机器学习知识梳理

感谢 tang\_Kaka\_back@新浪微博 的归纳整理，

我们先从下面4张图，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。

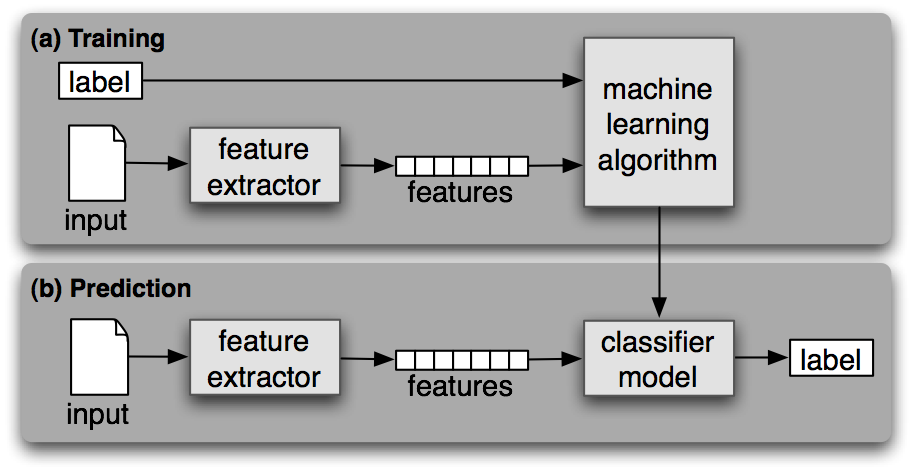
[](https://camo.githubusercontent.com/f814186058922d71de1dd9cf841299585ea64ae5/687474703a2f2f7777772e6e6c746b2e6f72672f696d616765732f737570657276697365642d636c617373696669636174696f6e2e706e67)

图1: 机器学习的例子：NLTK监督学习的工作流程图 (source: <http://www.nltk.org/book/ch06.html>)

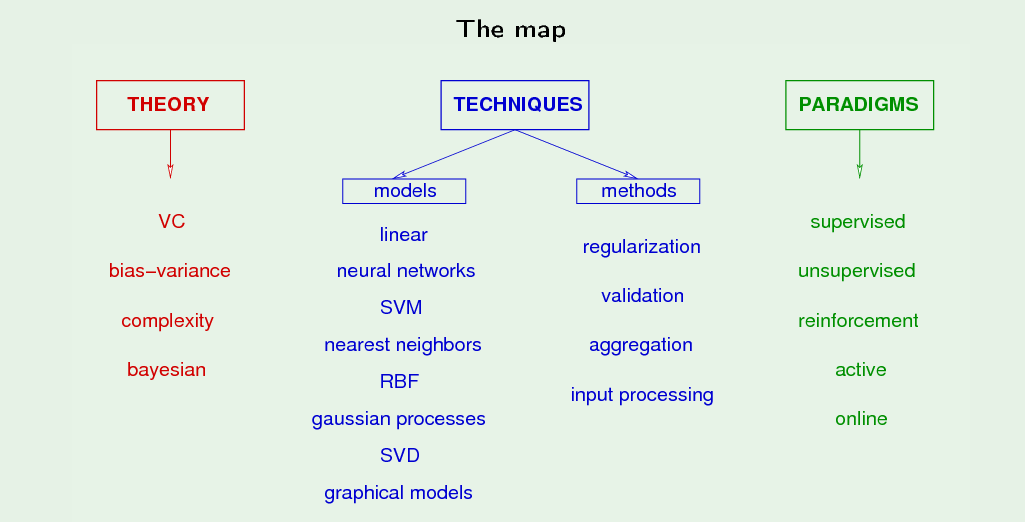
[](https://camo.githubusercontent.com/b7e7f6d502ce3d6fb9247943a846853525549120/687474703a2f2f776f726b2e63616c746563682e6564752f696d61676573312f6d61702e706e67)

图2: 机器学习概要图 by Yaser Abu-Mostafa (Caltech) (source: <http://work.caltech.edu/library/181.html>)

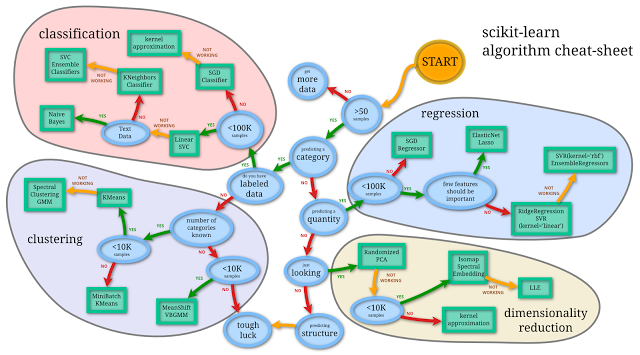


图3: 机器学习实战：在python scikit learn 中选择机器学习算法 by Nishant Chandra

(source: <http://n-chandra.blogspot.com/2013/01/picking-machine-learning-algorithm.html>)



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: <http://nirvacana.com/thoughts/becoming-a-data-scientist/>)

继续，

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](http://www.guokr.com/post/512037/) (2013) 作者 [白马](http://www.guokr.com/group/i/0373595356/) -- [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](http://ourcoders.com/thread/show/2837/) -- [起步体悟] 研究生型入门者的亲身经历，尤其要看[reyoung](http://ourcoders.com/user/show/25895/reyoung/)的建议
* [tornadomeet 机器学习 笔记](http://www.cnblogs.com/tornadomeet/tag/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/) (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](https://machinelearningmastery.com/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](http://machinelearningmastery.com/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](http://machinelearningmastery.com/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](http://blog.jobbole.com/56256/) 译者 [programmer\_lin](http://www.jobbole.com/members/linwenhui/)
* 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

更多攻略

* [What's the easiest way to learn machine learning @quora](http://www.quora.com/Whats-the-easiest-way-to-learn-machine-learning) (2013)
* [What is the best way to study machine learning @quora](http://www.quora.com/What-is-the-best-way-to-study-machine-learning) (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](http://www.quora.com/Is-there-any-roadmap-for-learning-Machine-Learning-ML-and-its-related-courses-at-CMU) (2014)

## 课程资源

Tom Mitchell 和 Andrew Ng 的课都很适合入门

## 入门课程

#### 2011 Tom Mitchell(CMU)机器学习

[英文原版视频与课件PDF](http://www.cs.cmu.edu/~tom/10701_sp11/lectures.shtml) 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

#### 2014 Andrew Ng (Stanford)机器学习

[英文原版视频](https://www.coursera.org/course/ml) 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

### 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](http://work.caltech.edu/lectures.html)

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](https://www.coursera.org/course/ntumlone)

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]

### 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](http://v.163.com/special/opencourse/machinelearning.html) | [英文版视频@youtube](https://www.youtube.com/playlist?list=PLA89DCFA6ADACE599) | [课件PDF@Stanford](http://cs229.stanford.edu/materials.html)

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](http://wenku.baidu.com/course/view/49e8b8f67c1cfad6195fa705) ｜ [课件PDF@龙星计划](http://bigeye.au.tsinghua.edu.cn/DragonStar2012/download.html)

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)

<http://www.kdnuggets.com/> 数据挖掘名站

<http://www.datasciencecentral.com/> 数据科学中心网站

## 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](http://en.wikipedia.org/wiki/Machine_learning#Machine_learning_and_data_mining)

* 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?](http://www.quora.com/What-are-some-good-machine-learning-jokes)

* 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](http://machinelearningmastery.com/best-machine-learning-resources-for-getting-started/)

* [The Discipline of Machine Learning](http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf)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](http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf) 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。

更多人工智能学习材料，请扫描下方二维码：

