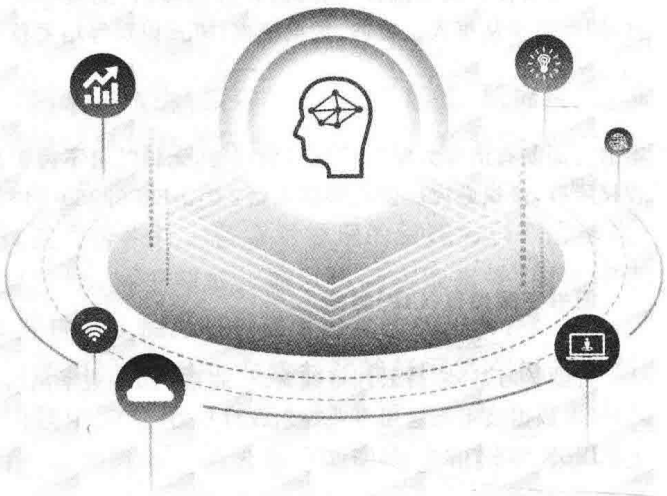
大数据技术与应用专业规划教材

机器学习基础

◎ 吕云翔 马连韬 刘卓然 张 凡 张程博 编著

清華大學出版社

大数据技术与应用专业规划教材



机器学习基础

◎ 吕云翔 马连韬 刘卓然 张 凡 张程博 编著

清華大學出版社
北京

内 容 简 介

本书全面系统地介绍了机器学习的基本概念、预备知识、主要思想、研究进展、基础技术、应用技巧，并围绕当前机器学习领域的热点问题展开讨论。全书共 11 章，主要内容包括决策树、神经网络、支持向量机、遗传算法、回归、聚类分析等。

本书可作为高等院校计算机、软件工程、智能科学与技术等专业研究生和高年级本科生的教材，同时对于从事人工智能、数据挖掘、模式识别等相关技术人员也具有较高的参考价值。

图书在版编目(CIP)数据

机器学习基础/吕云翔等编著. —北京：清华大学出版社，2018
(大数据技术与应用专业规划教材)
ISBN 978-7-302-49659-5

Ⅰ. ①机… Ⅱ. ①吕… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2018)第 033867 号

责任编辑：魏江江　薛　阳
封面设计：刘　键
责任校对：时翠兰
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954
印 装 者：三河市君旺印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10.75　　**字　　数**：226 千字
版　　次：2018 年 11 月第 1 版　　**印　　次**：2018 年 11 月第1次印刷
印　　数：1～1500
定　　价：29.80 元

产品编号：066058-01

前言

本书全面系统地介绍了机器学习的基本概念、预备知识、主要思想、研究进展、基础技术、应用技巧，并围绕当前机器学习领域的热点问题展开讨论。章节安排由浅入深，涵盖回归问题、分类问题、监督学习、无监督学习。具体内容包括决策树、神经网络、支持向量机、遗传算法、集成学习、聚类分析等。各章对原理的叙述力求概念清晰、表达准确，突出理论联系实际，富有启发性，易于理解。辅以代码实践指导，引领读者快速迈进机器学习领域，通过动手实践进一步加深对机器学习算法的理解。

本书注重对数学分析方法和理论的探讨，而且也非常关注神经网络在模式识别、信号处理以及控制系统等实际工程问题中的应用。它完美结合了基础理论与应用实践，可作为高等院校计算机、软件工程、智能科学与技术等专业研究生和高年级本科生的教材，同时对于从事人工智能、数据挖掘、模式识别的相关技术人员也具有较高参考价值。

大数据时代是机器学习最美好的时代。希望本书不仅可以帮助读者深入理解机器学习的概念，在理论分析与实际应用技术的结合中，掌握主流解决方案，更能以一种全新的视角理解在实际软件工程中机器学习的总体思想，在人工智能的大时代中夺得先机！

本书的作者为吕云翔、马连韬、刘卓然、张凡、张程博，另外，曾洪立、吕彼佳、姜彦华进行了素材整理及配套资源制作等。由于机器学习是一门新兴学科，机器学习的教学方法本身还在探索之中，加之作者的水平和能力有限，书中难免存在疏漏之处，恳请各位同仁和广大读者给予批评指正。也希望各位能将实践过程中的经验和心得与我们交流(yunxianglu@hotmail.com)。

作　者

2018 年 3 月

于北京航空航天大学

目 录

第1章

绪　论

1.1　从两个问题谈起

问题一：人工智能、知识工程、机器学习、神经网络、深度学习、数据挖掘，它们是什么关系？

人与动物根本的区别在于是否拥有智能。日常生活中，人们一直在本能地使用非常复杂而又高效的智能算法——识别出同学的长相、根据云的形状预测天气、把要传达的信息组织成一句话等。当我们希望机器也能聪明地完成类似的事情时，就需要利用**人工智能**(Artificial Intelligence，AI)。

人工智能中首先包括**知识工程**(Knowledge Engineering)，即根据已有知识，利用规则去解决问题。例如，我们写一段程序规定，如果鼻子眼睛之间的距离超过一个值，那么就识别为某个人的脸。如果我们把世界上人类的知识都转化为规则，那是不是就诞生了全知全能的 AI？但显然我们无法穷举规则。

机器学习(Machine Learning)是人工智能的另一部分，也是核心技术。其利用经验，建立统计模型、概率模型，去解决问题。具体地讲，机器学习就是对某个实际问题建立**计算模型**(Computational Model)，并利用已知的经验(Experience)来提升模型效果(Performance)的一类方法。我们经常听到的贝叶斯、神经网络、支持向量机都是机器学习的工具。当我们要处理、分析的数据中存在一定模式，我们想把其中的知识写成规则、形式化地确定下来，但又无法穷尽时，就可以尝试机器学习的方法。比如把医生多年学习、工作的经验知识，确定为一个模型，来进行疾病诊断。

机器学习方法在大型数据库中的应用称为**数据挖掘**(Data Mining)。在数据挖掘中，需要处理大量的数据以构建简单有效的模型，如具有高精度的预测模型。具体应

用如：零售业中分析历史数据，来构建市场应用模型；制造业中的学习模型用于故障检测；物理学、天文学、生物学中的海量数据分析等。

机器学习中，受到人脑神经元认知原理的启发，人们设计了人工**神经网络**(Neural Network，NN)。利用数据不断训练得到一个模型，将输入映射为输出。研究者们从数学上证明，多层嵌套的神经网络配合非线性激活函数可以模拟任意连续函数。当神经网络最早提出时，人们非常兴奋，因为如此简单的模型却能干很复杂的事情。大家认为机器学习全新的时代来临了，真正的人工智能即将实现。但是随着研究进展，大家发现，由于计算资源和数据的限制，网络做不大，在人工给定特征时，性能上还是比不过传统机器学习模型。

在神经网络最早提出时，隐藏层的层数很少。随着研究进展，人们发现层数的增加对提升网络模型能力非常有帮助。这样的模型提供了一个层次建模的功能，可以对输入的数据逐层提取特征。同时，神经网络研究者们的思路发生了变化，他们指出现代人工智能的关键是表示学习，希望神经网络能实现从数据到表示，即深度学习(Deep Learning)。深度学习相对于传统神经网络，可以简单理解为多层网络的堆叠。

传统机器学习方法在做图像分类识别时，需要研究者提供人工指定的特征值。而对于深度学习方法，我们可以直接把原始的图像提供给网络。图像中包含分类所需要的全部原始信息，网络将自己在训练过程中调整权重，学习使用怎样的特征来描述原始输入，把经验固化在网络。21 世纪大数据、云计算的背景让这个思路得以实现。

需要注意的是，截至目前，脑科学作为一个逐渐探索的领域，还没有人能完全回答人类神经元的工作机理，NN 的模型也还十分简单，且 AI 的发展并不是主要由认知推动的，数学等相关技术才是 AI 真正有效的手段。AI 的发展对认知也有一定的推动作用。虽然也许未来随着人们对脑科学的认识更加深入，AI 的能力会得到提升，但现在作为机器学习的相关研究者，你并不需要非常了解人脑认知。

到目前为止，在有监督学习方面，深度学习几乎超越了任何其他传统方法。在数据量大的领域，特定场景、特定需求的弱人工智能将会很快成为主流。但是现在我们距离“终结者”那样的人工智能还非常遥远。

人工智能拓展故事：Google DeepMind AlphaGo

2016 年，关于谷歌“阿尔法狗”的新闻曾经刷爆了大家的屏幕。

2014 年 4 月到 2015 年 9 月，AlphaGo 以英国棋友“DeepMind”的名义在弈城围棋网上对弈，水平维持在职业七段到八段之间。2015 年 9 月 16 日首次上升到职业九段。2015 年 10 月，分布式版 AlphaGo 以 5∶0 击败了欧洲围棋冠军、华裔法籍棋士樊麾。这是计算机围棋程序第一次在十九路棋盘且分先的情况下击败职业围棋棋手。2016 年 3 月，AlphaGo 挑战世界冠军、九段棋士李世乭，并以 4∶1 取得胜利。这次对战在网络上引发了人们对人工智能的广泛讨论。2016 年 7 月，世界职业围棋排名网站 GoRatings 公布最新世界排名，AlphaGo 以 3612 分，超越 3608 分的柯洁成为新的

世界第一。2016年12月到2017年1月，AlphaGo以“Master”名义注册弈城围棋网和腾讯野狐围棋网，以60战全胜的战绩击败中、日、韩顶尖围棋高手。

AlphaGo最初通过模仿人类玩家，尝试匹配职业棋手的过往棋局，其数据库中约含3000万步棋着。一旦它达到了一定的熟练程度，它就开始和自己对弈大量棋局，使用强化学习进一步改善自身。

一盘围棋平均约有150步，每一步平均约有200种可选的下法。围棋的分支因子大大多于国际象棋等其他游戏，计算机要在围棋中取胜比在其他游戏中取胜要困难得多。诸如暴力搜寻法、Alpha-Beta剪枝、启发式搜索等传统人工智能方法在围棋中很难奏效。

AlphaGo结合深度神经网络与蒙特卡洛树搜索算法，根据大量人类对弈棋局，模拟人类围棋下法，让人工智能算法学会如何评估棋局、选择落子。

当AlphaGo训练达到一定水平后，它将会进行大量自我对弈，利用增强学习进一步提升实力，超越人类已有围棋经验。这使得AlphaGo的棋着从“看起来像人类高手”提升至“人类无法完全理解”的境界。

问题二：为什么需要机器学习？

人类总是希望并善于借用外部的力量来替代自己。工业化时代的人类使用各种机械及电气装置将自己从重复的体力劳动中解放出来；到了20世纪，人类建造了可编程的电子计算机，并把各种计算、推理规则编码到计算机里，使得简单重复的脑力劳动可以被替代。

但是当人类试图让机器朝着更加自动化、智能化的方向发展时，却发现许多并非传统算法可以解决的问题：现实世界的运行方式并不总是可以总结、提炼成能编码到计算机里的规则的。这是因为一方面，某些时候这种规则是潜在的、难以使用严格的数学方法定义的，例如语音识别；另一方面，某些问题是人类自身也难以解决，而寄希望于机器强大的计算能力来解决的，例如医学诊断。

我们把实际问题抽象成其一般形式：给定问题的场景设定作为**输入**，馈送(Feed)到某个模型中，并随后从这个模型中得到反馈(Feedback)回来的解决方案作为**输出**。将输入记为 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$，输出记为 $\boldsymbol{y}=\{y_1,y_2,\cdots,y_n\}$，则模型为一个从输入空间到输出空间的映射：

$$f(\boldsymbol{x})\rightarrow \boldsymbol{y}$$

所有需要求解的实际问题都可以归纳到以上形式。例如，在手写数字识别中，$\boldsymbol{x}$是手写数字的图片，$\boldsymbol{y}$是识别出来的数字；在统计机器翻译中，$\boldsymbol{x}$是源语言的一个句子，$\boldsymbol{y}$是目标语言句子的条件概率。而机器学习要做的事情，就是要利用已知的经验来优化模型 $f(\cdot)$的效果。这些经验以**观测样本点**(Observed Samples)集合的形式出现，观测样本点构成的集合称为**数据集**(Dataset)。

通常将优化一个模型的过程称为**训练**(Training)或者**学习**(Learning)；检验模型效果的过程称为**测试**(Testing)；若该模型为参数化的模型，则还需要通过对不同参数下的模型表现进行检验来选择模型参数，这个过程称为**开发**(Development)或者**验**

证(Validation)。我们通常把所得到的数据集划分为互不相交的几个集合：数据集中绝大多数样本点被用于训练，这些样本点的集合称为**训练集**；剩余的少量样本点用于测试，这些样本点的集合称为**测试集**；若需要选择模型参数，我们还需要与测试集的大小相仿的少量样本点的集合作为**开发集**。

在实际场景中应用机器学习方法时，首先需要回答以下两个问题。

(1) 选择何种模型？

(2) 如何最优化该模型？

本书致力于向读者介绍一些常见的机器学习模型以及它们的最优化算法，使得读者在了解这些机器学习算法的原理后，在实际应用场景中能够选择恰当的模型和算法来解决实际问题。

1.2 模型评估与模型参数选择

如何评估一些训练好的模型并从中选择最优的模型参数？若对于给定的输入 $\boldsymbol{x}$，某个模型的输出 $\hat{\boldsymbol{y}}=f(\boldsymbol{x})$ 偏离真实目标值 $\boldsymbol{y}$，那么就说明模型存在误差；$\hat{\boldsymbol{y}}$ 偏离 $\boldsymbol{y}$ 的程度可以用关于 $\hat{\boldsymbol{y}}$ 和 $\boldsymbol{y}$ 的某个函数 $L(\boldsymbol{y},\hat{\boldsymbol{y}})$ 来表示，作为误差的度量标准，这样的函数 $L(\boldsymbol{y},\hat{\boldsymbol{y}})$ 称为损失函数。

在某种损失函数度量下，训练集上的平均误差被称为训练误差，测试集上的误差称为**泛化误差**。由于我们训练得到一个模型的最终目的是为了在未知的数据上得到尽可能准确的结果，因此泛化误差是衡量一个模型泛化能力的重要标准。

之所以不能把训练误差作为模型参数选择的标准，是因为训练集可能存在以下问题：①训练集样本太少，缺乏代表性；②训练集中本身存在错误的样本，即**噪声**。如果片面地追求训练误差的最小化，就会导致模型参数复杂度增加，使得模型**过拟合**(Overfitting)，如图 1.1 所示。

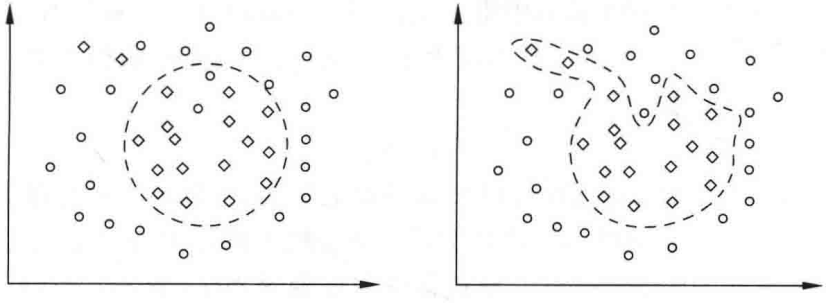


图 1.1 拟合与过拟合

为了选择效果最佳的模型，防止过拟合的问题，通常可以采取的方法如下：

(1) 使用验证集调参；

(2) 对损失函数进行正则化。

1.2.1 验证

模型不能过拟合于训练集，否则将不能在测试集上得到最优结果；但是否能直接以测试集上的表现来选择模型参数呢？答案是否定的。因为这样的模型参数将会是针对某个特定测试集的，那么得出来的评价标准将会失去其公平性，失去了与其他同类或不同类模型相比较的意义。

这就好比我们要证明某一位学生学习某门课程的能力比别人强（模型算法的有效性），那么就要让他和其他学生听一样的课、做一样的练习（相同的训练集），然后以这些学生没做过的题目来考他们（测试集与训练集不能交叉）；但是如果我们直接在测试集上调参，那就相当于让这个学生针对考试题目来复习，这样与其他学生的比较显然是不公平的。

因此参数的选择（即**调参**）必须在一个独立于训练集和测试集的数据集上进行，这样的用于模型调参的数据集被称为**开发集**或**验证集**。

然而很多时候我们能得到的数据量非常有限。这个时候可以不显式地使用验证集，而是重复使用训练集和测试集，这种方法称为**交叉验证**。常用的交叉验证方法如下。

（1）简单交叉验证。在训练集上使用不同超参数训练，使用测试集选出最佳的一组超参数设置。

（2）K-重交叉验证（K-fold Cross Validation）。将数据集划分成 K 等份，每次使用其中一份作为测试集，剩余的为训练集；如此进行 K 次之后，选择最佳的模型。

1.2.2 正则化

为了避免过拟合，需要选择参数复杂度最小的模型。这是因为如果有两个效果相同的模型，而它们的参数复杂度不相同，那么冗余的复杂度一定是由于过拟合导致的。为了选择复杂度较小的模型，一种策略是在优化目标中加入**正则化项**，以惩罚冗余的复杂度：

$$\min_{\theta} L(\boldsymbol{y},\hat{\boldsymbol{y}};\boldsymbol{\theta})+\lambda \cdot J(\boldsymbol{\theta})$$

其中，$\boldsymbol{\theta}$ 为模型参数，$L(\boldsymbol{y},\hat{\boldsymbol{y}};\boldsymbol{\theta})$ 为原来的损失函数，$J(\boldsymbol{\theta})$ 是正则化项，λ 用于调整正则化项的权重。正则化项通常为 $\boldsymbol{\theta}$ 的某阶向量范数。

1.3 机器学习算法分类

模型与最优化算法的选择，很大程度上取决于我们能得到什么样的数据。如果我们能得到的数据集中，样本点只包含模型的输入 $\boldsymbol{x}$，那么就需要采用非监督学习的算

法；如果这些样本点以$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$这样的输入-输出二元组的形式出现，那么就可以采用监督学习的算法。

1.3.1 监督学习

在监督学习中，我们根据训练集$\{\langle \boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)} \rangle\}_{i=1}^{N}$中的观测样本点来优化模型$f(\cdot)$，使得给定测试样例$\boldsymbol{x}'$作为模型输入，其输出$\hat{\boldsymbol{y}}$尽可能接近正确输出$\boldsymbol{y}'$。

监督学习算法主要适用于两大类问题：回归和分类。这两类问题的区别在于：回归问题的输出是连续值，而分类问题的输出是离散值。

1. 回归

回归问题在生活中非常常见，最简单的例如一个连续函数的拟合。

回归问题中通常使用均方损失函数来作为度量模型效果的指标，最简单的求解例子是最小二乘法。

第2章将介绍常见的几种回归模型。

2. 分类

分类问题也是生活中非常常见的一类问题，例如我们需要从金融市场的交易记录中分类出正常的交易记录以及潜在的恶意交易。

度量分类问题的指标通常为**准确率**(Accuracy)：对于测试集中D个样本，有k个被正确分类，$D-k$个被错误分类，则准确率为：

$$\text{Accuracy} = \frac{k}{D}$$

然而在一些特殊的分类问题中，属于各类样本的值并不是均一分布，甚至其出现概率相差很多个数量级，这种分类问题称为**不平衡类问题**。在不平衡类问题中，准确率并没有多大意义。例如，检测一批产品是否为次品时，若次品出现的频率为1%，那么即使某个模型完全不能识别次品，只要每次都“蒙”这件产品不是次品，仍然能够达到99%的准确率。显然我们需要一些别的指标。

通常在不平衡类问题中，我们使用***F*-度量**来作为评价模型的指标。以二元不平衡分类问题为例，这种分类问题往往是异常检测，模型的好坏往往取决于能否很好地检出异常，同时尽可能不误报异常。定义占样本少数的类为**正类**(Positive Class)，占样本多数的为**负类**(Negative Class)，那么预测只可能出现以下4种情况。

(1) 将正类样本预测为正类(True Positive，TP)；

(2) 将负类样本预测为正类(False Positive，FP)；

(3) 将正类样本预测为负类(False Negative，FN)；

(4) 将负类样本预测为负类(True Negative，TN)。

定义召回率(Recall)为：

$$R=\frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FN}|}$$

召回率度量了在所有的正类样本中模型正确检出的比率，因此也称为**查全率**。

定义**精确率**(Precision)为：

$$P=\frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FP}|}$$

精确率度量了在所有被模型预测为正类的样本中正确预测的比率，因此也称为**查准率**。

F-度量则是在召回率与精确率之间取调和平均数；有时候在实际问题上，若更加看重其中某一个度量，还可以给它加上一个权值 α，称为F_α-度量：

$$F_\alpha=\frac{(1+\alpha^2)RP}{R+\alpha^2 P}$$

特殊地，当 $\alpha=1$ 时，有：

$$F_1=\frac{2RP}{R+P}$$

可以看到，如果模型“不够警觉”，没检测出一些正类样本，那么召回率就会受损；而如果模型倾向于“滥杀无辜”，那么精确率就会下降。因此较高的 F-度量意味着模型倾向于“不冤枉一个好人，也不放过一个坏人”，是一个较为适合不平衡类问题的指标。

可用于分类问题的模型很多，例如 Logistic 回归分类器、决策树、支持向量机、感知机、神经网络等。本书将在第 2、4、5 章和第 7 章对以上算法进行介绍。

1.3.2 非监督学习

在非监督学习中，我们的数据集 $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ 中只有模型的输入，而并不提供正确的输出$\boldsymbol{y}^{(i)}$作为监督信号。

非监督学习通常用于这样的分类问题：给定一些样本的特征值，而不给出它们正确的分类，也不给出所有可能的类别；而是通过学习确定这些样本可以分为哪些类别、它们各自都属于哪一类。这一类问题称为**聚类**，将在第 9 章中介绍。

非监督学习得到的模型的效果应该使用何种指标来衡量呢？由于通常没有正确地输出 $\boldsymbol{y}$，我们采取一些其他办法来度量其模型效果。

(1) 直观检测，这是一种非量化的方法。例如，对文本的主题进行聚类，可以在直观上判断属于同一个类的文本是否具有某个共同的主题，这样的分类是否有明显的语义上的共同点。由于这种评价非常主观，通常不采用。

(2) 基于任务的评价。如果聚类得到的模型被用于某个特定的任务，可以维持该任务中其他的设定不变，而使用不同的聚类模型，通过某种指标度量该任务的最终结果来间接判断聚类模型的优劣。

(3) 人工标注测试集。有时候采用非监督学习的原因是人工标注成本过高，导致

标注数据缺乏，只能使用无标注数据来训练。在这种情况下，可以人工标注少量的数据作为测试集，用于建立量化的评价指标。

习题

1. 除了本章中提到的方法，还有什么办法可以防止过拟合的发生？
2. 是否训练数据量越大，越能得到良好的模型？为什么？