zure ML技术培训的御用指导书

从入门到实战训练，微软Azure解决方案架构师、日本微软公司数据科学家手把手教你轻松掌握Microsoft Azure机器学习技能，高效快捷创建机器学习工作区，易如反掌地进行开发

さわってわかる機械学習

Azure Machine Learning实践ガイド

微软Azure机器学习实战手册

[日] 千贺 大司（Hiroshi Senga）
山本 和贵（Kazuki Yamamoto）◎著
大泽 文孝（Fumitaka Oosawa）

[日] 笹木 幸一郎（Koichiro Sasaki）
佐藤 直生（Naoki Sato）◎编审

贾硕 魏宁◎译

中国人民大学出版社

さわってわかる機械学習
Azure Machine Learning実践ガイド

微软Azure机器学习实战手册

千贺 大司（Hiroshi Senga）
[日] 山本 和贵（Kazuki Yamamoto）◎ 著
大泽 文孝（Fumitaka Oosawa）

[日] 笹木 幸一郎（Koichiro Sasaki）
佐藤 直生（Naoki Sato）◎ 编审

贾硕　魏宁 ◎ 译

中国人民大学出版社
• 北京 •

图书在版编目（CIP）数据

微软 Azure 机器学习实战手册 /（日）千贺大司，（日）山本和贵，（日）大泽文孝著；贾硕，魏宁译 . -- 北京 : 中国人民大学出版社，2017.11

ISBN 978-7-300-25095-3

Ⅰ . ①微… Ⅱ . ①千… ②山… ③大… ④贾… ⑤魏… Ⅲ . ①机器学习—手册 Ⅳ . ① TP181-62

中国版本图书馆 CIP 数据核字（2017）第 257498 号

微软 Azure 机器学习实战手册

千贺大司

[日] 山本和贵 著

大泽文孝

贾 硕 魏 宁 译

Weiruan Azure Jiqi Xuexi Shizhan Shouce

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511770（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http: //www.crup.com.cn		
	http: //www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京宏伟双华印刷有限公司		
规　　格	170mm×230mm　16 开本	**版　　次**	2017 年 11 月第 1 版
印　　张	14.75　插页 1	**印　　次**	2017 年 11 月第 1 次印刷
字　　数	145 000	**定　　价**	65.00 元

推荐序 1

入门捷径，升级指南

AI 人才是分层的。跟很多其他领域的开发技术不一样，目前无论在中国还是在国际上，学校教育在人工智能方面被证明是比较成功的。因为学校教育相对比较严谨，因此对于变化特别快的领域，学校教育有先天的劣势。比如 Web 前端的开发，几乎每半年就有一个新的风向，让学校教育去跟着这个变化跑，显然是不合适的。但是人工智能、机器学习领域不太一样，它建构在坚实的数学和统计学基础之上，虽然创新层出不穷，但是基础相对是稳固的、成体系的，特别适合高等学校教学模式。

问题在于，学校教育的供给量严重不足，导致市场上人工智能人才严重匮乏。供需的缺口抬升了人工智能技术人才的薪资水平，也促使越来越多的人想学习人工智能。对于半路出家的技术人员来说，未必要像在学校里那样一板一眼地往前走，而是要尽可能发挥自己有经验的长处，通过不断的实践来学习。

一般来说，一个人要入门，首先要学会一点 Python 语言，刷一刷数学基础，然后从最简单的线性回归算法开始，一边学理论一边做例子，一步步掌握更复杂的算法，直至攻克深度学习。目前，很多面向开发者的机器学习课程都是这样一个模式。这个模式怎么样？事实证明是可行的，但是并不是没有问题。我就听到一些学习者跟我反映，说做案例和作业的时候经常感到很困难，因为太多的难点纠缠在一起了。比如也许你对算法本身理解了，但是卡在 Python 上，或者卡在数据处理上。另一方面，局部问题容易攻破，整体视图难以形成。

我觉得这些还是工具问题，是工具可以解决、应该解决的问题。

Azure Machine Learning Studio 就是这样一个工具。微软在可视化开发工具方面一直走在最前面。面对机器学习的大潮，微软推出了 Azure Machine Learning Studio。我第一次看到有人在我面前展示它的用法，一方面觉得很惊讶，能够把那么复杂的机器学习流程变得如此轻松易用；另一方面也觉得有些疑惑，如果大家都拖拖拽拽就做完了，那谁还能学到真正的底层算法呢？这样的“人才”能放心用吗？

看了这本书的大致内容，我的疑问得到了解答。Azure Machine Learning 作为当前可视化操作程度最好的工具，实际上并不是代替用户思考，也不会损害用户的控制力，而是帮助你尽快形成一个整体解决方案。从这本书来看，这个工具相当强的功能在于数据的探索性分析（EDA）和整理（wrangling），这无疑对任何机器学习开发者来说都是一个巨大的福利。而通过可视化的工具，用户可以在 Machine Learning Studio 中很快建立一个整体的架构，获得全局视图。这些工作的重要性，怎么强调都不过分。在一个机器学习系统中，特别是大数据环境下，良好的数据质量，合理的整体架构，往往比算法重要得多。当然，在这个基础之上，用户仍然可以进行细致的调参，甚至进入到代码层面进行精细的调整。

因此，在我看来，无论是初学者还是有一定经验、正在爬坡的人工智能学习者，可以尽早接触 Azure Machine Learning 工具，它能帮助你提高效率，合理分配精力，尽快形成大局观。这对于初学者来说，是一条入门捷径，而对于稍有经验者来说，也是升级的指南。

蒋涛

CSDN 创始人、极客邦创投合伙人

推荐序 2

程序员的未来之路

5 年前，当我们还在微软工作的时候，在一个讨论会上曾经有人提出这样一个观点：“真正程序员的门槛在提高，而不是降低。未来，只有那些开发工具系统、开发平台系统，或者使用机器学习方法解决实际问题的角色才能被称为程序员。”在当时，99% 的程序员用 Visual Studio，或者用 Emacs；用 Asp.Net、PHP，或者用 MFC，日复一日写着平淡无奇的业务逻辑代码。那时那刻，这个看似“极端”的观点，对于大部分程序员来说就像晴天霹雳，宣告了 99% 传统程序员黯淡的职业前景。

此时此刻，回首过去这 5 年，那个“极端”的观点还并未成真。然而，它的确精准地预测了一个正在加速发生的大趋势：机器学习新方法日新月异，机器学习工具层出不穷。与之相应的，我们看待传统问题的角度也发生天翻地覆的变化，越来越多的实际问题可以转化为机器学习问题来解决。特别是，在最近一波深度学习掀起的 AI 热潮中，底层工具框架如 TensorFlow、MXNet、Caffe2 等将机器学习系统的构建过程完全 API 化，从而进一步将机器学习，特别是深度学习提升到编程范式的层次。构建网络结构即为编程，编程即为网络结构构建。

在这样的大趋势下，传统程序员们该如何“升级”自我，迎接下一个 5 年挑战？ 5 年前，这的确是一个问题。如果一个传统程序员想投身机器学习的开发，他首先需要采购一台或多台足够强大的服务器，来提升计算效率，缩短开发周期。除此之外，对于大部分实际问题，十有八九他要重新造一些轮子，或者在一些还远未成熟到可以上天入地的开源工具上缝缝补补，才能完成一些基础工作。更糟

糕的是，数据在云端（公有云或者私有云），把机器学习框架和大数据相连接可并不总是一件“小”事情。

值得庆幸的是，伴随着公有云平台 5 年来突飞猛进的发展。机器学习工具也如其他基础设施一样成为公用云不可分割的一部分，比如本书所介绍的 Azure Machine Learning。在 Azure 上，Azure Blob、Azure SQL 和 Azure ML 无缝集成。我们无需再费尽心思地思考如何连接数据和工具，因为他们无时无刻不相连。同时 Azure ML 以一个 PaaS 的姿态展现在我们面前，需要多少计算力支持模型训练对于我们是完全透明的，也无需操心。最后，最重要的是各种强大且成熟的轮子已经全面就位，从数据清洗、特征工程、模型选择到模型部署，甚至连结果展示（Data visualization on PowerBI）、推荐解决方案（Recommendation engines for end to end scenario），以及最终盈利模式（Selling model on Azure Marketplace）都已经融入到 Azure ML 上下游中。

在此向每一位立志在未来成为一名“真正程序员”的小伙伴们推荐《微软 Azure 机器学习实战手册》一书．希望大家能借用 Azure ML 这杆利器，完成自我升级。

魏颢

码隆科技 研发副总裁

编审序

在这一年的时间，很多顾客向我们表达了想在微软 Azure 平台开展机器学习（Machine Learning）的计划，或想通过微软 Azure Machine Learning（Azure ML）的技术进行机器学习验证的意愿。网络新闻以及 IT 杂志多次介绍到“机器学习”一词，现在“机器学习”已经成为在社交媒体以及书签服务当中备受关注的流行语。

然而，实际上，真正开发过机器学习服务系统平台或业务 App 的人并不多。与关注机器学习的人数相比，从事研发的人数少得甚至可以忽略不计。很多人都在考虑，如果有机会，希望可以尝试一下创建机器学习模型，但是这些人要么不知道从何处入手，要么不清楚机器学习建模的效果怎样。我想，捧着这本书看到这里的读者是不是也都面临着同样的问题呢。

微软的 Azure ML 是可以快速创建机器学习 App 的云端服务平台，其中还包含可以使用该服务基本功能的免费套餐。Azure ML 可以以图表的形式掌握“现在在做什么”“得到了什么样的结果”等信息；并且作为标准功能，Azure ML 具备各种数据统计方式及多种机器学习算法。除此之外，Azure ML 还能够以 REST（Representational State Transfer）应用程序编程接口（Application Programming Interface，API）的形式公开已完成的机器学习处理方法，并拥有使用浏览器或者 Excel 的任意数据对 REST API 进行检测的辅助功能。换言之，有一个浏览器就可以通过四则运算完成神经元网络等各种各样的处理，也可以从 App 上进行实际操作检验已完成的处理。这些特色可以大大提升初学者的学习曲线。以前，别说将机器学习编入到实际的 App 程序当中，就连实现一边运行一边学习这一目标都很

难。但是现在，通过使用Azure ML进行实际操作，就可以轻松踏入机器学习的大门。

本书通过实际接触机器学习服务来加以理解，我觉得这是了解机器学习服务的第一步。实际上，本书作者大泽先生在书中加入了他本人接触 Azure ML 时不理解的以及难理解的内容，所以我觉得这会有助于其他学习者的理解。另外，FIXER 公司曾多次举办过 Azure ML 的实操研讨会，虽然是收费活动，但是每次都座无虚席。本书当中也加入了在收费学习会上说明的操作要素，因此书中内容具有很强的实操性。所以希望大家能够打开浏览器，一边翻阅本书，一边对照 Azure ML 的操作流程来学习。

“机器学习”以及包含机器学习的“人工智能”（Artificial Intelligence，AI）成为了流行语，还有很多人觉得“只要输入数据就会得到最合适的答案”，这是目前的现状。但是，今后随着越来越多的人对如何使用机器学习服务系统有了更深入的了解，并且随着具有机器学习服务相关知识和经验的技术人员不断增多，预计机器学习的实际应用会变得更加普及。其中，适合使用机器学习服务的对象和难以使用机器学习的对象分类会按照具体实例变得更加详细。现在也开始出现了将机器学习纳入到系统的“需求建议书”中的方案，相信跨越鸿沟的那一天就要到来了。不管在系统当中使不使用机器学习，只要进行实践就可以用自己的语言来评价好坏，并且可以判断为了进行更深入的研究还需哪些技能 [实际上本书中的实践并不是以深度学习（deep learning）、学习模式的改善、各项业务域名的对象为前提的，要想达到熟练水平还需要其他途径]。

我相信，本书可以帮助大家更好地理解机器学习进而推动实际应用。那么接下来，我们就进入机器学习的世界开始翱翔吧！

笹木幸一郎 佐藤直生
微软日本股份有限公司

前言

大概从2014年开始，在我们周围越来越多地听到和看到“机器学习”这个词。微软公司推出的通过图形用户界面（Graphical User Interface，GUI）工具就可以轻松实现机器学习的Azure ML于2014年6月首次对外发布，并于2015年2月开始提供通用版本（General Availability，GA），之后我感到“机器学习”这一概念快速传播开来。

2015年5月，在微软日本股份有限公司举办的面向日本国内技术人员的最大盛会“de：code2015”上，我们几位介绍了Azure ML成功预测出超过100万用户脱离智能手机游戏（退会）这一案例。并且于同年10月，我们在日经BP社主办的学习交流会“从零开始了解‘机器学习’实践讲座”中担任了讲师，就Azure ML如何实操进行了现场解说。通过这些活动，一方面大众对我们FIXER公司有了更多的了解，另一方面FIXER公司也获得了来自日本知名企业的诸如“希望使用机器学习预测器械、机器故障并进行预防”“希望使用机器学习创造机器人人工智能”等委托项目。

本书旨在将机器学习应用到现实的商业当中，并将其转变为商品或服务，而不是单纯地将机器学习捧为流行语。换言之，我们出版本书的目的并不是追求学术价值，而是为了让大家能够使用、活用机器学习，不落后于时代变革的潮流，甚至能够引领时代潮流。希望通过本书，工程师以及商业人士能够发明出使用机器学习的新型服务，或者从数据中发现以前被忽略的新视角。

以前，一提到机器学习，就会想到是那些被称为“数据科学家”的专业人士

使用的专业工具，但是如今情况会有所不同。奋战在商界的企业家们可以对数据进行直接分析，让使用数据的服务以及搭载人工智能的服务开始成为可能。可以说，企业家和数据科学家之间在认知以及理解上的障碍已经消除。初级的系统工程师和开发商很难涉足的数据分析、推荐引擎以及人工智能的开发和使用难度也会大幅下降。

“统计”一词自公元前诞生于埃及以来已经发展了 3000 多年，机器学习的理论基础自出现至今已经过了 40 多年，但在商业中的实际应用可以说依然非常受限。我们几位常年从事股票数据的分析，通过各种方式对市场动向及个别股票产品进行预测，但是仅仅依据从金融工程学以及统计学中导出的现有理论，很难获得高水平成果。

简单一提的是，过去在未来市场预测方面能够取得较高水平成果的方式，是把几十台服务器联接起来，使用计算机进行大量的运算，分析离散数据而不是分析函数数据。而现在，随着摩尔定律的不断发展，计算机的处理性能以及计算资源也在不断扩大。自从进入了云端时代，即使是个人也可以在短时间内以较低成本同时使用几十台甚至几百台服务器。

与此同时，现在可以以较低的成本储存大量数据。比如，当今世界很多人都使用智能手机，谷歌、苹果公司的以及手机 App 开发人员每时每刻都能收到来自世界各地的几亿部智能手机中的大量数据。除此之外，每隔几分钟或者几小时，就能收到来自几百万辆、几千万辆汽车以及家电产品的注册信息。如果是在 10 年之前，收集、存储如此巨大的数据是不可能的。10 年前，1TB 容量的企业版高速存储器价格超过 1 亿日元，但是现在，不到 1 万日元的硬盘（Hard Disk Drive，HDD）的容量就已经超过了 1TB。2016 年 4 月，Azure 的存储服务价格标准为：使用 99.9% 的服务级别协议（SLA）用三块硬盘备份的设备，1GB 平均每月 228 日元。

当今时代，机器学习已经能够使用存有大量计算资源以及大量数据的系统，像以“分析并活用大量数据”为标题的新闻急剧扩散开来，预示着近几年的发展动向。另外，机器学习自诞生以来经过了几十年发展，相关的观点及方法论再次受到世人瞩目。以前，受计算能力的影响，只能做一些很简单的事情，并未取得很大成果。但是现在可以使用大量数据让机器在短时间内学习，可以说现在和以往已经截然不同，能够让计算机对未来进行预测和判断。

另外，进行机器学习的环境以及工具以前是专业人员所擅长的，但是微软公司通过 GUI 基础让其得到飞跃性简化，现在很多人都可以使用机器学习工具。1975 年，比尔·盖茨和保罗·艾伦创建了微软公司，并凭借 BASIC 代码名声大噪。之后的几年又向 Visual Basic 发展，推动社会发展成为谁都能轻松使用 Windows 和 GUI 软件的世界。另外，在 Windows 95 中安装配备 TCP/IP 和网站浏览器，为网络的普及作出了巨大的贡献。现在，通过云端服务的微软 Azure 以及在此基础上运作的机器学习“Azure ML”，使得世界逐渐向计算机为具有知性的人类提供援助的方向发展。

新的历史一页才刚刚翻开，让我们一同朝着机器学习创造的新世界迈进吧！

千贺大司

目 录

第 1 章　什么是机器学习 /1

明晰机器学习 /2

机器学习概述 /2

机器学习流行的“原因” /4

将机器学习用于商业的方法 /5

消除对机器学习的误解 /9

机器学习通过数据进行判断 /9

机器学习是“系统” /11

机器自己会变聪明吗 /12

必须决定“特征向量” /12

开启机器学习之旅 /14

机器学习专用工具 /14

无须编程就可以使用的 Azure ML /15

即使如此，依然想编程 /17

通过判断目标来选择分类器 /17

第 2 章　收集数据 /19

使用公司内部数据 /20

日志文件等历史数据 /20

非时间类型数据 /22

使用公开数据 /22

DATA.GO.JP/22
DATA.GOV/23
Twitter/23
GitHub/35

第 3 章　通过 Azure ML 创建机器学习模型 /39

Azure ML 的基本操作 /40
注册 Azure ML Studio/40
在工作区进行操作 /41
机器学习的方法 /43
在 Azure ML 中进行机器学习的流程 /43
创建机器学习模型时 Experiment 的编辑界面 /45
机器学习模型的构成和种类 /47
学习逻辑 /47
计算逻辑 /48
学习组件的种类 /48

第 4 章　使用回归分析预测数据 /53

什么是回归分析 /54
本模拟所实现目标 /54
本模拟所建模型 /55
上传用于分析的数据集 /57
下载 CSV 文件样本 /57
将 CSV 文件作为数据集进行上传保存 /59
新建 Experiment/62
添加和调整所要分析的数据集对象 /64
添加数据集 /65
将范围缩小至使用列 /70

修复受损数据 /75
分离学习用数据和评价用数据 /80
构建学习逻辑 /83
构成回归分析的组件 /83
使用已训练模型预测评价用数据 /87
使用评分模型进行数据预测 /88
确认预测值 /91

第 5 章　尝试使用已建回归分析模型 /95

使用已训练模型进行计算 /96
上传用于计算的数据集对象 /96
在评分模型右上方输入数据即可得出结果 /97
保存已训练模型，使其在其他 Experiment 中也可以使用 /99
保存已训练模型 /100
使用已训练模型进行预测 /102
新建用于预测的 Experiment /102
创建可进行数据预测的机器学习模型 /103
观察运行结果 /105
以 CSV 形式输出 /106
数据转换组件 /107

第 6 章　提高预测精度 /111

提高预测精度的方法 /112
确认目前的预测精度 /113
使用评估模型对分析结果进行评价 /113
确认评价结果 /115
更改参数提高精确度 /117
更改 Linear Regression 的参数 /117

优化学习组件 /119

可用于回归分析的学习组件种类/119

更改为贝叶斯线性回归 /120

使用有限的学习数据进行检验 /123

使用 “Cross Validate Model” 组件 /125

确认 “Cross Validate Model” 的评价结果 /126

第 7 章　通过统计分类进行判断 /129

什么是统计分类 /130

本模拟所实现目标 /130

本模拟所建模型 /131

用统计分类创建分类机器学习模型 /132

新建数据集 /132

新建 Experiment/134

创建数据集 /134

构建学习逻辑 /137

预测和评价 /139

确认和反思学习结果 /141

确认使用评价用数据得出的结果/141

评价统计分类的学习结果 /142

使用其他统计分类学习组件 /146

第 8 章　用聚类方法判别相似数据 /151

什么是聚类 /152

本模拟所实现目标 /152

本模拟所建模型 /154

创建可通过聚类分析分组的机器学习模型 /156

新建数据集 /156
新建 Experiment/157
添加数据集 /158
构建学习逻辑 /161
确认分组结果 /164
将用于评价的数据加入到已训练的学习模型中 /167

第 9 章　活用实验结果 /173

Web API 化 /174
数据可视化/178

第 10 章　让机器越来越聪明 /179

进行模型的二次学习 /180
用 Web API 更新公开的分类器（模型更新）/187

附录　使用 Azure ML 的方法 /201

创建环境 /202
创建 Microsoft 账户 /202
激活订阅 /203
登录 Azure/208
云优化您的业务 /208
创建工作区 /210
访问 Azure ML Studio/211
关于收费 /213
免费使用 /214