HZ BOOKS
华章IT

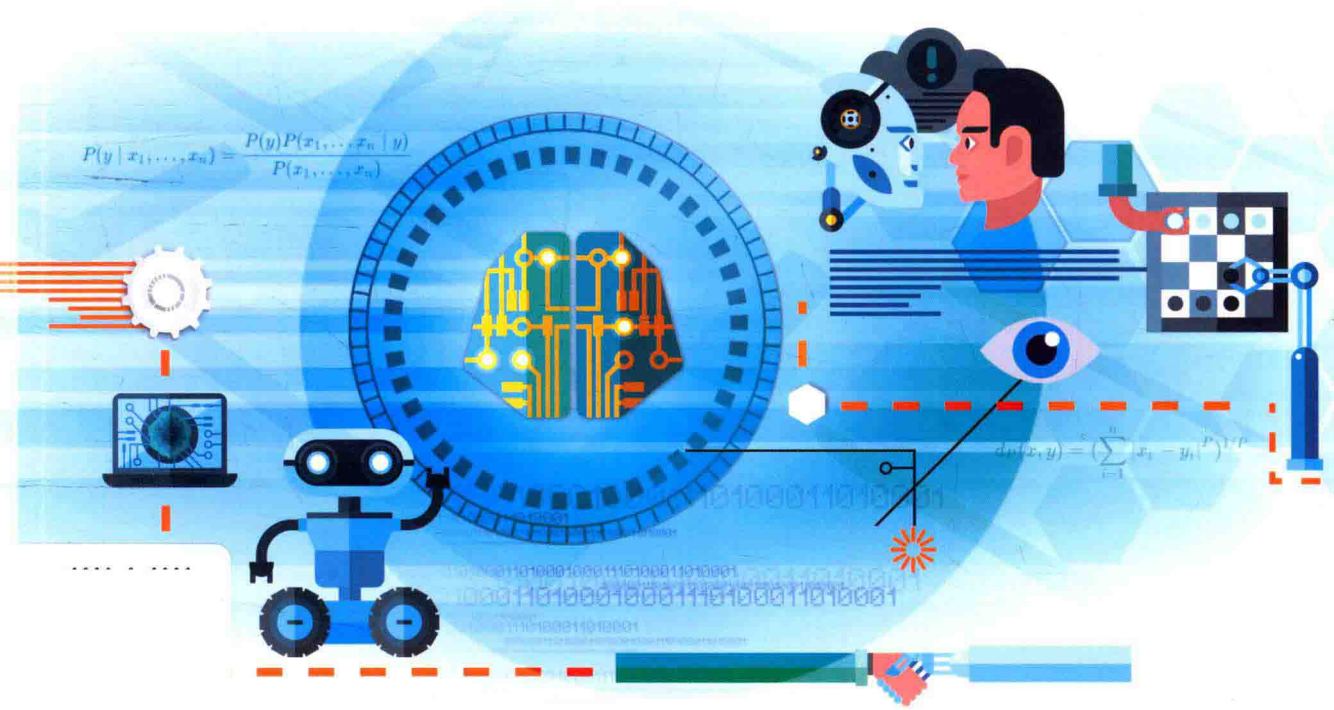
智能系统与技术丛书

从生活案例中理解算法，发现算法的乐趣，再把算法应用到机器学习中。
零基础掌握算法精髓，快速进入人工智能开发领域。

The First Book of Machine Learning

机器学习算法的数学解析与Python实现

莫凡 编著



机械工业出版社
China Machine Press

智能系统与技术丛书

The First Book of Machine Learning

机器学习算法的数学解析与Python实现

莫凡 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习算法的数学解析与 Python 实现 / 莫凡编著 . —北京：机械工业出版社，2020.1
（智能系统与技术丛书）

ISBN 978-7-111-64260-2

I. 机… II. 莫… III. ①机器学习 - 算法 ②软件工具 - 程序设计 IV. ①TP181
②TP311.561

中国版本图书馆 CIP 数据核字（2019）第 269887 号

机器学习算法的数学解析与 Python 实现

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：佘　洁　　责任校对：殷　虹
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2020 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：13.5
书　　号：ISBN 978-7-111-64260-2　　定　　价：89.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

PREFACE

前　言

这是一本介绍机器学习的书，按常理来说，我应该首先介绍学习机器学习的重要性。可是，有必要吗？我记得约五年前，机器学习还是一个很有科幻色彩的术语，而现在技术学习圈几乎整版都换成了机器学习的各种模型，国内很多大学已经开始设立人工智能专业，机器学习当仁不让地成为核心课程。据说相关学者已经将该知识编制成课本，即将走入中学课堂。机器学习的火热，连带着让长年不温不火的 Python 语言也借机异军突起，甚至掀起一阵 Python 语言的学习热潮。机器学习已经成为“技术宅”的一种必备技能，因此，实在没必要再占篇幅介绍它的重要性。

但是，学习机器学习的路途是坎坷和颠簸的，唯一不缺的就是让你半途而废的借口。机器学习今日的成就是站在巨人的肩膀上取得的，因此，当你终于下定决心学习机器学习时，很多人会给你开出一串长长的学习清单：机器学习涉及大量向量和矩阵运算，所以线性代数是肯定要学的；机器学习的很多模型算法都以统计知识作为背景，所以统计学和概率论也是必修的；许多重要环节依赖微分运算，那本好久不看的《高等数学》是不是到了重出江湖的时候了？

想想看，如果告诉你学习机器学习，首先得把《线性代数》《概率统计》《高等数学》统统翻一遍，然后你才只是刚刚摸到学习机器学习的起跑线，如果不擅长数学，你得需要多大的毅力才能坚持下来，把机器学习学明白？

真的很难，如果开始学习机器学习时我就知道后面会承受这么多“痛苦”，也许我根本就不会开始。特别是如果你也是利用业余时间来自学机器学习，那么真的称得上煎熬：当你已经为别的事情绞尽脑汁，好不容易有了那么一点属于自己的时间，想要学习充电时，结果鼓起勇气翻开书本，扑面而来的全是各种难以理解的数学公式和闻所未闻的专业术语，你就能立即体会到什么是无力感。

那时我总是在想，能不能有一本教机器学习的书对读者友好一点。首先不要假设读者擅长数学，认为读者一上来就可以看懂各种高深的数学公式，在介绍机器学习具体模型算法时要能按照从宏观到微观的顺序介绍。刚接触新的知识领域，先把模型算法的主要原理和基本结构讲清楚，让读者在脑海里勾勒出基本的轮廓，明确各种概念之间的关系，然后才深入各个细枝末节展开介绍，这样读者才不至于觉得自己一直在各种陌生的公式里转来转去，最后看得晕头转向。最后我还想再贪心一点，希望这本书的文字能够稍微有趣一点，最好能像弹幕评论那样在不经意间引人会心一笑，毕竟我是利用睡前的时间来学习机器学习，辛苦了一天，身体和精神都很疲惫，文字太生硬的话恐怕是啃不动的。

我找了很久，可惜直到最后也未能找到这样的一本书。现在，我决定自己动手来写一本。不过，这本书也并不能让你在短期内就全面掌握机器学习的各种知识。机器学习不但自成体系，自身就拥有枝繁叶茂的知识结构，而且也从多门大学科里汲取养分，又带有交叉学科的一些特点，可能将一个子问题深入研究下去就能发展成一门新学科——从神经网络发展到深度学习就是一个很好的例子。弱水三千，一本书哪怕写得再凝练透彻，也只能取一瓢饮。学习机器学习犹如建造大厦，总是需要从最基础的开始学，筑牢根基，然后一本一本地往上堆叠各有侧重的书本，才可能最终构建出完整的知识体系。

每一本书都有自己的使命。初学机器学习时，遇到的最大问题是迷茫，我深有体会。面对机器学习领域数量繁多又互有交叉的知识点，就像身处一大片繁茂的森林，没有指南设备很难不迷失方向，而大量好不容易挤出来的宝贵时间就浪费在辨别方向上了。在本书中，我负责为你踹开机器学习世界的大门，绘制出这个庞大而陌生的世界中的“山河湖海”，总体是怎样的，哪里是重点，哪里是难点，哪些点用到了哪些学科知识，点和点之间的关系又是怎样的，我都迫不及待想要一一清楚地告诉你。为了完成这个使命，我会竭尽全力，但也请原谅我无法“送佛到西”，正如前面所述，每个知识点深入下去，可能又是一片茂密的森林，机器学习涉及的知识点众多，我希望通过本书能让你清楚地看到兴趣所在，不过知识点背后仍然有很长很长的路，还请加倍努力。

最后，我想谈一谈“要不要亲手实现一遍机器学习算法”这个争议很大的问题。我推崇学以致用，用机器学习算法解决实际问题才是本书的最终目的，所以本书将会涉及

如何在实际中使用书中提及的机器学习算法的问题。对于这个问题，一般会有两种选择，一种是让读者亲手从头实现一遍算法，另一种则是直接使用现成的算法库。对于这个问题，如何选择争议很大，本书中选择的是后者。

学习机器学习的动机很多，可能是实际工作需要，可能是兴趣爱好，也可能是学业要求，从每种动机的角度看，这个问题都可能有不同的答案。我认同许多人所说的求知不能太功利这一观点，不过大家的时间和精力毕竟有限，就算不去追求投入产出比，至少也应该有一个学这门知识想要达到的目的。机器学习是更偏重于应用的学问，在当下的发展也确实使得机器学习越来越像一门技能，而不仅仅是技术。初学算法时我最想学的是里面的“最强算法”，不过在第 1 章我将介绍，机器学习算法没有最强的，只有最合适的，对于不同的问题，对应会有不同的最合适算法。所以，我们更需要关注的应该是问题，而不是算法本身。在本书中我选择介绍市面上成熟的机器学习算法包，通过现成的算法包，就能够根据实际要解决的问题直接选择所需要的机器学习算法，从而把注意力集中在对不同算法的选择上。

本书的目标读者是想要学习机器学习的学生、程序员、研究人员或者爱好者，以及想要知道机器学习是什么、为什么和怎么用的所有读者。本书第 1 章介绍机器学习总体背景，第 2 章介绍配置环境，第 3 章到第 10 章彼此独立，每一章介绍一种具体的机器学习算法，读者可以直接阅读想要了解的算法，第 11 章介绍了集成学习方法，这是一种组合机器学习算法的方法，也是当前在实际使用中常见又十分有效的提升性能的做法。

各章详细内容如下：

第 1 章首先介绍机器学习究竟是什么，特别是与“人工智能”“深度学习”这些经常在一起出现的术语究竟有什么关系，又有什么区别。本章也将对机器学习知识体系里的一些常用术语进行简要说明，如果读者此前并不了解机器学习，则可以通过本章了解相关背景知识。

第 2 章对当前机器学习算法常用的 Python 编程语言以及相关的 Python 库进行介绍，同时列举一些常用的功能。

第 3 章开始正式介绍机器学习算法，要介绍的第一款机器学习算法是线性回归，本章将对回归问题、线性模型和如何用线性模型解决回归问题，以及对机器学习解决问题的主要模式进行介绍。

从第 4 章开始，介绍当下机器学习应用最广的分类问题，第一款解决分类问题的算法是 Logistic 回归分类算法，即用线性模型结合 Logistic 函数解决分类问题。

第 5 章介绍 KNN 分类算法，这款算法不依赖太复杂的数学原理，因此一般被认为是最直观好懂的分类算法之一。

第 6 章介绍朴素贝叶斯分类算法，它基于贝叶斯公式设计，理论清晰、逻辑易懂，是一款典型的基于概率统计理论解决分类问题的机器学习算法。

第 7 章介绍决策树分类算法，这是一款很重要的算法，从思想到结构都对程序员非常友好，当前 XGBoost 等主流机器学习算法就是在决策树算法的基础上，结合集成学习方法设计而成的。

第 8 章介绍支持向量机分类算法，这是一款在学术界和工业界都有口皆碑的机器学习模型。在深度学习出现之前，支持向量机被视作最被看好的机器学习算法，能力强、理论美，也是本书中最为复杂的机器模型。

第 9 章介绍无监督学习的聚类问题，以及简单好懂的聚类算法——K-means 聚类算法。

第 10 章介绍神经网络分类算法，当前大热的深度学习就是从神经网络算法这一支发展而来的，而且大量继承了神经网络的思想和结构，可以作为了解深度学习的预备。

第 11 章介绍集成学习方法，以及如何通过组合两个以上的机器学习模型来提升预测效果。

我自己也经常阅读各类书籍，常常看到不少作者提到写书不易，待自己写作了一本书之后，才真正体会到写书真是一段漫长的“马拉松”，只有真正经历了才能明白其中所需要的决心和毅力。本书能顺利写作完成，首先要感谢我的妻子，她的一句“真想看你写完的这本书”是我克服白天工作的疲惫，坚持写下来的最大动力；我还要感谢我的父母，他们培养了我学习新知识的兴趣，更让我懂得了学习新知识的最大乐趣在于分享，继而深深地埋下了写作本书的梦想种子；最后我需要特别正式地感谢本书的策划编辑吴怡女士，这个世界上大大小小的进程都需要一个第一推动力——吴怡女士促使了我写作本书的梦想变成现实。

CONTENTS

目　录

CHAPTER 1

第1章 机器学习概述

你身处的世界，是一个已经被机器学习包围了的世界。大到自动驾驶，小到邮件过滤，就连你不久前浏览过的APP，只要里面带有“推荐”栏目，背后依靠的都是机器学习。作为一门科学，机器学习已经在人们的生活中大展拳脚，而前面仍然有着广阔的发展空间。学习机器学习，现在正是时候。本章将介绍机器学习的背景知识，包括基本概念和原理、常用的机器学习算法、性能衡量指标、数据对算法结果的影响等，为理解后续章节打下基础。如果你此前只听过“机器学习”这四个字，但还不太了解具体所指和涉及什么内容，那本章正是为你私人订制。

1.1 什么是机器学习

在开始我们的机器学习大冒险之前，应该先明白什么是机器学习。即使两年前，很多人听到“机器学习”这个词的第一反应恐怕也是一脸错愕，可能脑海中浮现的是一张泛着冷光的钢铁人脸，或者科幻电影里的某一个场景。但现在完全不同了，网络上铺天盖地地在谈机器学习，货架上摆满了相关的书籍，学者、名人轮番发表对相关话题的看法，连之前一些其他领域的公众号都开始提醒你“小心工作不要被机器抢了去”。只是大家都没有直面那个问题：什么是机器学习？

想回答这个问题，要从机器学习、人工智能和深度学习三者的关系说起。机器学习、

人工智能和深度学习都是最近很火的词，有的人用截然不同的态度评价它们，好像三者并无联系，有的人却认为它们不过是新瓶装旧酒，都是商家宣传推广的噱头。

这些看法未免有些片面，机器学习、人工智能和深度学习的目标都是让算法模拟“智能”，但层次范围不同。用北京市的环线来形容三者的关系实在最形象不过了（见图 1-1）：人工智能（Artificial Intelligence）涵盖范围最广，三环以内都可以叫人工智能，它关注的问题和方法也最杂，包括知识推理、逻辑规划以及机器人等方面。机器学习（Machine Learning）住在二环，是人工智能的核心区域，也是当前发展最迅猛的一部分，子算法流派枝繁叶茂，但思想比较统一，本书就是对机器学习“查户口”。至于当下“网红”——深度学习（Deep Learning），其实原本是从机器学习的神经网络子算法分支发展出来的一系列成果，知识体系一脉相承，只不过近年大出风头，干脆重新起了个名字“单飞”了，以后有机会再详细介绍。

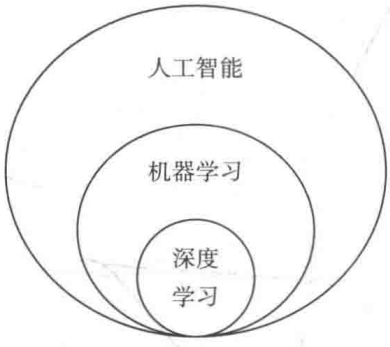


图 1-1　人工智能、机器学习和深度学习三者是包含关系

既然同宗同源，基本原理自然也是大同小异。不过机器学习这几年实在太热，人们担心出现机器代替人类而带来失业风险、机器过于智能以致人类无法控制等问题，这在客观上确实给人们理解机器学习带来一定困难——我们面对的到底是一门新的技术，还是一类新的物种？

其实，虽然机器学习是在这几年才成为热议话题，但其背后的知识和技术体系绝不是这几年才横空出世。今天我们所使用的机器学习知识，都是经年累月的研究成果。很多在今天大放光彩的机器学习算法，甚至在 20 世纪 50 年代就已经被提出了，只不过近

几年机器性能突飞猛进，这些算法有了强力硬件的支持，才得以像猛虎出柙一般取得瞩目的成绩。至于作为这些算法背后支撑的数学知识体系，那就更久远了。

那么机器学习到底是什么呢？没有什么特别的，形象地说，就是一个“苦命”的外包程序员。以前我们写程序时，先要知道输入什么数据、输出什么数据，然后设计功能函数，最后一行一行敲代码来实现。现在有了机器学习，我们写程序就简便多了，即把输入输出数据一股脑儿全扔过去，这位外包程序员二话不说就吭哧吭哧把后面的活全干了。我们给这个过程起了一个好听的名字——“机器学习”。

1.2 机器学习的几个需求层次

“机器学习可难懂了，首先数学知识是必需的，线性代数、概率论、微积分等一门都不能落下，还要掌握编程技术，最好是先用 Python 做一个项目……”

也许每一位打算深入了解机器学习的爱好者或多或少都听过类似的“忠告”。机器学习确实是一门算法科学，数学也确实是它背后的源泉和依靠。不过我们学知识，不是因为要学而学，而是因为有用才学的。机器学习不是“屠龙之技”，它从诞生开始就立足于解决实际问题。你要解决什么样的问题，才决定你需要学习什么样的知识，以及学到什么程度。

目的明确并不等于功利。曾经听过一个寓言：一位农夫发现墙上的钉子快要脱落了，想用锤子敲打一下，可这时他才发现家里的锤子坏了，他又去修锤子，结果修锤子的工具也坏了……这位倒霉的农夫不断陷入这样的循环，于是到了最后，他已经完全忘记自己最初的目标，因为大家发现他正在森林里跑来跑去，想要锯倒某棵大树。开始学习机器学习时，说要“向死而生”好像过于严重，但知道自己需要什么，带着目的去学习确实才是最有效率的。

也许你还需要时间思考自己到底要什么，也许你只是走过路过，顺道过来看看。那么在这里，我将根据自己遇到过的实际问题，分享我对机器学习知识的三个需求层次。

设计需求层次：这个层次有一个更接地气的名字——包工头层次。前面我们说过，机器学习就像外包程序员，我们作为包工头，只需要考虑两件事，一件事是我们要做什么，另一件事是程序能做什么。至于程序怎么做，那是后面的事。

譬如你现在要设计一个购物网站，需求很简单，也很直接，就是希望看到用户踊跃掏腰包。你问机器学习，咱能让用户多掏腰包吗？机器学习说："没问题，我能实现一个商品推荐功能，朴素贝叶斯分类算法通过挖掘共现频率……"你一挥手赶紧打断它："我不需要了解这些过程，是骡子是马拉出来遛遛就知道了。"这就是设计需求层次。

调用需求层次：上面的例子很理想，当然以机器学习现有的技术，还远不能像演二人转一样一唱一和这么智能。不过在实际工作中，又确实需要这样一个拟人的角色，这就是调用者。现在机器学习算法和算法库是现成的，就好比已经摆好了琳琅满目的上等食材，不过还需要有一位大厨把它们做成一桌好菜，才能最后端上桌。

数学需求层次：有人笑谈，在很多人心目中，数学都是"猛于虎"的存在，甚至有人说书中多一条数学公式，就会少一半的读者。虽然对于一部分人来说，数学确实让之生畏，但不要忘记数学还是一门语言，是一门由符号组成的"数学语"。与口头语言不同的是，"数学语"准确严谨，特别是需要描述一些抽象概念时，数学的优势更加明显。

对于机器学习算法中涉及的一些不太直观的理念，我们当然可以进行口语化的描述，但口语自身也存在很多局限，用来描述抽象概念反而可能让人感觉"隔靴搔痒"，更重要的是口语存在很多歧义，你理解的意思未必是我想表达的意思，特别是对于一些比较复杂的概念，可能失之毫厘而谬以千里。本书虽然力争有趣，但还是努力想成为一本严谨的读物，而不只是科普性质的"简史"。所以，在介绍新算法时会尽力用直白、形象的语言建立算法的总体图景，然后再使用"数学语"精工细作，确保每一个微小的细节都能严丝合缝。

这里首先需要说明一点，机器学习算法会涉及各种各样的数学表达式，但也许与你想的不太一样，就算同一个含义的数学式子也可能有不同的表达形式，就像同一事件可以有不同的描述方法。为了方便应用，在有多种方式可选的情况下，本书将优先采用

Scikit-Learn 说明文档中的数学表达式。

不难想象，这三种需求是从机器学习算法的思想原理、运行流程到数学解析的层层递进和不断深入，就像一只在高空翱翔的雄鹰向目标俯冲时所看到的画面，这也是本书介绍每一种机器学习算法所遵循的主要路径。不过，是不是必须先读完像《线性代数》这些数学教程才能看懂数学解析呢？能不能既了解其中原理，又不必先啃数学书呢？这个要求好像很贪心，也确实是一个挑战。本书就来挑战一下，遵循“现学现用”原则，对于马上需要用到的数学知识，我们就地现学。幸运的是，刨除细枝末节，机器学习的主要原理所用到的数学知识反而较为集中，确实存在在尽可能少地介绍数学背景的同时，尽可能多地介绍机器学习算法的数学原理的可能。

1.3 机器学习的基本原理

从本节开始，我们正式踏入机器学习的领域。不妨对照图 1-1 的关系图，把机器学习领域想象成一个圈，圈外是人工智能，用各种千奇百怪的方法模拟“智能”，而圈内则简单得多，只有一种方法，就是“学习”。

“这（机器学习）是一个妨碍你理解的名字。”类似说法在本书中还将一再重复。

那么叫什么更容易理解呢？改叫“统计模型训练”就好了很多。因为机器学习的过程不是让机器蹲在小板凳上读书识字，而是更接近于马戏团里的动物训练。

我们都看过马戏表演，各种动物根据训练员的指示，或者踢球或者跳舞，十分有趣。但不妨深入考虑一下：毕竟“爬说语”只有哈利·波特这样的巫师才能掌握，我们一般无法与动物直接沟通，那么训练员是怎样做到与动物配合无间的呢？具体过程很复杂，但说起来很简单，就是利用反馈激励机制。譬如训练海豹，训练员给海豹一个信号要它拍手，最开始海豹当然不知道要做什么，它可能做出各种动作，如点头、扭动身体，但只要它无意中做出了拍手的动作，训练员就会奖励它一条小鱼。海豹希望吃到小鱼，但它没有那么聪明，无法立即明白听到信号只要拍手就能吃小鱼，需要训练员花费大量

的时间，不断给它反馈。久而久之，海豹形成了条件反射，听到信号就拍手，训练就成功了。

机器学习的过程与此类似，所以机器学习的一项主要工作就叫作“训练模型”。可是仔细一想也许会发现，“训练”这个词在机器学习中虽然常见，但感觉还是无法与机器很好地契合。那么干脆再换个词——“拟合”。

拟合是机器学习的主要工作。在学习机器学习的知识之前，你也许已经接触过很多算法，譬如冒泡排序算法或MD5消息摘要算法，它们和机器学习算法有一个很大的区别，即这些算法的结果值是“算”出来的，只要确定了输入，输出就是一个定数。而机器学习算法的结果是“猜”出来的，猜的结果受很多因素影响，具体后面再讲。如果给机器学习算法总结一个本质，我认为最符合的就是一个“猜”字。

机器学习的过程就是不断回答两个问题：“我猜是什么”和“我猜中没有”。这两个问题推动着学习过程不断进行，根据“我猜是什么”的结果回答“我猜中没有”，再根据“我猜中没有”的结果回答“我猜是什么”。

不知道大家是否玩过“猜数字”游戏。游戏规则很简单，首先裁判选定一个数字，接着参赛选手也报一个数字，裁判回答他猜大了或猜小了，不断重复这个过程，直到最后猜中。

接下来，游戏内容不变，我们引入两个机器学习的术语——“算法模型”和“损失函数”，这两个术语的具体含义马上就会解释，现在只要简单地把这两个术语当作两个名字。把参赛选手替换成“算法模型”，把裁判的回答替换成“损失函数”，那么猜数字的过程就是一个完整的机器学习过程：算法模型输出一个数值，损失函数经过计算，回馈一个偏差结果，算法模型根据这个偏差结果进行调整，再输出一个数值，周而复始，直到正确为止。这就是机器学习的学习过程，这个过程在机器学习里称作“拟合”。

拟合可以说是机器学习中最重要的概念之一，甚至有人认为机器学习算法中所谓的“学习”，本质就是拟合数据。在机器学习中，除了拟合外还有两个很重要的概念，分别