't need this. Read this instead: https://www.1point3acres.com/bbs ... 76429&extra=&page=1
- **Data scientist** who's more junior in modeling or focuses on causal inference: 3, some of 4. This list will give more resources: https://www.1point3acres.com/bbs ... 76429&extra=&page=1
- **Machine learning scientist**: 2, 3, expert knowledge of 4,
- **Machine learning engineer:** 1, 2, some 3, good knowledge of 4, 5, some knowledge of 6
- **Machine learning systems engineer:** 1, 2, maybe 3, some knowledge of 4, expert knowledge of 5 and 6

1. CS Fundamentals 计算机基础

Introduction to Python or Introduction to Java
Data structure (地里关于Berkeley 61B的板块 or Coursera specialization)
Design and analysis of algorithms（ Coursera Part 1, Coursera Part 2)

Database (Stanford archived DB course or Using Database with Python)
Discrete math (Coursera Discrete Math specialization)

Operating System (Book: Modern Operating Systems)



2. Programming languages 编程语言

为什么与上一节计算机基础分开来说：因为老是遇到同学说我会numpy啊为什么你说我基础不行。。。。实际上是可以在不太懂基础的情况下楞刷题，或者写些基本能用的代码的。但是稍微深一点的地方会感觉基础知识的缺乏会让人难以在ML道路上上升到一定高度。

Must have :Python, Java
必须一门Python（因为ML好多framework就是python)，和一门compiled language.

仅仅只会python，作为scientist是够了，作为engineer就会有显著差距。

Good to have: C++ 不一定是必须的。但是如果做的工作对速度有比较强的要求，那还是需要会C++的

Will likely encounter at work: scala/go/js 这些看每家tech stack不一样，会有不同涉及。遇到了再学即可。

Will likely encounter in academic settings: Matlab/Octave, R, Julia 这些是学校里面用的多，我至今（2018年底）没看见公司里面用julia的。

3. Math Fundamentals 数学和统计基础

Linear algebra 线性代数，必须的
Calculus 这个大家应该都上过
Optimization 优化：必须的
Statistics 基础统计（不是概率论 which is also good to have，这里是特指统计）

(real analysis and functional analysis might be useful, but is not required)
有空的话学实分析和泛函也可以，但是不是必须的

4. Machine Learning, from intro to advanced
这部分稍微区分一下从入门到进阶

4a. Intro 入门

- 这门课可能没有人不知道了 Introductiont to Machine Learning by Andrew Ng
- 这本书是ESL的简单版，作为直觉培养和思路练成，仍然是不错的，但是那里面的编程就很轻很轻了，真的只够本科生用 Very light but still a good book: an introduction to statistical learning

- 深度学习，也是Ng这课来入门还是不错的
Introudction to Deep Learning by Andrew Ng
- 另外个人比较喜欢Udacity的第一门旗舰课程讲AI的，基于斯坦福的一门本科生课程。会稍微设计一点比前几门入门课更宽广的概念，虽然很浅但是对了解domain很有好处。
Introduction to Artificial Intelligence by Udacity

4b. Advanced 进阶

- 前面几门主要还是supervised learning，下面这门稍微宽广一点，并不完全是ML，但是也是因此感觉对知识面和*落地*有帮助
Data mining & other topics: Mining Massive Data Sets http://web.stanford.edu/class/cs246/

- 这门课可惜没有录像，关于实战的部分讲的还是不错的，而且是其他课程都没有涉及到，但是工作里面的确需要的部分
Cornel course (slides only) adv ml http://www.cs.cornell.edu/courses/cs6780/2010fa/lectures.html

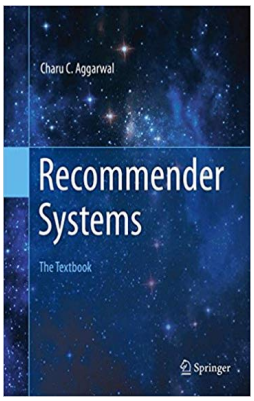
- Book:
经典ESL 不必多说，统计角度 Elements of Statistical Learning

经典不必多说，CS角度 Pattern Recognition and Machine Learning

- 下面看几个应用大方向

- Info retrieval & search engine 信息提取和搜索
Some intro here:
https://www.searchtechnologies.com/blog/relevancy-ranking-301
http://web.stanford.edu/class/cs276/
UIUC course slides https://github.com/SSQ/Coursera-UIUC-Text-Retrieval-and-Search-Engines-Lecture-Slides

- Recommender systems 推荐系统
Book: Recommender Systems: The Textbook by Charu Aggarwal



- Image: 图像识别，如今主要就是**CNN**了
Andrej版的CS231n堪称经典 stanford CS231n Convolutional Neural Networks for Visual Recognition

- NLP 自然语言处理
CS224n: Natural Language Processing with Deep Learning
others to be added
NLU 暂时不知道哪里有比较好的课

- Reinforcement Learning, Deep Reinforcement Learning 加强学习
(book and course TBD)

- Lots more stuff in DL here

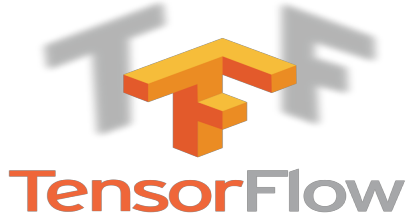
经典课本（但是我觉得读起来还是蛮晦涩。。。不知道我是不是一个人）Deep learning book by Ian Goodfellow and Yoshua Bengio

之前学DL的时候的一些收集 看这里

4c. Frameworks

非深度学习，最常用的肯定就是 General ML framework: sci-kit-learn

深度学习的目前很多了 DL: Tensorflow(Keras), Caffe, Pytorch(Caffe2)



TF的看狗家自己的内容，或者Ng那个课；Pytorch的看fast.ai
Up & coming 或者已经下去了的: theano, MXnet, dl4j

5. Scaling considerations: Big data, distributed systems etc 数据量大了面临的问题

做小数据ML（笔记本上跑跑regression or classification，产生个报告给别人看）严格来说算不上ML，其实主要只能算是modeling（统计建模）
high dimensional data 是另外一个故事，这里先按下不提。

只说业界，面试会考系统设计的地方，需要用到的机器学习系统：

从最最小白的地方看起：（非科班同学不妨看看，科班的可以绕过）
This blog post: thorough intro to distributed systems
https://hackernoon.com/a-thorough-introduction-to-distributed-systems-3b91562c9b3c
And this
https://www.youtube.com/watch?v=BkSdD5VtyRM

System Design: 虽然这个是准备面试用的，但是作为大致入门也是差不多了
Grokking system design interview (for brushing up fundamentals and case studies)

DDIA 堪称经典 Book Designing Data Intensive Applications : for an in-depth look, refer back to fundamental knowledge in OS

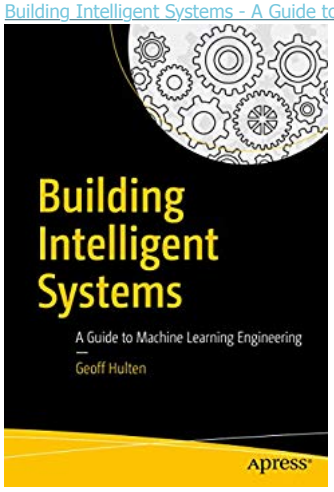
Distributed OLTP and OLAP

https://www.youtube.com/watch?v=tcyDgOejU5g
https://www.youtube.com/watch?v=AoK8_QEi5U0

6. ML Systems & Platforms 机器学习系统和平台

- 这门课还是可惜没有录像只有slides
Adv ML systems (Cornell, slides only)
http://www.cs.cornell.edu/courses/cs6787/2017fa/

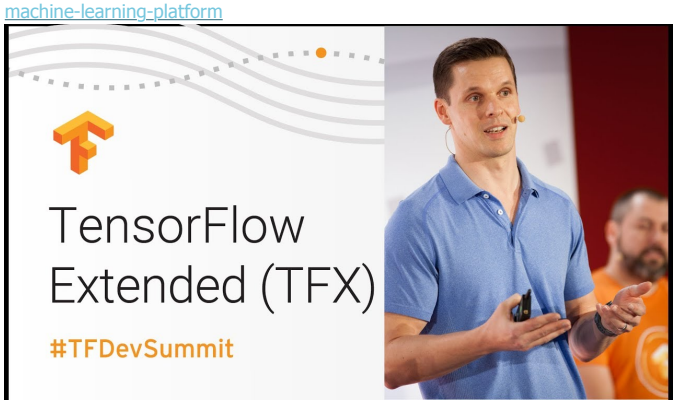
- Book: 这本书我也只有翻过，还不知道到底多好
Building Intelligent Systems - A Guide to Machine Learning Engineering



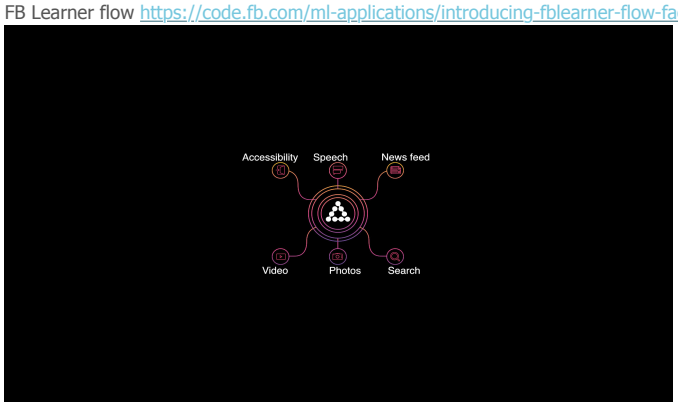
- NIPS2018: to find some talks http://learningsys.org/nips18/schedule.html

ML Systems 什么是机器学习系统
- Prod ML, paper 1 (tech debt) https://ai.google/research/pubs/pub43146,
- paper 2 (test score), https://ai.google/research/pubs/pub46555
- tbd

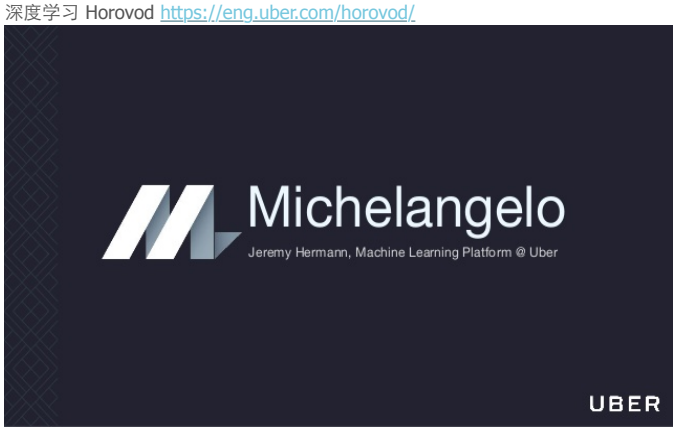
ML Platforms 机器学习平台
- Google:
TFX https://www.tensorflow.org/tfx/ & KDD talk https://www.kdd.org/kdd2017/papers/view/tfx-a-tensorflow-based-production-scale-machine-learning-platform



- Facebook:
FB Learner flow https://code.fb.com/ml-applications/introducing-fblearner-flow-facebook-s-ai-backbone/



- Uber:
Michelangelo https://eng.uber.com/scaling-michelangelo/
深度学习 Horovod https://eng.uber.com/horovod/



- linkedin:
Pro ML https://twimlai.com/twiml-talk-200-productive-machine-learning-at-linkedin-with-bee-chung-chen/

- Airbnb:
Bighead https://databricks.com/session/bighead-airbnbs-end-to-end-machine-learning-platform

居然缺logo

另外这个podcast很不错，建议不要错过
Podcast ***TwimlAI*** is featuring a lot of these systems lately, a fantastic listen
https://twimlai.com/shows/

其他还有 Amazon, Netflix etc, ...

机器学习

本主题由 K姐 于 2018-12-16 07:51 设置高亮

评分

参与人数 32	大米 +110	理由	收起
鲁雪松	+ 5	很有用的信息!	
jincute	+ 1	赞一个	
jason0123lin	+ 3	很有用的信息!	
Skinnycook	+ 3	很有用的信息!	
SuperGuy10	+ 3	很有用的信息!	
czhrabbit	+ 3	很有用的信息!	
johnson139	+ 3	给你点个赞!	
femton	+ 2	给你点个赞!	
pzqkent	+ 3	给你点个赞!	
mk48	+ 3	给你点个赞!	
JCY	+ 3	很有用的信息!	
Steinhafen	+ 3	很有用的信息!	
ikazu	+ 3	很有用的信息!	
hellohuijia	+ 3	点赞!	
shiruizhi	+ 3	给你点个赞!	

查看全部评分

相关帖子

深度学习主机推荐	[研究向]哪些学校在搞贝叶斯优化/Meta Learning/神经架构搜索?
只读研究生申machine learning人工智能相关工作比较尴尬	R新手求助！关于二项分布随机数
想要通过自学和交流提升自己！ML自律小组	Dimitri P. Bertsekas --- Reinforcement Learning 2019 教材，幻灯片
Get Cloud Server for less than $1 per year	诚咨Facebook research intern coding test
求问怎么在Google Drive上解压6GBzip文件	机器学习-网站paperwithcode
[转载] Setting up deep learning environment the easy way on	finance本is硕转方向选择求指导——sde还是big data/machine
经典教材! 一起学习! Deep Learning By Ian Goodfellow etc	请问现在真的是机器学习比较好吗
请教如果可以拿到facebook AI方向的research intern offer	optimization online class求推荐
机器学习-读书看文献记录贴	Data Scientist 转行做 ML Engineer, Ask Me Anything
[转载] 牛人林达华推荐有关机器学习的数学书籍	刷Andrew Ng deeplearning.ai系列五节课求课友一起打卡

本帖被以下淘专辑推荐:

· JOB | 主题: 22, 订阅: 0

收藏 141 评分 分享 1 淘帖 1

一亩三分地管理员 + 所有产品线 PM
玩转Data Science
ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复 微信 使用道具 举报

楼主 | K姐 发表于 2019-1-9 09:15:13 | 只看该作者 来自 9垅

本楼: 【顶】 0% (0) 0% (0) 【踩】
全局: 顶 95% (4787) 4% (236) 踩

MIT，和无人车的内容统一更新在这一楼

good stuff here:
https://deeplearning.mit.edu/
lots of practitioner's talk on actual, industry scale systems that's hard to find elsewhere

无人车
Part of MIT course, self driving car (slightly out of date but still good stuff)https://ocw.mit.edu/resources/re ... 5/unit-8.-robotics/

Here's a 2017 talk by Cruise,
https://www.youtube.com/watch?v=s-8cYj_eh8E

百度 Apollohttps://www.youtube.com/watch?v=jiZhSIrmODk

2019 MIT 的最新总结
https://www.youtube.com/watch?v=53YvP6gdD7U

一亩三分地管理员 + 所有产品线 PM
玩转Data Science
ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复 评分 举报

moernongdu 发表于 2019-1-7 05:07:34 | 只看该作者 推荐

本楼: 【顶】 100% (1) 0% (0) 【踩】
全局: 顶 93% (15) 6% (1) 踩

关于 - Reinforcement Learning, Deep Reinforcement Learning 加强学习
(book and course TBD)
我贡献一个素材 Deepmind （就是搞alpha狗）的课程
https://www.youtube.com/playlist?list=PLqYmG7hTraZDNJre23vqCGIVpfZ_K2RZs

回复 评分 举报

llj20dd 发表于 2019-1-1 02:12:57 | 只看该作者 推荐

本楼: 【顶】 100% (1) 0% (0) 【踩】
全局: 顶 100% (2) 0% (0) 踩

十分感谢，正准备从DA转到偏ML方向！共同进步！

回复 评分 举报

我的人缘 8
892 主题 | 1万 帖子 | 7万 积分
一匹黑马
发消息

我的人缘 0
7 主题 | 139 帖子 | 263 积分
发消息

我的人缘 0
0 主题 | 4 帖子 | 108 积分
发消息

我的人缘 0

0 主题 | 102 帖子 | 213 积分

发消息

yangyijane 发表于 2018-12-13 02:35:55 | 只看该作者 沙发

本楼: 【顶】 0% (0) 0% (0) 【踩】
全局: 顶 100% (1) 0% (0) 踩

很喜欢会一直跟下去的。

回复 评分 举报

我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马

发消息

楼主 | K姐 发表于 2018-12-16 09:08:11 | 只看该作者 板凳

本楼: 【顶】 0% (0) 0% (0) 【踩】
全局: 顶 95% (4787) 4% (236) 踩

顺手的工具们
Docker https://www.docker.com/
Kubernetes https://kubernetes.io/ Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications. 目前（end 2018）看起来是同类产品里面完全胜出的那个。
目前(end of 2018) AWS和GCP里面用一种就差不多了。AWS 市场占有比例更高

评分

参与人数 1	大米 +5	理由	收起
DL	+ 5	很有用的信息！	

查看全部评分

一亩三分地管理员 + 所有产品线 PM
玩转Data Science
ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复 评分 举报

我的人缘 0

10 主题 | 241 帖子 | 360 积分

发消息

happydreamer 发表于 2018-12-19 00:27:49 | 只看该作者 地板

本楼: 【顶】 0% (0) 0% (0) 【踩】
全局: 顶 96% (29) 3% (1) 踩

Great, thanks!

回复 评分 举报

sunson55 发表于 2018-12-19 11:09:51 | 只看该作者 地下室

本楼: 【顶】 0% (0) 0% (0) 【踩】
全局: 顶 100% (1) 0% (0) 踩

好详细实用，多谢楼主。

我的人缘 0

0 主题 | 59 帖子 | 333 积分

发消息

回复 评分 举报

楼主 | K姐 发表于 2018-12-19 15:24:35 | 只看该作者 下水道

本楼: 【顶】 0% (0) 0% (0) 【踩】
全局: 顶 95% (4787) 4% (236) 踩

Kubeflow https://www.infoworld.com/article/3301549/kubernetes/kubeflow-brings-kubernetes-to-machine-learning-workloads.html

actually quite a lot of other talks at this SciPy 2018 conf is very good, especially good for a general overview on the kind of problem out there, and showing at least one of the solutions out there.

我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马

发消息

一亩三分地管理员 + 所有产品线 PM
玩转Data Science
ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复 评分 举报

还有一些的帖子被系统自动隐藏，点此展开

下 一 页 »

返回列表 1 2 3 1 / 3 页 下一页

snowsmile1211
快速发帖

| 上传 高级模式

发表回复 本版积分规则

提醒：发帖可以选择内容隐藏，部分板块支持匿名发帖。请认真读完以下全部说明：

■隐藏内容方法 - 不要多加空格： [hide=200]你想要隐藏的内容比如面经[/hide]
■意思是：用户积分低于200则看不到被隐藏的内容
■可以自行设置积分值，不建议太高（200以上太多人看不到），也不建议太低（那就没必要隐藏了）
■建议只隐藏关键内容，比如具体的面试题目、涉及隐私的信息，大部分内容没必要隐藏。
■微信/QQ/电子邮件等，为防止将来被骚扰甚至人肉，以论坛私信方式发给对方最安全。
■匿名发帖的板块和方法：http://www.1point3acres.com/bbs/thread-405991-1-1.html

回帖后跳转到最后一页

Powered by **Discuz!** X3
© 2001-2013 Comsenz Inc. Design By HUXTeam