《机器学习实战》

图书基本信息

书名：《机器学习实战》

13位ISBN编号：9787115317957

10位ISBN编号：711531795X

出版时间：2013-6

出版社：人民邮电出版社

作者：Peter Harrington

页数：332

译者：李锐,李鹏,曲亚东,王斌

《机器学习实战》

内容概要

机器学习是人工智能研究领域中一个极其重要的研究方向，在现今的大数据时代背景下，捕获数据并从中萃取有价值的信息或模式，成为各行业求生存、谋发展的决定性手段，这使得这一过去为分析师和数学家所专属的研究领域越来越为人们所瞩目。
本书第一部分主要介绍机器学习基础，以及如何利用算法进行分类，并逐步介绍了多种经典的监督学习算法，如k近邻算法、朴素贝叶斯算法、Logistic回归算法、支持向量机、AdaBoost集成方法、基于树的回归算法和分类回归树（CART）算法等。第三部分则重点介绍无监督学习及其一些主要算法：k均值聚类算法、Apriori算法、FP-Growth算法。第四部分介绍了机器学习算法的一些附属工具。
全书通过精心编排的实例，切入日常工作任务，摒弃学术化语言，利用高效的可复用Python代码来阐释如何处理统计数据，进行数据分析及可视化。通过各种实例，读者可从中学会机器学习的核心算法，并能将其运用于一些策略性任务中，如分类、预测、推荐。另外，还可用它们来实现一些更高级的功能，如汇总和简化等。

《机器学习实战》

作者简介

Peter Harrington
拥有电气工程学士和硕士学位，他曾经在美国加州和中国的英特尔公司工作7年。Peter拥有5项美国专利，在三种学术期刊上发表过文章。他现在是Zillabyte公司的首席科学家，在加入该公司之前，他曾担任2年的机器学习软件顾问。Peter在业余时间还参加编程竞赛和建造3D打印机。

《机器学习实战》

书籍目录

目　录

《机器学习实战》

章节摘录

版权页： 插图： 7.1.1 bagging：基于数据随机重抽样的分类器构建方法 自举汇聚法（bootstrap aggregating），也称为bagging方法，是在从原始数据集选择S次后得到S个新数据集的一种技术。新数据集和原数据集的大小相等。每个数据集都是通过在原始数据集中随机选择一个样本来进行替换而得到的①。这里的替换就意味着可以多次地选择同一样本。这一性质就允许新数据集中可以有重复的值，而原始数据集的某些值在新集合中则不再出现。 在S个数据集建好之后，将某个学习算法分别作用于每个数据集就得到了S个分类器。当我们要对新数据进行分类时，就可以应用这S个分类器进行分类。与此同时，选择分类器投票结果中最多的类别作为最后的分类结果。 当然，还有一些更先进的bagging方法，比如随机森林（random forest）。有关这些方法的一个很好的讨论材料参见网页接下来我们将注意力转向一个与bagging类似的集成分类器方法boosting。 7.1.2 boosting boosting是一种与bagging很类似的技术。不论是在boosting还是bagging当中，所使用的多个分类器的类型都是一致的。但是在前者当中，不同的分类器是通过串行训练而获得的，每个新分类器都根据已训练出的分类器的性能来进行训练。boosting是通过集中关注被已有分类器错分的那些数据来获得新的分类器。 由于boosting分类的结果是基于所有分类器的加权求和结果的，因此boosting与bagging不太一样。bagging中的分类器权重是相等的，而boosting中的分类器权重并不相等，每个权重代表的是其对应分类器在上一轮迭代中的成功度。 boosting方法拥有多个版本，本章将只关注其中一个最流行的版本AdaBoost。 下面我们将要讨论AdaBoost背后的一些理论，并揭示其效果不错的原因。 7.2训练算法：基于错误提升分类器的性能能否使用弱分类器和多个实例来构建一个强分类器？这是一个非常有趣的理论问题。这里的“ 弱 ”意味着分类器的性能比随机猜测要略好，但是也不会好太多。这就是说，在二分类情况下弱分类器的错误率会高于50%，而“ 强 ”分类器的错误率将会低很多。AdaBoost算法即脱胎于上述理论问题。AdaBoost是adaptive boosting（自适应boosting）的缩写，其运行过程如下：训练数据中的每个样本，并赋予其一个权重，这些权重构成了向量D。一开始，这些权重都初始化成相等值。首先在训练数据上训练出一个弱分类器并计算该分类器的错误率，然后在同一数据集上再次训练弱分类器。在分类器的第二次训练当中，将会重新调整每个样本的权重，其中第一次分对的样本的权重将会降低，而第一次分错的样本的权重将会提高。为了从所有弱分类器中得到最终的分类结果，AdaBoost为每个分类器都分配了一个权重值alpha，这些alpha值是基于每个弱分类器的错误率进行计算的。其中，错误率 ε 的定义为： 而alpha的计算公式如下： AdaBoost算法的流程如图7—1所示。

《机器学习实战》

编辑推荐

《机器学习实战》面向日常任务的高效实战内容，介绍并实现机器学习的主流算法。《机器学习实战》没有从理论角度来揭示机器学习算法背后的数学原理，而是通过“原理简述+问题实例+实际代码+运行效果”来介绍每一个算法。学习计算机的人都知道，计算机是一门实践学科，没有真正实现运行，很难真正理解算法的精髓。这本书的最大好处就是边学边用，非常适合于急需迈进机器学习领域的人员学习。实际上，即使对于那些对机器学习有所了解的人来说，通过代码实现也能进一步加深对机器学习算法的理解。《机器学习实战》的代码采用Python语言编写。Python代码简单优雅、易于上手，科学计算软件包众多，已经成为不少大学和研究机构进行计算机教学和科学计算的语言。相信Python编写的机器学习代码也能让读者尽快领略到这门学科的精妙之处。

《机器学习实战》

精彩短评

1、要具备：Python语言、算法、Hadoop等知识
2、比较入门的书。
3、实战性较强，可以快速看到产出，就是代码丑了点。
4、入门级别的书吧，没有太复杂的数学推倒，代码写得一般般，适合新手对ml的初步了解
5、没有理论，只有实践，所以配合着《统计学习方法》看，收获很大！
6、挺棒的一本书，非常非常基础，只需要掌握基础Python和一些线性代数知识即可，主要是聚类分类预测方面的内容，主要讲了原理，代码还有可优化空间，SVM真的很难懂啊，线性代数还是很重要滴。
7、好痛苦
8、还行，都是些入门的东西，说实话不如上博客园上逛一逛来的快来的实在
9、第一部分介绍了机器学习的基础知识，然后讨论如何使用机器学习算法进行分类；第二部分讨论连续型数值的回归预测问题；第三部分讨论无监督学习；第四部分介绍了机器学习算法使用到的附属工具。
10、不错的一本书。适合入门。也可以顺便熟悉python。但是算法原理基本上没怎么讲！
11、凑合吧 错误太多 胜在有代码 对码农来说天生的亲切感。
12、这本书不错，相对来说比较适合入门。如果想搞机器学习或者数据挖掘，想要了解相关的算法，可以看看这本书。但是这本书讲得并不深，只是把算法简单地用python实现了出来，如果想了解更深刻应该去看相关的论文。
13、作为一本机器学习的书可以打4分，因为介绍的概念还算清晰；作为一本Python的书可以打3分，因为从代码上看感觉作者的编码风格比较奇特
14、在这本书的帮助下,1 小时理解了决策树算法的核心思路,赚了一个 HHKB
15、学完cs229之后看得第一本机器学习的书。主要有代码，可以顺便熟悉一下python，内容上也把最基础的都覆盖到了。
16、这本书，读起来整体给人的感觉一般，（也可能是翻译质量的原因），书中理论部分讲的较为浅显，除此之外的大部分内容都在解释自己的python代码，而且其源代码的质量并不让人那么易懂。感觉使用价值不是太大~ 当然看看还是很有益处的...
17、搞清楚自己想要什么。
18、书中的理论介绍不深，算法相关的代码也是适合于较小数据集的简单实现。但是，对于接触了较多的理论，而不知道从何入手编程实现的读者来说，还是非常值得一读。通过书中的介绍，可以对这些算法的实现思路有初步了解。

本书也是一本很好的入门书籍。

不得不说——翻译得已经算是很好了。
19、厉害了
20、与偏重理论推导过程的书籍不同，本书用通俗易懂的语言描述了机器学习算法的核心思想。然后将大量篇幅用在了以python 为编程语言用机器学习算法来解决实际问题。适合数学不太好，但又希望了解机器学习的同学。推荐一下～
21、不错的书,建议做过IR和数据挖掘的人看
22、注重实战，有很多实用代码，但正如题目一样，理论方面不是重点，所以很多时候不知其所以然，需要与其他理论书籍结合。
23、看过此书的英文版章节，期待中文版！
24、想把算法从matlab里放到实际工程中，这个是不二的选择！
25、适合工程开发人员了解使用
26、原理上的东西写的太浅
27、很好的入门书籍，理论推倒介绍的比较浅，但是没种算法都有代码实现~
28、对学习机器学习帮助极大，不仅停留在公式推导上，对提高代码能力很有帮助。

29、写得很详细,还有的代码,很不错的,推荐!机器学习的好书!
30、理论结合实际
31、学习机器学习很重要的工具，而且是基于python的，很有用
32、好书，只是自己对python数据结构还不够熟练，看着不顺畅；感觉算法里面有错误；再次谢谢作者和译者带来的这本好书
33、数据挖掘入门的不错的书籍，看着代码，比较好可以了解到算法的本质。唯一缺陷的是有些要点没展开论述，对于某些数学功底差的朋友可以比较苦逼了，同事推荐下吴恩达的公开课。
34、內容不錯然而代碼部分有點小坑啊。。。
35、代码乱成一团，我也是服。。
36、翻译太渣
37、用代码来说明问题，讲的很好的。
38、看到一半了可以当课外书籍看 写得简单易懂 很有用~
39、里面机器学习的经典算法基本都用Python实现了一遍，但是没有深网的内容
40、如其名 比较实用 不过理论不强 配合统计学习方法食用尤佳
41、攒灰一年，为了写Adaboost的作业才拿出来，本打算靠它迅速解决(抄完)作业，没想到书里的代码太坑，不重写一遍简直难受，最后作业自己做的。。
42、代码不能认真给注释吗
43、有人说这本书不好，不科学，其实我觉得还是不错的。对于我这种编程小白来说，真是手把手教我怎么用python。如果想要科学性可以看李航的《统计学习方法》和周志华的《机器学习》，这本书既然叫“实战”了，确实是偏重实战一些。理论上跳跃比较大，可以看其他书作为补充。
44、学习机器学习，看了目录就买了，回来看了一点，很不错，推荐
45、例子相当完整，讲解清晰，翻译也不错，获益良多。
46、作为python处理的机器学习方面的书，算是很好的。
47、机器学习入门必备书籍推荐啊！虽然只有这一本完全不够，可是没有这一本入门会比较困难，这本书算是降低机器学习入门门槛的实用书籍。别指望着这一本书就能让你成为高手，但是这本书很适合初学者
48、python好学，用python来编写机器学习算法比用别的含有静态类型检查的语言需要考虑的东西更少。这本书对于算法的讲解水准和其他书区别不大，但详细的代码解释是很大的亮点，跟着敲下来，无论是对python本身还是对相应的学习算法的理解都会加深不少。
49、一个我都差不多能YY出来啥意思的讲代码的书，大概真的很适合初学者了，赞一个
50、深入浅出，很好的书，正好用的上。
51、历时1个月，终于读完。问自己，当初在学校的时间都在干嘛呢？全书分为4个部分，分别是分类（有监督学习，包括KNN/决策树/朴素贝叶斯/逻辑斯蒂回归/svm/改变样本权重的bagging和adaboosting）、回归（有监督学习，线性回归、局部加权、特征维度比样本个数多时缩减系数，如岭回归、lasso等，树回归，这块掌握不太好）、无监督学习（kmeans、apriori/fp-growth）以及其他工具（PCA/SVD/MAPREDUCE）。基本上都比较清楚了，过段时间再刷一遍代码吧
52、挺好的 不过性能上没考虑到
53、……又看了一半（大雾）
54、不适合小白。。。。结合《统计学习方法》食用，效果更佳～～
55、talk is cheap,show me the code
56、很不错，值得一读，是一本好书
57、这本书真的对于机器学习爱好者很好，而且是用python实现，对于初学者很有意义
58、学没学过线性代数都忘了，基本上没一点数学底子了，这尼玛怎么学机器学习，读完也就刚了解个概念（悲伤脸）
59、很好的一本书适合大家学习
60、评论里说这本书理论太强的一定是在逗。这本书简直是python和numpy的小教程。要深入原理的话还是配着别的书一起看比较好。作者字里行间一些憋不住的冷笑话还是蛮好笑的。
61、以实际例子讲解主要的机器学习算法，适合engineer背景的机器学习的上手
62、认真的说，这本书写得不是特别好，但是比较适合入门了。由于原理没有说清楚，代码写得也一

般（大量的缩写变量名，没有注释），所以我对很多算法都没有理解。接下来看周志华的《机器学习》
63、书中将所有的相关算法都使用python语言实现，对算法原理简要介绍，但是对于数学公式没有进行推导，建议配合周志华的《机器学习》一起看。
64、所谓的实战没有意义，理论也没有讲清楚
65、这本书很适合希望从理论回归实践的研究人员，将理论算法用python代码完成，不过书上代码质量很差，还是得自己多练练手。
66、书很不错，里面的理论知识很好，值得 一看
67、贝叶斯的那个例子很典型。 机器学习十大算法
68、好难懂啊。。。
69、我。终于把所有的代码都写了一遍！结果发现软肋是数学，又要开始恶补数学了！
70、处于原理和直接sdk之间的自己实现算法. 不去看数学原理的书, 看这个真的没什么用.

《机器学习实战》

精彩书评

1、尽管评论里对这本书褒贬不一，我觉得这些都是根据每个人不同的能力背景出发而给的评论。而对于我这样能力的人来说，这本书可以说是最适合了。我是什么能力状况呢，计算机专业背景，有那么几年开发经验，但是机器学习方面是小白。看这本书需要一定的编程经验，但不需要很强，想我这样就行（不经常写代码）。书中的代码示例一般都不长，很好理解。本书也需要一定的数学知识，主要是线性代数和概率论，但一样不用很熟。想我这样毕业n年，忘了差不多的人，看一下附录里的知识点温故而知新，就没问题了。很多人说书太理论性，其实这是我看的机器学习资料中，最接近实践的了。书上的内容写的很清楚，个人觉得结合例子一起看，也很易懂。很崇拜作者怎么能把这么一个外行人看来如此复杂的知识写的那么清楚。译者的水平也很高，很多技术书籍的译者我就不说了，看的那个累啊，读数时都是偏科生。但是这本书不错。我个人建议，是看一遍中文版的书先，然后再看遍英文版的，巩固下知识，也有助于了解英文中的名词，帮助你将来深入看英文资料。总的来书，本书还算是比较适合机器学习初学者的，个人非常推荐。大牛就可以飘过了。
2、这本书最大的优点在于有源码实现，很赞，但是理论部分太差了，看了逻辑回归和支持向量机两章，发现好多理论都没讲，就比如逻辑回归中的Cost函数都没说，如果不了解，源码读起来也是一头雾水，所以对于初学者还需要一本理论较强的书，推荐李航博士的统计机器学习方法，刚好配套～
3、为什么我会力荐这本书？也许书中分类器都非常的简单，数学理论都非常的粗浅（为了看明白书中SVM分类器的训练过程，不得不去复习了二次凸优化解法，自己推导被作者略去的中间过程），算法测试也只在轻量级的数据集上完成。不过，大可不必像其他评论一样对贬低本书。聪明的读者会知道自己没有什么，自己需要学习什么。如果更加喜欢背后深奥的统计学理论和凸优化理论，可以去看《Machine Learning: A Probabilistic Perspective》，如果对自己的数学水平足够自信的话。这本书能让你明白：那些被吹捧得出神入化的分类算法，竟然实现起来如此简单；那些看是高深的数学理论，其实一句话就能道明其本质；一切复杂的事物，出发点都是非常简单的想法。我说不出这本书适合什么样的读者，但是却明白它不适合谁：学过一点机器学习或者模式识别或者数据挖掘，完全不具备统计推断和凸优化知识，又想找一条捷径，想从菜鸟摇身一变成大师的人；对编程不感兴趣的人，或者没有动手实践习惯的人；不喜欢独立思考，希望别人把答案摆在自己面前的人。祝君学运昌盛
4、这本书基本上是基于一个例子讲解一种机器学习算法，但是朴素贝叶斯那一章就存在重大错误了！书页眉下面标注使用伯努利模型，但计算条件概率那段代码却是混合使用伯努利模型与多项式模型，网上流传已久的代码与算法描述页都是错误的，不知道为什么只有几个人提到这个错误了
5、本书强调的是机器学习算法的Python实现，并未深入涉及这些算法的数学证明或推演。理解机器学习的算法本书就需要基本的数学基础。个人感觉还是先找本机器学习的书籍理解这些算法数学原理，然后在根据这本书编写Python代码，有助理解算法精髓。
6、不夸夸其谈概念理论，有理论也有实战，对于熟悉机器学习各种算法有很大的帮助作用。虽然是一本基础类书籍，但是读了之后还是收获较大。一直在找实用性较高的机器学习方面的书，这本算是找到了。书中对算法的讲解简单清晰有条理，实战的例子也选得很恰当。
7、很好很强大，5块钱你买不了上当，5块钱你买不了吃亏。本书情节跌宕起伏，个别章节少儿不宜，需参看大神博客，才能打通。比如“ svm三层境界 ”等。至于实战嘛，这是必须的，有完整的python代码和现成的数据供你把玩。比什么《推*系统实战》丰富精彩多了。
8、理论没讲太明白，直接上算法，甚至还有公式缺失，代码不敢恭维就像大家说的一样 先看看线性代数、概率论、统计学再来看看这书吧我这10多年 php、java、c#、js通吃，本想python应该不难，竟然代码部分有东西看不懂了，不得不拿起本python的书对着看...
9、特别适合新手，特别适合新手，特别适合新手。长度适中，举例形象，概念浅显通俗。难得有一个条理清楚 逻辑不迷糊 不堆砌代码打哈哈的书。基于这个理由bonus给五星，以后给别人推荐就这本了。尤其是前面几章，介绍机器学习的基本概念。作者给我们指明了一个做ML的基本要求：“ 机器学习的目的在于提炼数据背后的隐含规律。你们应当让机器说人话，而不是说鬼话 ”。我觉得对于初学者这种概念非常有意义。可惜冲着过来看这本书的都是ML expert，根据书名和期望值就匆匆打了个低分就离开了。
10、这本书的最大好处是让你能够用最基本的pyton语法，从底层上让你构建代码，实现我们常说的比如邮件过滤，数据分类的应用。很多时候你要写最基本的代码和结构去做这些工作，而不是像kaggle

的tutorial或者其他的工程大多数告诉你一个lib库函数去调用，你能看到底层在干什么，决策树是怎么弄出来的，gradient descent是怎么弄出来的，知道机器学习是如何从低实现的。缺点就是评论上面说的各种，理论上不严密，其实这个有错误到是其次，主要是理论讲的不是很清楚不是很透彻。需要参考一些公开课老师讲的或者一些比较理论上的书籍，你才会发现原来作者不声不响加上的一些东西是这么来的。 有些作者写的机器学习的代码也是只对那个样本有效，遇到一般性的样本就会出现问题。但是总体来说，我觉得这本书还是告诉了我们机器学习从代码上是个什么样子。的确，它在理论上欠缺，而且讲的逻辑不是很清楚，但是市面上根本不缺理论的书，这本书有自己独特的位置。开始你按照它的代码走，到后来你觉得，诶这哥们写的代码有点问题啊，不过到最后还是得感谢他领你入门了。我觉得他的利还是大于弊，不过最好结合理论书和理论的公开课来一起学习，这样互补性很好。
11、现在刚读到第三章，决策树，感觉这本书主要是给出代码，并对代码作出解释，而对背后的数学原理讲解很少， 个人感觉读代码其实就已经知道干了什么，只是那些说明可以帮助理解代码，同时指明了代码阅读顺序。 所以这本书需要结合其他讲解相关算法的资料一起看。不过这正好让人更加直观的理解那些统计学知识了

《机器学习实战》

版权说明