86 | 程序员练级攻略：机器学习和人工智能

2018-7-26 陈皓

我之前写过一篇机器学习的入门文章，因为我也是在入门和在学习的人，所以，那篇文章和这篇机器学习和人工智能方向的文章可能都会有点太肤浅。如果你有更好的学习方式或资料，欢迎补充。

基本原理简介

我们先来介绍一下机器学习的基本原理。

机器学习主要有两种方式，一种是监督式学习（Supervised Learning），另一种是非监督式学习（Unsupervised Learning）。下面简单地说一下这两者的不同。

监督式学习（Supervised Learning）。所谓监督式学习，也就是说，我们需要提供一组学习样本，包括相关的特征数据和相应的标签。我们的程序可以通过这组样本来学习相关的规律或是模式，然后通过得到的规律或模式来判断没有被打过标签的数据是什么样的数据。

举个例子，假设需要识别一些手写的数字，我们要找到尽可能多的手写体数字的图像样本，然后人工或是通过某种算法来明确地标注上什么是这些手写体的图片，谁是1，谁是2，谁是3…… 这组数据叫样本数据，又叫训练数据（training data）。然后通过机器学习的算法，找到每个数字在不同手写体下的特征，找到规律和模式。通过得到的规律或模式来识别那些没有被打过标签的手写数据，以此完成识别手写体数字的目的。

非监督式学习（Unsupervised Learning）。对于非监督式学习，也就是说，数据是没有被标注过的，所以相关的机器学习算法需要找到这些数据中的共性。因为大量的数据是没被被标识过的，所以这种学习方式可以让大量的未标识的数据能够更有价值。而且，非监督式学习，可以为我们找到人类很难发现的数据里的规律或模型，所以也有人称这种学习为“特征点学习”，其可以让我们自动地为数据进行分类，并找到分类的模型。

一般来说，非监督式学习会应用在一些交易型的数据中。比如，你有一堆堆的用户购买数据，但是对于人类来说，我们很难找到用户属性和购买商品类型之间的关系。所以，非监督式学习算法可以帮助我们找到它们之间的关系。比如，一个在某年龄段的女性购买了某种肥皂，有可能说明这个女性在怀孕期，或是某人购买儿童用品，有可能说明这个人的关系链中有孩子，等等。于是，这些信息会被用作一些所谓的精准市场营销活动，从而可以增加商品销量。

我们这么来说吧，监督式学习是在被告诉过了正确的答案后的学习，而非监督式学习是在没有被告诉正确答案时的学习。所以，非监督式学习是在大量的非常乱的数据中找寻一些潜在的关系，这个成本也比较高。非监督式学习经常被用来检测一些不正常的事情发生，比如信用卡的诈骗或是盗刷。也被用在推荐系统，比如买了这个商品的人又买了别的什么商品，或是如果某个人喜欢某篇文章、某个音乐、某个餐馆，那么他可能会喜欢某个车、某个明星或某个地方。

在监督式学习算法下，我们可以用一组“狗”的照片来确定某个照片中的物体是不是狗。而在非监督式学习算法下，我们可以通过一个照片来找到其中有与其相似的事物的照片。这两种学习方式都有些有用的场景。

关于机器学习，你可以读一读 Machine Learning is Fun! ，这篇文章（中文翻译版）恐怕是全世界最简单的入门资料了。

- Data Science Simplified Part 1: Principles and Process
- Data Science Simplified Part 2: Key Concepts of Statistical Learning
- Data Science Simplified Part 3: Hypothesis Testing
- Data Science Simplified Part 4: Simple Linear Regression Models
- Data Science Simplified Part 5: Multivariate Regression Models
- Data Science Simplified Part 6: Model Selection Methods
- Data Science Simplified Part 7: Log-Log Regression Models
- Data Science Simplified Part 8: Qualitative Variables in Regression Models
- Data Science Simplified Part 9: Interactions and Limitations of Regression Models
- Data Science Simplified Part 10: An Introduction to Classification Models
- Data Science Simplified Part 11: Logistic Regression

相关课程

接下来，我们需要比较专业地学习一下机器学习了。

在学习机器学习之前，我们需要学习数据分析，所以，我们得先学一些大数据相关的东西，也就是Data Science相关的内容。下面是两个不错的和数据科学相关的教程以及一个资源列表。

UC Berkeley’ s Data 8: The Foundations of Data Science 和电子书 Computational and Inferential Thinking 会讲述数据科学方面非常关键的概念，会教你在数据中找到数据的关联、预测和相关的推断。

Learn Data Science ，这是GitHub上的一本电子书，主要是一些数据挖掘的算法，比如线性回归、逻辑回归、随机森林、K-Means聚类的数据分析。然后，donnemartin/data-science-ipython-notebooks 这个代码仓库中用TensorFlow、scikit-learn、Pandas、NumPy、Spark等把这些经典的例子实现了个遍。

Data Science Resources List ，这个网站上有一个非常长的和数据科学相关的资源列表，你可以从中得到很多你想要的东西。

之后，有下面几门不错的在线机器学习的课程供你入门，也是非常不错。

吴恩达教授（Andrew Ng）在 Coursera 上的免费机器学习课程 非常棒。我强烈建议从此入手。对于任何拥有计算机或科学学位的人，或是还能记住一点点数学知识的人来说，都应该非常容易入门。这个斯坦福大学的课程请尽量拿满分。可以在 网易公开课 中找到这一课程。除此之外，吴恩达教授还有一组新的和深度学习相关的课程，现在可以在网易公开课上免费学习——Deep Learning Specialization。

Deep Learning by Google ，Google的一个关于深度学习的在线免费课程，其支持中英文。这门课会教授你如何训练和优化基本神经网络、卷积神经网络和长短期记忆网络。你将通过项目和任务接触完整的机器学习系统TensorFlow。

卡内基梅隆大学汤姆·米切尔（Tom Mitchell）的机器学习 英文原版视频与课件PDF 。

2013年加利福尼亚理工学院亚瑟·阿布-穆斯塔法（Yaser Abu-Mostafa）的Learning from Data 课程视频及课件PDF，内容更适合进阶。

关于神经网络方面，YouTube上有一个非常火的课程视频，由宾夕法尼亚大学的雨果·拉罗歇尔（Hugo Larochelle）出品的教学课程 - Neural networks class - Université de Sherbrooke 。

除此之外，还有很多的在线大学课程可以供你学习。比如：

斯坦福大学的《统计学学习》、《机器学习》、《卷积神经网络》、《深度学习之自然语言处理》等。

麻省理工大学的《神经网络介绍 》、《机器学习》、《预测》等。

更多的列表，请参看——Awesome Machine Learning Courses。

相关图书

《Pattern Recognition and Machine Learning》，这本书是机器学习领域的圣经之作。该书也是众多高校机器学习研究生课程的教科书，Google上有PDF版的下载。这本书很经典，但并不适合入门来看。GitHub上有这本中的 Matlab 实现。

下面这两本电子书也是比较经典的，其中讲了很多机器学习的知识，可以当做手册或字典。

《Understanding Machine Learning: From Theory to Algorithms》。

《The Elements of Statistical Learning - Second Edition》。

《Deep Learning: Adaptive Computation and Machine Learning series》 中文翻译为《深度学习》。这本书由全球知名的三位专家伊恩·古德费洛（Ian Goodfellow）、友华·本吉奥（Yoshua Bengio）和亚伦·考维尔（Aaron Courville）撰写，是深度学习领域奠基性的经典教材。
全书内容包括3部分：第1部分介绍基本的数学工具和机器学习的概念，它们是深度学习的预备知识；第2部分系统深入地讲解现今已成熟的深度学习方法和技术；第3部分讨论某些具有前瞻性的方向和想法，它们被公认为是深度学习未来的研究重点。这本书的官网为“deeplearningbook.org”，在GitHub上也有中文翻译 - 《Deep Learning 中文翻译》。

《Neural Networks and Deep Learning》（中文翻译版），这是一本非常不错的神经网络的入门书，在豆瓣上评分9.5分，从理论讲到了代码。虽然有很多数学公式，但是有代码相助，就不难理解了。其中讲了很多如激活函数、代价函数、随机梯度下降、反向传播、过度拟合和规范化、权重初始化、超参数优化、卷积网络的局部感受野、混合层、特征映射的东西。

《Introduction to Machine Learning with Python》，算是本不错的入门书，也是本比较易读的英文书。其是以Scikit-Learn框架来讲述的。如果你用过Scikit这个框架，那么你学这本书还是很不错的。

《Hands-On Machine Learning with Scikit-Learn and TensorFlow 》，这是一门以TensorFlow为工具的入门书，其用丰富的例子从实站的角度来让你学习。这本书对于无基础的人也是适合的，对于小白来说虽然略难但是受益匪浅。

相关文章

除了上述的那些课程和图书外，下面这些文章也很不错。

YouTube 上的 Google Developers 的 Machine Learning Recipes with Josh Gordon ，这9集视频，每集不到10分钟，从Hello World讲到如何使用TensorFlow，非常值得一看。

还有 Practical Machine Learning Tutorial with Python Introduction 上面一系列的用Python带着你玩Machine Learning的教程。

Medium上的 Machine Learning - 101 ，讲述了好些我们上面提到过的经典算法。

Medium上的 Marchine Learning for Humans。

Dr. Jason Brownlee 的博客 ，也非常值得一读，其中好多的 “How-To” ，会让你有很多的收获。

Rules of Machine Learning: Best Practices for ML Engineering ，一些机器学习相关的最佳实践。

i am trask ，也是一个很不错的博客。

关于Deep Learning中的神经网络，YouTube上有介绍视频 Neural Networks。

麻省理工学院的电子书 Deep Learning。

用Python做自然语言处理Natural Language Processing with Python。

最后一个是Machine Learning和Deep Learning的相关教程列表，Machine Learning & Deep Learning Tutorials。

下面是一些和神经网络相关的不错的文章。

The Unreasonable Effectiveness of Recurrent Neural Networks ，这是一篇必读的文章 ，告诉你为什么要学RNN，以及展示了最简单的NLP形式。

Neural Networks, Manifolds, and Topology ，这篇文章可以帮助你理解神经网络的一些概念。

Understanding LSTM Networks ，解释了什么是LSTM的内在工作原理。

Attention and Augmented Recurrent Neural Networks ，用了好多图来说明了RNN的attention机制。

Recommending music on Spotify with deep learning ，一个在Spotify的实习生分享的音乐聚类的文章。

相关算法

下面是10个非常经典的机器学习的算法。

对于监督式学习，有如下经典算法。

1. 决策树（Decision Tree），比如自动化放贷、风控。
2. 朴素贝叶斯分类器（Naive Bayesian classifier)，可以用于判断垃圾邮件、对新闻的类别进行分类，比如科技、政治、运动、判断文本表达的感情是积极的还是消极的、人脸识别等。
3. 最小二乘法（Ordinary Least Squares Regression），是一种线性回归。
4. 逻辑回归（Logisitic Regression），一种强大的统计学方法，可以用一个或多个变量来表示一个二项式结果。可以用于信用评分，计算营销活动的成功率，预测某个产品的收入。
5. 支持向量机（Support Vector Machine，SVM），可以用于基于图像的性别检测、图像分类等。
6. 集成方法（Ensemble methods），通过构建一组分类器，然后通过它们的预测结果进行加权投票来对新的数据点进行分类。原始的集成方法是贝叶斯平均，但最近的算法包括纠错输出编码、Bagging和Boosting。

对于无监督式的学习，有如下经典算法。

1. 聚类算法（Clustering Algorithms）。聚类算法有很多，目标是给数据分类。有5个比较著名的聚类算法你必需要知道：K-Means、Mean-Shift、DBSCAN、EM/GMM、和 Agglomerative Hierarchical。
2. 主成分分析（Principal Component Analysis，PCA）。PCA的一些应用包括压缩、简化数据便于学习、可视化等。
3. 奇异值分解（Singular Value Decomposition，SVD）。实际上，PCA是SVD的一个简单应用。在计算机视觉中，第一个人脸识别算法使用PCA和SVD来将面部表示为"特征面"的线性组合，进行降维，然后通过简单的方法将面部匹配到身份。虽然现代方法更复杂，但很多方面仍然依赖于类似的技术。
4. 独立成分分析（Independent Component Analysis，ICA）。ICA是一种统计技术，主要用于揭示随机变量、测量值或信号集中的隐藏因素。

如果你想了解更全的机器学习的算法列表，你可以看一下Wikipedia上的 List of Machine Learning Algorithms。

在 A Tour of Machine Learning Algorithms ，这篇文章带你概览了一些机器学习算法，其中还有一个"脑图"可以下载，并还有一些How-To的文章供你参考。

对于这些算法，SciKit-Learn有一些文档供你学习。

1. Supervised learning

2.3 Clustering

2.5. Decomposing signals in components (matrix factorization problems)

3. Model selection and evaluation

4.3. Preprocessing data

相关资源

对于初学者来说，动手是非常非常重要的，不然，你会在理论的知识里迷失掉自己，这里有篇文章"8 Fun Machine Learning Projects for Beginners"，其中为初学者准备了8个很有趣的项目，你可以跟着练练。

学习机器学习或是人工智能你需要数据，这里有一个非常足的列表给你足够多的公共数据 – 《Awesome Public Datasets》，其中包括农业、生物、天气、计算机网络、地球科学、经济、教育、金融、能源、政府、健康、自然语言、体育等。

GitHub上的一些Awesome资源列表。

Awesome Deep Learning

Awesome - Most Cited Deep Learning Papers

Awesome Deep learning papers and other resources

小结

总结一下今天的内容。我首先介绍了机器学习的基本原理：监督式学习和非监督式学习，然后给出了全世界最简单的入门资料 Machine Learning is Fun!。随后给出了与机器学习密切相关的数据分析方面的内容和资料，然后推荐了深入学习机器学习知识的在线课程、图书和文章等，尤其列举了神经网络方面的学习资料。最后描述了机器学习的十大经典算法及相关的学习资料。

在机器学习和人工智能领域，我也在学习，也处于入门阶段，所以本文中推荐的内容，可能在你看来会有些浅。如果你有更好的信息和资料，欢迎补充。目前文章中给出来的是，我在学习过程中认为很不错的内容，我从中受益良多，所以希望它们也能为你的学习提供帮助。

从下篇文章开始，我们将进入前端知识的学习，包括基础和底层原理、性能优化、前端框架、UI/UX设计等内容。敬请期待。

下面是《程序员练级攻略》系列文章的目录。

分布式架构工程设计(分布式架构)

微服务

容器化和自动化运维

机器学习和人工智能

前端基础和底层原理（前端方向）

前端性能优化和框架（前端方向）

UI/UX设计（前端方向）

技术资源集散地

精选留言 31

斯盖丸
1532611035

网上资料虽多，但质量良莠不齐。靠自己找实在费时费力。左老师帮我们辛辛苦苦挑选出来有什么不好呢？里面照样凝结了作者思路的精华呀。

还是希望左老师可以做自己，不用理会一些个杂音。

songyy
1532581248

期待练级系列尽快到尾声 每篇都是链接的罗列，内容太多需要慢慢消化。但作为读者，总觉得这种优质网上资源可以自己找到，我更希望看到作者自己的东西呀。

因为作者自己的东西，才是网上找不到的，是花钱买专栏的目的

knull
1532609097

最近耗子总贴链接，感觉在交任务。许多人这样想，我也是这许多人中一个。后来，看到耗子哥的回复，我想通了。网上的确都有，而且好多好多，但是耗子哥帮我们筛选了。（百度，Google存在价值不就是网上大量数据中找有用信息么）。所以，谁给的知识不重要，重要的是的确好，的确对我们有用。那就够了。买这课程是学东西，不是来听专场的。

Alan
1532693489

老师，课程有点高大上，能不能接地气点？ 绝大多数人都是码农。

> 作者回复 上6000元的北大青鸟？

yongxiang
1532596710

感谢皓哥的程序员练级攻略，提供了一条由浅入深的学习路径。省去了新手在浩瀚的资料中寻找有用资料的痛苦。真的是压箱底的宝贝，够我学习10年了

空白格
1532666897

作者的内容整理，包括推荐学习资料的由浅入深是很好的，但看到最近的几篇文章都是这样的，就不太好了。机器学习，或者大数据这些想学习的可以去看其他更专业的栏目，希望老师把自己的擅长的内容先整理出来

super
1536283188

耗子哥，说实在的您列的资料太多了。这里面只需要把2-3本书读会了就可以了，ng的，bishop的。即使80%算法岗位的人都没能好好学习这两个人的课程。不列出重点等于白说，这些资料网上都能找到，会误导初学者的。

怀特
1545280076

感觉这些东西，可以写到wiki上，作为百科知识的一种。
一个人看这些东西，看完就老头子了吧。

Wayne
1542584766

资料太多了，根本不知道怎么选。

北极点
1532906459

学习的有个时间成本就是寻找有用的信息！大神能帮我们整理这些，很赞！

aiselo
1532672979

有没有读者群？想加入大部队😁

云学
1532654937

谢谢作者能够把自己看过的认为好的内容分享出来，在信息量爆炸的时代，寻找好的且适合每个学习阶段的资料要花很多时间，真是感谢

多米
1532564539

要做到术业有专攻的好。

Geek_fa497b
1532564119

请问练级攻略系列是到尾声了？

neilyoyoyoyo
1562141846

如果是图像与视觉相关领域的话，斯坦福的cs231n即可帮助绝大多数人入门。

stormrong
1551508800

耗子叔，我是自动化专业，工作两年，目前从事嵌入式开发，想转行后台开发或者人工智能，希望听听您的建议！

凌空
1549080306

耗子叔，怎么看你发的链接地址？想复制出来到电脑上看

godtrue
1547168095

哈哈，有人不亦乐乎，有人疲于奔命，我突然想起来孙悟空学艺的情景了。七十二变还是三十六变，看好奇心啦？

caohuan
1540542592

耗子哥 给的资料太全了，我要的 入门、进阶的视频、书籍 以及数据 全都有，我只需要拿来这个地图，去寻找，省去大量搜索，想问下 耗子哥 怎么做到 拥有这么全的资料库，这得耗费多少的精力、还有怎么做到消化掉它们，牛人的世界 我只能仰望了。

入门阶段 我选择 Machine Learning is fun 和练习8个有趣的项目。

super
1536283292

资料太多了，标出两三本重点即可。ng的，bishop的足够了

gaodai001
1533648707

没想到11年前大学毕业论文聚类分析，尽然可以用在无监督机器学习上啊，当时就是用DBSCAN，去除噪点，岩层分析，用matlab做各种入参的效果图示分析

Ansir
1532757202

阿法狗——监督式，阿法元——非监督式，是吗?

xkyang
1532743042

学习资料非常关键，真正经典的素材只占 1%。

Michael
1532667272

计算机的东西太广了，像前端的东西我一直没做过，也不准备看了，实在学不过来
人工智能，机器学习，是很大一块
目前从事的工作内容跟这些都没关系是该继续深入工作相关的内容，还是抽点时间看看机器学习呢
耗子哥对学习内容取舍方面有什么建议吗？

作者回复 看你自己的好奇心吧

晓冬
1532652823

期待大数据相关的练级攻略，是大数据现在已经不火了么？

ydp
1532637580

希望写写大数据相关

香香
1532624115

谢谢作者，希望能有关于区块链等 关于共识机制的练级攻略， 不知道会不会有？

小肥羊yeah
1532575306

真的是很全面了！

Chang
1532566969

通过Andrew的Coursera入门，深入浅出，强烈推荐！

637
1532565895

有一定机器的基础，想了解Alpha Go使用的算法 和 reinforcement learning的相关知识，皓哥有推荐的文章吗？

铁憨憨
1532565670

很期待前端篇