<<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

出版时间：人民邮电出版社

作者：Peter Harrington

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。

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

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。

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。

版权说明