机器学习

机器学习该怎么入门？

本人大学本科，对机器学习很感兴趣，想从事这方面的研究。在网上看到机器学习有一些经典书如Bishop的PRML， Tom Mitchell的machine learning，还有pattern classification，不知该如何入门？那本书比较容易理解？

3 条评论 分享

49 个回答 按投票排序

417

张松阳，**syzhang.me**

紧箍当头、Charm Young、知乎用户 等人赞同

好东西不敢独享，转载一发。

正在学习林轩田的机器学习基石和吴恩达的机器学习，感觉讲的还不错，数学基础还是蛮重要的。

> 机器学习入门资源不完全汇总
> 感谢贡献者： tang_Kaka_back@新浪微博
> 欢迎补充指正，转载请保留原作者和原文链接。 本文是 机器学习日报 的一个专题合集，欢迎订阅：请给hao@memect.com 发邮件，标题＂订阅机器学习日报＂。
> 机器学习入门资源不完全汇总 基本概念
> 机器学习 机器学习是近20多年兴起的一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法。
> 下面从微观到宏观试着梳理一下机器学习的范畴：一个具体的算法，领域进一步细分，实战应用场景，与其他领域的关系。
> 图1: 机器学习的例子：NLTK监督学习的工作流程图 (source: nltk.org/book/ch06.html)
> 图2: 机器学习概要图 by Yaser Abu-Mostafa (Caltech) (source: Map of Machine Learning (Abu-Mostafa))
> 图3: 机器学习实战：在python scikit learn 中选择机器学习算法 by Nishant Chandra (source: In pursuit of happiness!: Picking the right Machine Learning Algorithm)
> 图4: 机器学习和其他学科的关系： 数据科学的地铁图 by Swami Chandrasekaran (source: Becoming a Data Scientist)

机器学习入门资源不完全汇总　入门攻略
大致分三类： 起步体悟，实战笔记，行家导读

- 机器学习入门者学习指南 @果壳网　(2013) 作者 白马　-- [起步体悟] 研究生型入门者的亲身经历
- 有没有做机器学习的哥们？能否介绍一下是如何起步的 @ourcoders　-- [起步体悟] 研究生型入门者的亲身经历，尤其要看reyoung　的建议
- tornadomeet 机器学习 笔记　(2013) -- [实战笔记] 学霸的学习笔记，看看小伙伴是怎样一步一步地掌握“机器学习”
- Machine Learning Roadmap: Your Self-Study Guide to Machine Learning　(2014) Jason Brownlee -- [行家导读] 虽然是英文版，但非常容易读懂。对Beginner,Novice,Intermediate,Advanced读者都有覆盖。
    - A Tour of Machine Learning Algorithms　（2013） 这篇关于机器学习算法分类的文章也非常好
    - Best Machine Learning Resources for Getting Started　（2013） 这片有中文翻译 机器学习的最佳入门学习资源 @伯乐在线　译者 programmer_lin
- 门主的几个建议
    - 既要有数学基础，也要编程实践
    - 别怕英文版，你不懂的大多是专业名词，将来不论写文章还是读文档都是英文为主
    - [我是小广告][我是小广告]订阅机器学习日报，跟踪业内热点资料。

机器学习入门资源不完全汇总　更多攻略

- 机器学习该怎么入门 @知乎 (2014)
- What's the easiest way to learn machine learning @quora　(2013)
- What is the best way to study machine learning @quora　(2012)
- Is there any roadmap for learning Machine Learning (ML) and its related courses at CMU Is there any roadmap for learning Machine Learning (ML) and its related courses at CMU　(2014)

机器学习入门资源不完全汇总　课程资源
Tom Mitchell 和 Andrew Ng 的课都很适合入门
机器学习入门资源不完全汇总　入门课程机器学习入门资源不完全汇总　2011 Tom Mitchell(CMU)机器学习英文原版视频与课件PDF　他的《机器学习》在很多课程上被选做教材，有中文版。

- Decision Trees
- Probability and Estimation
- Naive Bayes
- Logistic Regression
- Linear Regression
- Practical Issues: Feature selection，Overfitting ...
- Graphical models: Bayes networks, EM，Mixture of Gaussians clustering ...
- Computational Learning Theory: PAC Learning, Mistake bounds ...

- Semi-Supervised Learning
- Hidden Markov Models
- Neural Networks
- Learning Representations: PCA, Deep belief networks, ICA, CCA ...
- Kernel Methods and SVM
- Active Learning
- Reinforcement Learning 以上为课程标题节选

机器学习入门资源不完全汇总　2014 Andrew Ng (Stanford)机器学习
英文原版视频　这就是针对自学而设计的，免费还有修课认证。“老师讲的是深入浅出，不用太担心数学方面的东西。而且作业也非常适合入门者，都是设计好的程序框架，有作业指南，根据作业指南填写该完成的部分就行。”（参见白马同学的入门攻略）"推荐报名，跟着上课，做课后习题和期末考试。(因为只看不干，啥都学不会)。" (参见reyoung的建议）

1. Introduction (Week 1)
2. Linear Regression with One Variable (Week 1)
3. Linear Algebra Review (Week 1, Optional)
4. Linear Regression with Multiple Variables (Week 2)
5. Octave Tutorial (Week 2)
6. Logistic Regression (Week 3)
7. Regularization (Week 3)
8. Neural Networks: Representation (Week 4)
9. Neural Networks: Learning (Week 5)
10. Advice for Applying Machine Learning (Week 6)
11. Machine Learning System Design (Week 6)
12. Support Vector Machines (Week 7)
13. Clustering (Week 8)
14. Dimensionality Reduction (Week 8)
15. Anomaly Detection (Week 9)
16. Recommender Systems (Week 9)
17. Large Scale Machine Learning (Week 10)
18. Application Example: Photo OCR
19. Conclusion

机器学习入门资源不完全汇总　进阶课程
2013年Yaser Abu-Mostafa (Caltech) Learning from Data -- 内容更适合进阶 课程视频,课件PDF@Caltech

1. The Learning Problem
2. Is Learning Feasible?
3. The Linear Model I
4. Error and Noise
5. Training versus Testing
6. Theory of Generalization
7. The VC Dimension
8. Bias-Variance Tradeoff
9. The Linear Model II
10. Neural Networks
11. Overfitting
12. Regularization
13. Validation
14. Support Vector Machines
15. Kernel Methods
16. Radial Basis Functions
17. Three Learning Principles
18. Epilogue

2014年 林軒田**(**国立台湾大学**)** 機器學習基石 **(Machine Learning Foundations)** -- 内容更适合进阶，華文的教學講解 课程主页
When Can Machines Learn? [何時可以使用機器學習] The Learning Problem [機器學習問題] -- Learning to Answer Yes/No [二元分類] -- Types of Learning [各式機器學習問題] -- Feasibility of Learning [機器學習的可行性]
Why Can Machines Learn? [為什麼機器可以學習] -- Training versus Testing [訓練與測試] -- Theory of Generalization [舉一反三的一般化理論] -- The VC Dimension [VC 維度] -- Noise and Error [雜訊一錯誤]
How Can Machines Learn? [機器可以怎麼樣學習] -- Linear Regression [線性迴歸] -- Linear `Soft' Classification [軟性的線性分類] -- Linear Classification beyond Yes/No [二元分類以外的分類問題] -- Nonlinear Transformation [非線性轉換]
How Can Machines Learn Better? [機器可以怎麼樣學得更好] -- Hazard of Overfitting [過度訓練的危險] -- Preventing Overfitting I: Regularization [避免過度訓練一：控制調適] -- Preventing Overfitting II: Validation [避免過度訓練二：自我檢測] -- Three Learning Principles [三個機器學習的重要原則]
机器学习入门资源不完全汇总　更多选择
2008年**Andrew Ng CS229** 机器学习 -- 这组视频有些年头了，主讲人这两年也高大上了.当然基本方法没有太大变化，所以课件PDF可下载是优点。 中文字幕视频@网易公开课 | 英文版视频@youtube |课件PDF@Stanford
第1集.机器学习的动机与应用 第2集.监督学习应用.梯度下降 第3集.欠拟合与过拟合的概念 第4集.牛顿方法 第5集.生成学习算法 第6集.朴素贝叶斯算法 第7集.最优间隔分类器问题 第8集.顺序最小优化算法 第9集.经验风

险最小化 第10集.特征选择 第11集.贝叶斯统计正则化 第12集.K-means算法 第13集.高斯混合模型 第14集.主成分分析法 第15集.奇异值分解 第16集.马尔可夫决策过程 第17集.离散与维数灾难 第18集.线性二次型调节控制 第19集.微分动态规划 第20集.策略搜索

2012年余凯**(**百度**)**张潼**(Rutgers)** 机器学习公开课 -- 内容更适合进阶 课程主页@百度文库 | 课件PDF@龙星计划

第1节Introduction to ML and review of linear algebra, probability, statistics (kai) 第2节linear model (tong) 第3节overfitting and regularization(tong) 第4节linear classification (kai) 第5节basis expansion and kernelmethods (kai) 第6节model selection and evaluation(kai) 第7节model combination (tong) 第8节boosting and bagging (tong) 第9节overview of learning theory(tong) 第10节optimization in machinelearning (tong) 第11节online learning (tong) 第12节sparsity models (tong) 第13节introduction to graphicalmodels (kai) 第14节structured learning (kai) 第15节feature learning and deeplearning (kai) 第16节transfer learning and semi supervised learning (kai) 第17节matrix factorization and recommendations (kai) 第18节learning on images (kai) 第19节learning on the web (tong)

机器学习入门资源不完全汇总 论坛网站机器学习入门资源不完全汇总 中文

我爱机器学习 我爱机器学习

mitbbs.com/bbsdoc/DataS... MITBBS— 电脑网络 - 数据科学版

机器学习小组 果壳 > 机器学习小组

cos.name/cn/forum/22 统计之都 » 统计学世界 » 数据挖掘和机器学习

北邮人论坛-北邮人的温馨家园 北邮人论坛 >> 学术科技 >> 机器学习与数据挖掘

机器学习入门资源不完全汇总 英文

josephmisiti/awesome-machine-learning · GitHub 机器学习资源大全

Machine Learning Video Library Caltech 机器学习视频教程库，每个课题一个视频

Analytics, Data Mining, and Data Science 数据挖掘名站

datasciencecentral.com/ 数据科学中心网站

机器学习入门资源不完全汇总 东拉西扯

一些好东西，入门前未必看得懂，要等学有小成时再看才能体会。

机器学习与数据挖掘的区别

- 机器学习关注从训练数据中学到已知属性进行预测
- 数据挖掘侧重从数据中发现未知属性

Dan Levin, What is the difference between statistics, machine learning, AI and data mining?

- If there are up to 3 variables, it is statistics.
- If the problem is NP-complete, it is machine learning.
- If the problem is PSPACE-complete, it is AI.
- If you don't know what is PSPACE-complete, it is data mining.

几篇高屋建瓴的机器学习领域概论, 参见原文

- The Discipline of Machine Learning Tom Mitchell 当年为在CMU建立机器学习系给校长写的东西。
- A Few Useful Things to Know about Machine Learning Pedro Domingos教授的大道理，也许入门时很多概念还不明白，上完公开课后一定要再读一遍。

> 几本好书
>
> - 李航博士的《统计学习方法》一书前段也推荐过，给个豆瓣的链接

发布于 2015-07-05　13 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

353

阿猫**Knight**，Perfekt

知乎用户、Jason、温融冰 等人赞同

我也谈谈自己的经验。

机器学习说简单就简单，说难就难，但如果一个人不够聪明的话，他大概很难知道机器学习哪里难。基本上要学习机器学习，先修课程是**algebra, calculus, probability theory, linear regression**。这几门科学好了再学**Machine learning**是事半功倍的。此外近代数学的东西也要懂， **functional analysis**啥的。其实不懂也行，只是现在文献总是喜欢引用里面的概念，懂一些读起来方便。（我就很讨厌**manifold learning**这个名字，把许多人都吓跑了）**real analysis**最好用心学，对序列或函数的收敛性的理解很能帮助你了解这些模型的精髓。**Optimization theory (ref. Convex optimization by Boyd)**也是重中之重，在前面几门课学好并有一定**python**基础的时候可以仔细读一读。

其实机器学习需要看的书不多，必读的是**elements of statistical learning**。这本书涵盖范围很广，且深入浅出，习题也有一定难度，适合自学。你看过这本之后就知道其他什么书可以看什么书不需要看了。

再下来就是练习，这个是重中之重。我觉得做**kaggle**的比赛最有效。可以仿照别人写写**code**，也可以自己想想办法，但最主要的是要能够迅速完成编程并给出结果。我见过许多人光讨论就可以几天，但真正动起手来就萎了。

最后就是读**source code**并自己实现几个**model from scratch**。这个比较难，但是确是最锻炼人的。具体语言应该是越基础越好，比如**C/C++**什么的。等你自己写完了一两个**model**，再去用别人的**package**就会觉得得心应手许多了。我真心觉得这个比上**coursera**那些课强多了。上**coursera**最大的缺点就是容易变得似懂非懂纸上谈兵。我自己**program**过**ensemble trees(C++)**和**deep learning solver(Python)**，受益颇多。至于读**source code**，我觉得**libsvm**写得很好啊，不过算法对大一大二新生是难了点。此外，基于**python**的工具包**scikit-learn**的**sourcecode**很好读，建议大家多看看。

我看回答中有提到**Matlab**，我觉的**matlab**处理字符很麻烦，现在很多**dataset**都需要处理字符，所以并不是好的选择。

补充一点就是要学会发散思维，学会如何从**data**中找**feature**。关于这个的教程很缺，需要大量练习及一些天赋。

说实话**machine learning**虽然门槛不高，但真心是聪明人的游戏。

编辑于 2014-07-19　25 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

知乎用户，厚积才可薄发

72

肖玉龙、欧阳江卉、pyanfield 等人赞同

刚好是一名小菜正在入门，日学习时间>8h/d，与楼主共勉，基础课程学习完了之后，动手实践吧！

2015/07/01： 根据自己上过的课程，更新课程列表

1. 数学基础

机器学习必要的数学基础主要包括：多元微积分，线性代数

1. Calculus: Single Variable | Calculus One （可选）
2. Multivariable Calculus
3. Linear Algebra

2. 统计基础

1. Introduction to Statistics: Descriptive Statistics
2. Probabilistic Systems Analysis and Applied Probability | 概率 (可选)
3. Introduction to Statistics: Inference

3. 编程基础

1. Programming for Everybody (Python)
2. DataCamp: Learn R with R tutorials and coding challenges (R)
3. Introduction to Computer Science:Build a Search Engine & a Social Network

4. 机器学习

1. Statistical Learning(R)
2. Machine Learning
3. 机器学习基石
4. 机器学习技法

下面是近期的给外行人读的泛数学科普书籍，由浅至深，作用除了感受数学之美之外，更重要的是可以作用每天学习的鸡血，因为这些书都比较好读.....

1.《数学之美》作者：吴军
2.《 Mathematician's Lament | 数学家的叹息》作者：by Paul Lockhart
3.《 Think Stats: Probability and Statistics for Programmers | 统计思维：程序员数学之概率统计 》 作者：Allen B. Downey
4.《 A History of Mathematics | 数学史 》作者：Carl B. Boyer
5.《 Journeys Through Genius | 天才引导的历程：数学中的伟大定理 》作者：William Dunham
6.《 The Mathematical Experience | 数学经验 》作者 Philip J.Davis、Reuben Hersh
7.《 Proofs from the Book | 数学天书中的证明 》作者：Martin Aigner、Günter M. Ziegler
8.《 Proofs and Refutations | 证明与反驳－数学发现的逻辑 》作者：Imre Lakatos

本文源：我的数据挖掘学习图谱

编辑于 2015-07-02　2 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

32

谢澎涛，CMU机器学习

全大葱、谢梦、周君沛 等人赞同

机器学习用到的数学并不难，很多较难的数学（如抽象代数、微分几何）目前在ML问题上也没有用武之地。相比数学，我觉得更重要的一点是对问题和数据的insight。很多经典漂亮的模型，如HMM、CRF、LDA都是建立在良好的motivation和insight之上，数学并不是瓶颈。至于怎么培养insight 恐怕很难说，目前能做的就是多读、多想、多试。

发布于 2015-03-14　2 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

740

知乎用户，Joint Ph.D of Beihang University and M…

陈老千、知乎用户、温融冰 等人赞同

我要翻译一把quora了，再加点我的理解，我相信会是一个好答案，链接我都放到一起了，没插入到正文中，要求其实比较高了，我觉得我自己都差很远很远~~~我尽量持续更新翻译质量以及自己理解

> 1. **Python/C++/R/Java** - you will probably want to learn all of these languages at some point if you want a job in machine-learning. Python's Numpy and Scipy libraries [2] are awesome because they have similar functionality to MATLAB, but can be easily integrated into a web service and also used in Hadoop (see below). C++ will be needed to speed code up. R [3] is great for statistics and plots, and Hadoop [4] is written in Java, so you may need to implement mappers and reducers in Java (although you could use a scripting language via Hadoop streaming [5])

首先，你要熟悉这四种语言。Python因为开源的库比较多，可以看看Numpy和Scipy这两个库，这两个都可以很好的融入网站开发以及Hadoop。C++可以让你的代码跑的更快，R则是一个很好地统计工具。而你想很好地使用Hadoop你也必须懂得java，以及如何实现map reduce

> 2. **Probability and Statistics**: A good portion of learning algorithms are based on this theory. Naive Bayes [6], Gaussian Mixture Models [7], Hidden Markov Models [8], to name a few. You need to have a firm understanding of Probability and Stats to understand these models. Go nuts and study measure theory [9]. Use statistics as an model evaluation metric: confusion matrices, receiver-operator curves, p-values, etc.

我推荐统计学习方法 李航写的，这算的上我mentor的mentor了。理解一些概率的理论，比如贝叶斯，SVM，CRF，HMM，决策树，AdaBoost，逻辑斯蒂回归，然后再稍微看看怎么做evaluation 比如P R F。也可以再看看假设检验的一些东西。

> 3. **Applied Math + Algorithms**: For discriminate models like SVMs [10], you need to have a firm understanding of algorithm theory. Even though you will probably never need to implement an SVM

> from scratch, it helps to understand how the algorithm works. You will need to understand subjects like convex optimization [11], gradient decent [12], quadratic programming [13], lagrange [14], partial differential equations [15], etc. Get used to looking at summations [16].

机器学习毕竟是需要极强极强数学基础的。我希望开始可以深入的了解一些算法的本质，SVM是个很好的下手点。可以从此入手，看看拉格朗日，凸优化都是些什么

> 4. **Distributed Computing**: Most machine learning jobs require working with large data sets these days (see Data Science) [17]. You cannot process this data on a single machine, you will have to distribute it across an entire cluster. Projects like Apache Hadoop [4] and cloud services like Amazon's EC2 [18] makes this very easy and cost-effective. Although Hadoop abstracts away a lot of the hard-core, distributed computing problems, you still need to have a firm understanding of map-reduce [22], distribute-file systems [19], etc. You will most likely want to check out Apache Mahout [20] and Apache Whirr [21].

熟悉分布计算，机器学习当今必须是多台机器跑大数据，要不然没啥意义。请熟悉Hadoop，这对找工作有很大很大的意义。百度等公司都需要hadoop基础。

> 5. **Expertise in Unix Tools**: Unless you are very fortunate, you are going to need to modify the format of your data sets so they can be loaded into R,Hadoop,HBase [23],etc. You can use a scripting language like python (using re) to do this but the best approach is probably just master all of the awesome unix tools that were designed for this: cat [24], grep [25], find [26], awk [27], sed [28], sort [29], cut [30], tr [31], and many more. Since all of the processing will most likely be on linux-based machine (Hadoop doesnt run on Window I believe), you will have access to these tools. You should learn to love them and use them as much as possible. They certainly have made my life a lot easier. A great example can be found here [1].

熟悉Unix的Tool以及命令。百度等公司都是依靠Linux工作的，可能现在依靠Windows的Service公司已经比较少了。所以怎么也要熟悉Unix操作系统的这些指令吧。我记得有个百度的面试题就是问文件复制的事情。

> 6. **Become familiar with the Hadoop sub-projects**: HBase, Zookeeper [32], Hive [33], Mahout, etc. These projects can help you store/access your data, and they scale.

机器学习终究和大数据息息相关，所以Hadoop的子项目要关注，比如HBase Zookeeper Hive等等

> 7. **Learn about advanced signal processing techniques**: feature extraction is one of the most important parts of machine-learning. If your features suck, no matter which algorithm you choose, your going to see horrible performance. Depending on the type of problem you are trying to solve, you may be able to utilize really cool advance signal processing algorithms like: wavelets [42], shearlets [43], curvelets [44], contourlets [45], bandlets [46]. Learn about time-frequency analysis [47], and try to apply it to your problems. If you have not read about Fourier Analysis[48] and Convolution[49], you will need

> to learn about this stuff too. The ladder is signal processing 101 stuff though.

这里主要是在讲特征的提取问题。无论是分类（classification）还是回归（regression）问题，都要解决特征选择和抽取（extraction）的问题。他给出了一些基础的特征抽取的工具如小波等，同时说需要掌握傅里叶分析和卷积等等。这部分我不大了解，大概就是说信号处理你要懂，比如傅里叶这些。。。

> Finally, practice and read as much as you can. In your free time, read papers like Google Map-Reduce [34], Google File System [35], Google Big Table [36], The Unreasonable Effectiveness of Data [37],etc There are great free machine learning books online and you should read those also. [38][39][40]. Here is an awesome course I found and re-posted on github [41]. Instead of using open source packages, code up your own, and compare the results. If you can code an SVM from scratch, you will understand the concept of support vectors, gamma, cost, hyperplanes, etc. It's easy to just load some data up and start training, the hard part is making sense of it all.

总之机器学习如果想要入门分为两方面：
一方面是去看算法，需要极强的数理基础（真的是极强的），从SVM入手，一点点理解。
另一方面是学工具，比如分布式的一些工具以及Unix~

Good luck.
祝好

[1] http://radar.oreilly.com/2011/04...

[2] NumPy — Numpy

[3] The R Project for Statistical Computing

[4] Welcome to Apache™ Hadoop®!

[5] http://hadoop.apache.org/common/...

[6] http://en.wikipedia.org/wiki/Nai...

[7] http://en.wikipedia.org/wiki/Mix...

[8] http://en.wikipedia.org/wiki/Hid...

[9] http://en.wikipedia.org/wiki/Mea...

[10] http://en.wikipedia.org/wiki/Sup...

[11] http://en.wikipedia.org/wiki/Con...

[12] http://en.wikipedia.org/wiki/Gra...

[13] http://en.wikipedia.org/wiki/Qua...

[14] http://en.wikipedia.org/wiki/Lag...

[15] http://en.wikipedia.org/wiki/Par...

[16] http://en.wikipedia.org/wiki/Sum...

[17] http://radar.oreilly.com/2010/06...

[18] AWS | Amazon Elastic Compute Cloud (EC2)

[19] http://en.wikipedia.org/wiki/Goo...

[20] Apache Mahout: Scalable machine learning and data mining

[21] incubator.apache.org/wh...

[22] http://en.wikipedia.org/wiki/Map...

[23] HBase - Apache HBase Home

[24] http://en.wikipedia.org/wiki/Cat...

[25] grep

[26] en.wikipedia.org/wiki/F...

[27] AWK

[28] sed

[29] http://en.wikipedia.org/wiki/Sor...

[30] http://en.wikipedia.org/wiki/Cut...

[31] http://en.wikipedia.org/wiki/Tr_...

[32] Apache ZooKeeper

[33] Apache Hive TM

[34] http://static.googleusercontent....

[35]http://static.googleusercontent....

[36]http://static.googleusercontent....

[37]http://static.googleusercontent....

[38] http://www.ics.uci.edu/~welling/...

[39] http://www.stanford.edu/~hastie/...

[40] http://infolab.stanford.edu/~ull...

[41] https://github.com/josephmisiti/...

[42] http://en.wikipedia.org/wiki/Wav...

[43] http://www.shearlet.uni-osnabrue...

[44] http://math.mit.edu/icg/papers/F...

[45] http://www.ifp.illinois.edu/~min...

[46] http://www.cmap.polytechnique.fr...

[47]http://en.wikipedia.org/wiki/Tim...

[48] http://en.wikipedia.org/wiki/Fou...

[49]http://en.wikipedia.org/wiki/Con...

编辑于 2014-02-17　29 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

165

王丰，**Data is power.**

Jason、知乎用户、Scott Tao 等人赞同

不同意第一名的答案。

知乎一直这样，精英主义太严重，好像不表现得专业点、不长点就不是好答案。说白了就是装b。知乎不是wikipedia（wikipedia适合查资料，不适合从零开始学习）。

题主的问题是这样的：

> 本人大学本科，对机器学习很感兴趣，想从事这方面的研究。在网上看到机器学习有一些经典书如Bishop的PRML， Tom Mitchell的machine learning，还有pattern classification，不知该如何入门？那本书比较容易理解？

注意题主是本科，现在感兴趣，希望能入门。让一个想要入门机器学习的本科生上来就学Hadoop，这靠谱吗？估计等题主熟练掌握了Hadoop、MapReduce、HBase，本科也快结束了。这种问题天生就不适合长答案。对一个初学者，无重点地给出很多资料，除了能打击学习兴趣、扰乱学习目标之外没有任何作用。

同意 @Darkscope@苗忆南 等人的答案。

题主既然是初学者，就要从简单入手。而大学本科的概率论线性代数对于基本的入手已经足够了。Andrew Ng的课非常好（很简单很适合入门），值得去学一学，网上的相关资料也特别多，比如这个：Machine Learning ，写得很详细。如果题主英文水平有限，那可以看一些中文书籍入门。

入门之后，对常用算法有了基本了解之后，就可以多学一些原理性知识了，比如统计理论、矩阵理论、信号处理、分布式计算等等。这时排名第一的答案才比较有用。

编辑于 2014-05-20　4 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

许靖，银行/数据挖掘/andriod开发

知乎用户、知乎用户、张驰 等人赞同

最近在学teradata的aster数据挖掘工具，果然商业的框架跟hadoop的易学性易用性可拓展性都有天壤之别啊。建议有兴趣的同学自查资料，软件贵的惊人，但是对于自身维护团队水平一般的企业还是值得买的

不邀自来，结合我现在的工作以及当年机器学习的经验来讲一下自己的一些感受。
首先从机器学习的全流程说起。
输入：机器学习按输入数据分可以分结构化数据(表)，半结构化数据(文本，日志等)，非结构化数据(图片，录像)，后两者对应的分支中比较出名分别是文本挖掘和图像挖掘，其实就是通过提取特征的方式 把半结构化数据非结构化数据转化为结构化数据，然后进行机器学习。
按输入数据是否预测状态分，可以分为有监督学习数据及无监督学习数据，直接决定后续模型是使用分类模型还是聚类模型。
同时如果数据量到达了T级，就要考虑时候hadoop框架了，这里要说明的是，hadoop框架只是解决大数据处理效率瓶颈的工具，除非你志向是做架构师，不然不用深究，掌握hadoop家族的sql处理工具hive和机器学习工具mahout就可以了，需要掌握java和sql，这是基础。

下一步，数据清洗和数据降维，在数据清洗方面不是学术界研究的重点，清洗方法有很多，主要通过描述性统计量填补缺失值和极端值，数据降维方面有较多比较出名的算法，如主成分分析，lasso，LDA线性判别分析，变量聚类等，数据降维是重点，因为维度过大容易产生维度灾难和过度拟合问题。
然后是数据分区(有监督学习才需要做)，分成训练集，验证集，测试集，分别用于训练模型，模型内修正，多模型效果对比用。不展开讨论。
接着建模，分为分类算法，聚类算法，规则关联算法，分类算法是大头，建议分别拜读支持向量机，概率图模型，神经网络(虽然我不喜欢但是google的深度学习就是用的神经网络)，决策树(C4.5)，逻辑回归(吐槽：线性回归什么的看不看得看个人水平)，混合高斯模型等，聚类算法有KNN，LDA潜在迪力克雷分析(做文本挖掘效果一流)，聚类算法研究不多，大多机器学习算法都是基于有监督学习的，即分类算法。规则关联算法有最简单的规则关联，路径关联分析，及协同过滤(推荐系统首选，输入数据量太大数据太稀疏跑数效率很低一般需要hadoop支持)，模型是机器学习的核心，还有一些优化模型，如EM模型等，建议阅读增加知识广度，要求学习人员要懂以下课程：概率论，高数，线性优化，线性代数还有英文，也可以边看论文边学。书籍的话推荐数据挖掘十大算法作为入门索引，然后根据每章节的参考文献找原文阅读，记得把公式推导一遍。
结果评估：本来不应该是重点，但是从个人学习和工作经历发现，有很多人不是很会看模型结果，甚至包括一些专业人士。所以如果有志做机器学习的千万要学会看结果，不然就丢人丢大了。
最直接的两个指标准确度percision和召回度recall，分别描述模型的精度和泛化性的，模型结果应在两者取得均衡。最小平方误差也是描述准确度的，算法不一样，大家更喜欢用这个评估模型精度。还有一些K-S值，基尼值，ROC值也是描述准确度和拟合度的，不同的软件会使用不同的指标，注意一下。还有就是提升度，事件捕获率这些就是描述分类预测下前百分之几数据的预测准确性情况的，不展开讨论。最直接的学法就是认真读论文的实验部分，看看它的指标评价量。
最后是工具和语言，按现在数据挖掘与机器学习的趋势，必须掌握java，python，原因是hadoop是处理大数据的框架，已成趋势，hadoop是java写的，挖掘工具mahout是java写的。python的好处在于底层基于C，语法简单，效率高，而且有很多开源的算法可以直接用，支持mapreduce。可以选择掌握R，同python，但是R效率木有python高，如果是针对商业分析建议学习SAS，因为它集成了一套专业的数据可视化及数据分析方案，大大方便了数据展示功能，这是作为商业分析所必须的。
最后补充一点，很多做数据挖掘和机器学习的人最后都走火入魔执着技术忽略了业务的重要性，忽略了模型可解释性的重要性，埋头苦干做出业务人员不能理解的黑箱模型，无法说服业务使用，最终沦为鸡肋的存在，所以在学习过程中要时刻设想技术适应的业务场景，在算法效果接近时选择高可解释性算法，做好结果的图形化展示，让业务理解，这才是王道。
补充，andrew ng的公开课不错，但还是建议先看论文。

2014/10/23更新：

这两天看到李航老师的《统计学习方法》，感觉写的非常好，适合入门，机器学习的基本概念都有，但是不太深入，中文书写，所有专业名词给出英文翻译。适合给初学者建立概念，可以系统的了解机器学习。

原答案：
强烈推荐这个UFLDL教程 - Ufldl 。这是Andrew Ng写的关于非监督特征学习与深度学习的教程，关键是有一

批无私且专业的网友，将其翻译成中文，并有中英文对照，与Andrew Ng商量后贴在了原网址上。非常感谢这些人啊。
对于一个初学者，如果单纯从英文教材（视频）入手的话，会比较吃力，很多概念都没建立起来，很多术语都没有掌握，而这个教程设计机器学习很多的基本概念，并附有matlab习题，通过循序渐进的练习，可以更快掌握基本概念。
另外这个的好处是不像一般教材，面面俱到，很多追究的太深，不利于初学者建立概念！有了这个的基础之后，再去看相关著作或者论文，肯定得心应手。

编辑于 2014-10-23　13 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

22

Raymain1944，INTP大学狗 | raymain1944.com

杨青、佟津乐、陈军 等人赞同

(●—●)我来弱弱的反对一下精英主义答案。
作为一名属于正常正态分布的中国大学僧，学习机器学习方面入门极力推荐《机器学习实战》。这本书电子版多看书城就有，排版很精美……(●—●)
最主要的是两个字，入门。以py为基础，直接从实例入手，第一章还没完你就能自己码一个小小的程序来让机器真正的“学习”一下。极力反对那些上来就扔一大堆相关学科理论知识的小伙伴们。(￣▽￣)咱就是入门一下，探个头看看这坑咱能进不？不行掉头就走……

发布于 2015-07-02　11 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

6

知乎用户，software engineer

杨艾森、杨旻、Hsiao Lan 等人赞同

分听课和看书两个部分来说：
入门：
视频可以看coursera上Andrew Ng的machine learning课
书国内的可以看李航的统计学习方法，综合了老外的基本ML经典，写得浅显易懂，书也比较薄，好读，学习曲线不会太陡。国外的建议先看pattern classification，较其他的简单一些。

进阶：
视频的话可以看看coursera上的一些数值计算和最优化课程
书老外的大部头多了，经典的比如PRML（patten recognition and machine learning），elements fo statistical learning（此书虽名叫基础，实则十分艰深难读，不推荐初学者学习）。这两本都能读完且读懂了，那是看最前沿的ml的paper也不会发怵了

发布于 2013-12-17　1 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

106

知乎用户

neil2wm、知乎用户、小言 等人赞同

coursera上的andrew ng教授的公开课，妥妥的，绝对靠谱 class.coursera.org/ml/c...

发布于 2013-05-06　28 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

知乎用户

48

aBcs、杨青、戴戴 等人赞同

本人原来是学EE的，现在马上大四选了个机器学习方向的毕业论文，准备毕业后读机器学习方面的研究生了。现在正是暑假，在家自学机器学习，我就按照我的学习步骤脚要说一下吧。这里推荐的都是一些很容易得到的资源，也是推荐的人比较多的。

1.我读的第一本书是大名鼎鼎的ESL作者写的另外一本简略版ESL：Introduction to Statistical Learning 。这本书删去了复杂的数学推导，以应用的形式把机器学习的基本算法介绍出来，好像常规的算法除了神经网络其他都涉及了，并且更好的是，每一章后面都有Lab，而且是用R语言来做的，对于想学习R语言的童鞋来说又是一大福利。这本书我正在看第二遍。

2.Andrew Ng的机器学习公开课，这里推荐的不是C站上的，而是斯坦福现场上课的录像，网易公开课上有(斯坦福大学公开课：机器学习课程)，原始的课程地址(CS 229: Machine Learning (Course handouts))可以下到讲义和作业以及答案，这里推荐好好研究一下讲义和作业，很有收获。这门课我正在看第二遍，也在研究作业作业题。

3.来自多伦多大学的机器学习与数据挖掘课程，讲义写的非常好，不过很难找到，这里给出网盘地址(Machine Learning and Data Mining Lecture Notes.pdf_免费高速下载)，原始课程地址(Professor Richard Zemel)上有PPT，作业和考试一些内容。

4.巨多人推荐的MIT线性代数公开课(麻省理工公开课：线性代数)，来自Gilbert Strang,在这里废话不多说，机器学习书籍里常见到很多矩阵运算还有SVD什么的，看完这个就不怕啦。

5.来自Sam Roweis的机器学习课程，很多人推荐他的PPT(CSC2515F : lectures)，写的超级详细和精致，适合打印出来没事的时候翻翻。

6.同样是机器学习的数学基础，统计学也是要学一学的，机器学习中常见的极大似然法就是一个统计学的基础理论，在这里推荐国立交大的陈邻安老师主讲的统计学(國立交通大學開放式課程(OpenCourseWare, OCW))和高等统计学(國立交通大學開放式課程(OpenCourseWare, OCW))，讲的好懂但不浅显。

GitHub上有个很好的关于机器学习和深度学习的合集(Qix/dl.md at master · ty4z2008/Qix · GitHub)，资料很多，慢慢淘会找到一些好东西的。
暂时我看的资料就这么多，又补充会随时回来更新的。

编辑于 2015-07-11　3 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

17

酸性沼泽软泥怪，曾获微博达人称号

杨青、谷可可、知乎用户 等人赞同

强烈推荐台大的这门《机器学习基石》。我感觉这门比Andrew Ng的讲得更深入更透彻更直观，而且作业题也需要更多的思考。给个coursera的链接：class.coursera.org/ntum...

发布于 2014-04-14　2 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

Spirit_Dongdong，向成为Data Scientist的目标努力!~

7

知乎用户、刘彬、知乎用户 等人赞同

尝试跟一下斯坦福的机器学期公开课吧
已经全部翻译完整了
v.163.com/special/openc...

发布于 2013-01-09 添加评论 感谢 分享 收藏 • 没有帮助 • 举报 • 作者保留权利

7

dontbeatmycat，征女友，有意私信。

陈淦燊、王海啸、Hattori 等人赞同

台大林轩田的《机器学习基石》和《机器学习技法》公开课，很适合入门，采用频率学派观点，作业的理论题有一定难度。

编辑于 2015-05-12 1 条评论 感谢 分享 收藏 • 没有帮助 • 举报 • 作者保留权利

8

黎清水，在用生命说

LongLong Jing、战军丞、知乎用户 等人赞同

Mitchell的机器学习书是懂的部分你不需要看 不懂的地方看不懂
PRML入门挺好的，主要是数学讲的细，循序渐进。如果Iz数学底子好可以直接上pattern classification，这本书主要是有翻译的，是个优点，但偏老了些。

顺便吐槽说PRML不适合本科生的，现在哪个机器学习实验室入门不是看PRML或者pattern classification啊，大一到大三进实验室都得乖乖看这两本。基础打扎实了到大四再分具体方向看Probabilistic Graphical Models这些进阶书。

发布于 2013-05-21 8 条评论 感谢 分享 收藏 • 没有帮助 • 举报 • 作者保留权利

6

知乎用户，from the Planet Krypton

夜行者、Petzold、知乎用户 等人赞同

- 什么是机器学习，请参考这里，

什么是机器学习？

- 建议视频、书和编程实践结合起来学习。
- 视频：coursera 林轩田 《機器學習基石 (Machine Learning Foundations) 》

网易公开课 Andrew Ng 《 斯坦福大学公开课 ： 机器学习课程 》

- 书籍：李航 《统计学习方法 (豆瓣) 》

Christopher M. Bishop 《Pattern Recognition And Machine Learning (豆瓣) 》

- 编程实践：《Peter Harrington 机器学习实战 (豆瓣) 》

《TOBY SEGARAN 集体智慧编程 (豆瓣) 》

- 从头入门的话，可以先看李航的《统计学习方法》和《机器学习实战》，Coursera上有吴恩达的ML的课程，最近才开你可以去跟着看看。我其实更推荐A站上的台大林老师的课，虽然第一部分结课了，但是现在还可以看，课后题认真做收获很大的。等这些都看完以后，补充点数学知识，就可以去看PRML或者MLAPP或者ESL，然后就要看你的方向了。

碰到不明白的数学理论，直接谷歌百度就可以了。
这些消化了，机器学习基本算入门了。

编辑于 2015-08-31　11 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 禁止转载

6

知乎用户，机器学习爱好者，《机器学习：实用案例解...

李小燚、李晓芸、陈榆丰 等人赞同

我怎么觉得《统计学习方法》不适合入门呢？有点像中国的教科书，一上来就列很多公式，然后开始一通推导，有些还不如PRML讲得通俗。

如果是工程师的话，我觉得Logistic Regression是一个比较适合入门的东西，通过它可以搞懂：
1、分类和回归，在这一个算法中都得到了体现；
2、正则，从贝叶斯观点理解最大似然和最大后验的区别；
3、最优化，最好搞懂如何通过梯度下降或者随机梯度下降得到LR模型参数；
4、特征选择，特征转换等特征工程；
5、模型评价。

从它入手，可以继续深入指数族；
通过SGD,可以对不同损失函数去尝试一下优化过程；
另外，LR模型应用很广，广告系统和推荐系统都喜欢用这玩意儿。

发布于 2014-04-27　1 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

知乎用户，没人要的逗逼民工

60

Yao LIU、Anything、张展雄 等人赞同

建议你跟着世界一流高校的一门课学习，而不要去看ng 的公开课。
stanford 一门ee xxx， machine learning，据说效果很好
berkeley有cs 289 A/B 供您选择
mit 想必也是有的
cmu 想必也是有的

这些课程都有网络版

教材的话，楼上有人说了，elements of statistical learning (esl)。这个说的高屋建瓴，适合数学基础好的人看；另外一本书是introduction to statistical learning with R。这个说的很实际，适合入门理解。
实际上书都写得很好，问题是你有多想学。你把他们都好好学会了，获得学校里面科班的训练，那你绝对比知乎上99%的半吊子水平要高了。剩下的就是学一学软件。
亚洲的精英教育鼓励大家追究刁钻的细节，却不清楚一门学问里最基本的普世的动机。却不知道，入门最基本的就是了解这个动机。刁钻的技术细节，一个理解动机的人很容易就能看懂

另附cs 289A 的内容大纲：

- Introduction: applications, methods, concepts
- Good Machine Learning hygiene: test/training/validation, overfitting
- Linear classification
  - ***Perceptron*** algorithm
  - Support vector machines (***SVMs***)
- Statistical learning background
  - Decision theory; Bayes risk
  - Probabilistic models vs no model
  - Generative and discriminative models
  - Controlling complexity: regularization, bias-variance trade-off, priors.
  - Resampling, cross-validation.
  - The multivariate normal distribution.
- Linear regression
  - ***Least squares***
  - Regularization: ***ridge regression, lasso***
- Brief primer on optimization
- Linear Classification, revisited

    - ***Logistic regression***
    - ***Linear Discriminant Analysis***
    - Support vector machines revisited
        - Algorithms
        - ***The kernel trick***
- Theoretical analysis of machine learning problems and algorithms
    - Generalization error bounds; VC dimension
- Nearest neighbor methods
    - ***k-nearest-neighbor***
    - Properties of high-dimensional spaces
    - Distance learning
    - Efficient indexing and retrieval methods
- Decision trees
    - ***Classification and regression trees***
    - Random Forests
    - ***Boosting***
- Neural networks
    - ***Multilayer perceptrons***
    - Variations such as convolutional nets
    - Applications
- Unsupervised methods
    - ***Clustering***
    - Density estimation
    - Dimensionality reduction
- Applications in ***Data Mining***
    - Collaborative filtering
    - The power and the peril of Big Data

编辑于 2015-03-16　13 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

8

知乎用户，rational

Petzold、陈henry、谷可可 等人赞同

李航的<<统计学习方法>>我觉得很不错,基础的模型和数学推导基本上都有了.PRML我觉得不是很适合初学者,涉及太多的基础理论.数学稍微不好就会看的很难受.

另外AndrewNg的course确实非常适合初学者,不过coursera上也有挺多其他的ML相关的course,我记得有个台湾大学的林老师讲svm讲的非常深入。

其它的,如果题主你是研究生,很可能你导师的课题是非常前沿的(在ML领域前沿的也就是deep learning了.....),很可能你老板不会太注意你的基础,而是直接让你在各种deep learning的framework上码起来,说实话这样对你并没有好处.

如果你的目标定在将来做ML领域的工作,而不是做理论研究,我建议题主,把logistic regression,svm,random forest这些算法弄明白,然后多写点代码.现在工程领域的ML基本上是逻辑斯特回归走天下,random forest都算比较复杂的算法了.

好了,最后还说一句,多写代码,数学其实没有那么重要,ML的数学基础无非就是优化理论,统计学相关的只是帮助你去理解和分析模型，而具体到模型求参则会用到优化理论。常用的优化方法如求解neural network的gradient descent，看代码非常容易明白。

编辑于 2015-09-06　4 条评论　感谢　分享　收藏 • 没有帮助 • 举报 • 作者保留权利

更多

我来回答这个问题

写回答...

我要回答

关注问题 5321 人关注该问题

相关问题 换一换

学习录音，如何入门？ 12 个回答

有哪些机器学习、图像识别方面的入门书籍？ 17 个回答

如何从入门开始学习OpenCV？ 9 个回答

学编程该如何入门？ 2 个回答

想学马术该如何入门？ 10 个回答