

微软Azure机器学习专业指南

Azure Machine Learning

Azure

机器学习

[加] 杰夫·巴恩斯 (Jeff Barnes) 著

高雪松 胡伟凤 马琳涛 译

Microsoft

Pearson

Azure Machine Learning

Azure

机器学习

[加] 杰夫·巴恩斯 (Jeff Barnes) 著

高雪松 胡伟凤 马琳涛 译

人民邮电出版社
北京

图书在版编目 (CIP) 数据

Azure 机器学习 / (加) 杰夫·巴恩斯
(Jeff Barnes) 著 ; 高雪松, 胡伟凤, 马琳涛译. — 北
京 : 人民邮电出版社, 2018. 11
ISBN 978-7-115-48869-5

I. ①A… II. ①杰… ②高… ③胡… ④马… III. ①
机器学习—指南 IV. ①TP181-62

中国版本图书馆CIP数据核字(2018)第157746号

版权声明

Authorized Translation from the English language edition, entitled AZURE MACHINE LEARNING, 1st Edition
by JEFF BARNES, published by Pearson Education, Inc, publishing as Microsoft Press, Copyright © 2015.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means,
electronic or mechanical, including photocopying, recording or by any information storage retrieval system,
without permission from Pearson Education, Inc.

CHINESE SIMPLIFIED language edition published by POSTS AND TELECOMMUNICATIONS PRESS,
Copyright ©2018.

本书中文简体字版由美国 Pearson Education 授权人民邮电出版社出版。未经出版者书面许可, 不得以任
何方式复制或发行本书任何部分。

版权所有, 侵权必究。

-
- ◆ 著 [加] 杰夫·巴恩斯 (Jeff Barnes)
 - 译 高雪松 胡伟凤 马琳涛
 - 责任编辑 吴晋瑜
 - 责任印制 焦志炜
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
 - 邮编 100164 电子邮件 315@ptpress.com.cn
 - 网址 <http://www.ptpress.com.cn>
 - 固安县铭成印刷有限公司印刷
 - ◆ 开本: 800×1000 1/16
 - 印张: 11.75
 - 字数: 222 千字 2018 年 11 月第 1 版
 - 印数: 1—2 400 册 2018 年 11 月河北第 1 次印刷
 - 著作权合同登记号 图字: 01-2016-6011 号
-

定价: 55.00 元

读者服务热线: (010)81055410 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京东工商广登字 20170147 号

内容提要

本书基于 Azure Machine Learning Studio 探讨了现代数据科学算法的背景、理论和实际应用。

全书共 8 章。第 1 章描述了在数据科学领域，Azure 机器学习如何通过实现完全托管的数据科学云服务迈出预测分析解决方案的关键一步；第 2 章讲解预测分析科学和方法论的基本概念；第 3 章探讨 Azure ML Studio 的基本原理；第 4 章基于一个实用的 Azure 机器学习预测模型，探讨可以用来调用 Azure 机器学习 Web 服务的客户端和服务端应用程序的类型；第 5 章深入探讨 Azure ML Studio 提供的一些复杂的机器学习算法；第 6 章探讨数据分析的挖掘方案，包括自主数据分析、确定数据的相关性、推断逻辑分组以及被广泛研究的用“从树木到森林”的算法处理混乱数据的方案；第 7 章介绍当今互联网中最强大和被广泛使用的预测分析的实现方法；第 8 章探索如何将“持续学习”纳入到预测模型工作流的实现机制上。

本书适合从事云计算、机器学习和数据科学相关行业的开发人员、工程师阅读，也适合作为大专院校相关专业和培训机构的教学用书。

读者对象

本书的重点是提供数据科学的基本信息和利用 Azure Machine Learning Studio 构建数据科学解决方案的技术。本书重点面向数据科学家和工程师，以及利用云计算和机器学习的开发人员和专业人士。Azure 机器学习可以帮助读者了解机器学习技术的数据科学背景并快速构建与测试机器学习解决方案。

本书通过具体的例子和详细的步骤展示如何构建并部署当今广泛使用的每一个关键的预测分析算法，以及它们在 Azure ML Studio 中相应的实现方法。这些材料不仅对没有 Azure 机器学习经验的开发人员极具意义，对于在数据科学领域有丰富经验的专家也非常有

前言

微软 Azure 机器学习是为开发者提供的一种云计算服务，开发者可以使用它建立预测分析模型（支持各种数据源的训练数据），并能轻松地将模型部署为云 Web 服务供“消费者”使用。Azure ML Studio 提供了丰富的功能来为许多端到端的工作流场景建立预测模型，包括公共数据源的易用访问，丰富的数据探索和可视化工具，主流机器学习算法的应用，强大的模型评估、试验和 Web 发布的工具。

本书概述了现代数据科学的理论和原理、相关工作流程，并且涵盖了许多现如今被广泛使用的机器学习算法。我们将使用现实生活中的真实数据构建一系列预测分析模型，评估不同的机器学习算法和建模策略，然后在数分钟之内便可将模型部署至 Azure 上以提供机器学习 Web 服务。本书还通过扩展 Azure 机器学习预测模型的实例，探讨如何用 Azure 机器学习 Web 服务创建服务器和各种客户端的应用程序。

本书给出的各种场景和端到端的示例旨在提供足够的信息，帮助你能够快速开始使用 Azure ML Studio，继而可以轻松地扩展实例中的场景来创建自己强大的预测分析实验。本书结尾详细介绍了如何应用“持续学习”的技术以编程的方式“重新训练”Azure 机器学习预测模型，而不需要任何人工干预。

读者对象

本书的重点是提供数据科学的基本信息和利用 Azure Machine Learning Studio 构建数据科学程序的原则和技术。本书重点面向数据科学家和工程师，以及初识云计算和机器学习的开发人员和 IT 专家。Azure 机器学习可以帮助机器学习的初学者像经验丰富的数据科学家那样快速地构建与测试机器学习解决方案。

本书通过具体的例子和详细的步骤演示来帮助读者理解当今被广泛使用的每一个关键的预测分析算法，以及它们在 Azure ML Studio 中相应的实现方法。这些材料不仅对没有 Azure 机器学习经验的开发者颇具意义，对于在数据科学领域有丰富经验的专家也非常有

用。不管是哪种情况，端到端的示例都能通过真实场景中具体的需求实现来帮助读者加强对机器学习概念的理解。虽然，本书各章在一定程度上彼此关联，但并不是非得完成某一章的动手实践后才能理解下一章，每一章都可独立阅读。

阅读前提

我们认为你至少了解部分云计算的概念和基本的 Web 服务。本书不要求读者具备较强的领域知识，但是如果你对每章主题所涉及的知识有所了解，将会获得更深入的理解。举例来说，如果你掌握一些 Web 开发的技能，那么关于创建 Azure ML 客户端和服务端应用程序的章节对你来说就会更容易理解。Azure Machine Learning Studio 为每个 Azure 机器学习实验自动生成使用预测分析 Web 服务的代码示例（用 C#、Python 和 R 语言实现）。如果你掌握 C#、Python 或者 R 中的任何一种开发语言，这对你都会很有帮助，但这并不是必要的前提。

本书可能不适合……

本书不适合那些想要探求数据科学算法所涵盖的深层数学和统计学理论的读者。本书的目标是尽可能地为广大读者传递 Azure Machine Learning Studio 的核心概念和实现细节，而这些读者可能不需要具备深厚的数学和统计学背景知识。

内容结构

本书探讨的主题如下。

- 第1章“数据科学介绍”，描述在数据科学领域，Azure机器学习如何通过实现完全托管的数据科学云服务迈出预测分析解决方案的关键一步。
- 第2章“Azure机器学习入门”，概括预测分析科学和方法论的基本概念。
- 第3章“Azure ML Studio的使用”，探讨Azure Machine Learning Studio的基本原理并帮助你开启数据科学的伟大之门。
- 第4章“创建Azure机器学习客户端应用程序和服务器应用程序”，基于一个实用的Azure机器学习预测模型，探讨可以用来调用Azure机器学习Web服务的客户端和服务端应用程序的类型。
- 第5章“回归分析”，深入探讨Azure ML Studio中提供一些复杂的机器学习算法。

- 第6章“聚类分析”，探讨数据分析的挖掘方案，包括自主数据分析、确定数据的相关性、推断逻辑分组，以及被广泛探讨的“从树木到森林”的算法处理混乱数据的方案。
- 第7章“Azure机器学习火柴盒推荐引擎”，介绍当今互联网中最强大和被广泛使用的预测分析的实现方法，以及其如何在众多消费行业中取得关键性的成功。
- 第8章“Azure机器学习模型重训练”，探索如何将“持续学习”纳入预测模型工作流的实现机制。

本书约定

为了使书中信息更具可读性和易于理解，本书使用以下的几个约定。

- 你可以在每个动手实践中遵循说明步骤创建特定的Azure资源。
- 当前Azure有两个管理门户网站：Azure管理门户和新的Azure预览版门户。本书在所有的操作中都使用原来的Azure管理门户。
- 两个按键之间的加号(+)表示这是组合键，需要同时按下。例如，“按Alt + Tab”意味着同时按下Alt和Tab键。

系统要求

本书中的许多示例，您只需借助互联网在浏览器（IE/O 或更高版本）中访问 Azure 管理门户即可。在第 4 章和其他几章中使用 Visual Studio 为你展示调用 Azure 机器学习 Web 服务的客户端应用程序和概念的示例。对于这些动手实践的示例，你需要事先安装 Visual Studio 2013。

系统要求如下。

- Windows 7 Service Pack 1、Windows 8、Windows 8.1、Windows Server 2008 R2 SP1、Windows Server 2012，或者Windows Server 2012 R2。
- 计算机至少支持1.6 GHz或者更快的处理器（建议2 GHz）。
- 1 GB（32位）或者2 GB（64位）内存（如果在虚拟机上运行则需增加512 MB 内存）。
- 20 GB的可用磁盘空间。
- 每分钟5400转的硬盘。

- 支持DirectX 9 的显卡，支持1024×768 或者更高分辨率的显示器。
- DVD-ROM驱动器（如果从DVD中安装Visual Studio）。
- 互联网连接。

根据你的 Windows 配置，你可能需要管理员权限才能安装或配置 Visual Studio 2013。

致谢

本书献给我的父亲，他在我编写本书期间永远地离开了我们，但他曾经英明地预见到计算机早晚会有“大动静”，并鼓励我在这个新领域中“踏浪前行”。到目前为止，浪潮仍未退去。

为了写这本书，我花费了许多夜晚和周末的时间。感谢我的妻子苏珊，她总能在我真正行动之前预知我想要做什么。感谢我的孩子瑞恩、布鲁克和尼古拉斯，他们给了我无尽的支持和鼓励。

特别感谢微软出版社团队给予我的支持和引导。最重要的是，与我的编辑德文·马斯格雷夫共事是非常愉悦的享受。他从本书构思开始就不断给予我建议和指导，最终促成本书与读者见面。

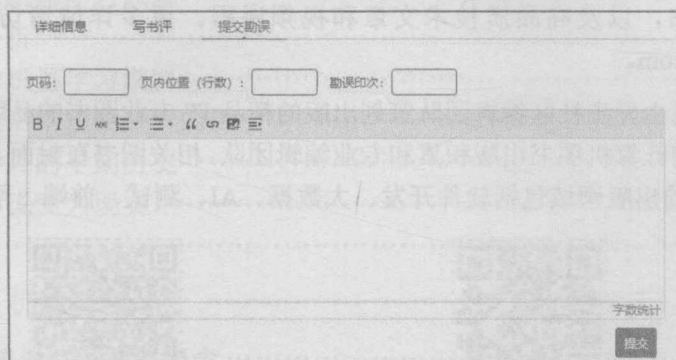
资源与支持

本书由异步社区出品，社区（<https://www.epubit.com/>）为您提供相关资源和后续服务。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，我们将赠予您异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。



与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件

给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 <https://www.epubit.com>。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等等。



异步社区



微信服务号

目录

- 第 1 章 数据科学介绍 1
 - 1.1 什么是机器学习 1
 - 1.2 当下的机器学习风暴 3
 - 1.3 预测分析 4
 - 1.4 无限的机器学习燃料 4
 - 1.5 日常生活中预测分析的例子 6
 - 1.6 机器学习的早期历史 7
 - 1.7 科幻小说变为现实 9
 - 1.8 总结 10
- 第 2 章 Azure 机器学习入门 11
 - 2.1 Azure机器学习核心概念 11
 - 2.2 先进的Azure机器学习工作流 12
 - 2.3 机器学习算法 13
 - 2.3.1 有监督学习 14
 - 2.3.2 无监督学习 18
 - 2.4 部署预测模型 19
 - 2.5 Azure机器学习带来的收益 19
 - 2.6 是什么，是怎样，为什么 20
 - 2.7 总结 21
- 第 3 章 Azure ML Studio 的使用 22
 - 3.1 Azure机器学习术语 22
 - 3.2 Azure机器学习入门 24
 - 3.3 Azure机器学习定价和可用性 26
 - 3.4 创建第一个Azure机器学习工作区 27

3.5	创建第一个Azure机器学习实验	31
3.6	从公共资源库下载数据集	31
3.7	数据上传至Azure机器学习实验	33
3.8	创建新的Azure机器学习实验	34
3.9	可视化数据集	36
3.10	分割数据集	40
3.11	模型训练	41
3.12	选择预测列	42
3.13	模型评分	44
3.14	模型计算结果的可视化	45
3.15	模型评估	46
3.16	保存实验	48
3.17	将训练的模型发布为Web服务准备工作	49
3.18	创建评分实验	52
3.19	将模型发布为Web服务	54
3.20	Azure机器学习Web服务的批处理	61
3.21	测试Azure机器学习Web服务	62
3.22	发布至Azure数据市场	64
3.23	总结	65
第4章	创建 Azure 机器学习客户端应用程序和服务器应用程序	66
4.1	为什么要创建Azure机器学习客户端应用程序	66
4.2	Azure机器学习 Web 服务的示例代码	68
4.3	C# 控制台应用程序示例代码	70
4.4	R的示例代码	75
4.5	不仅仅是简单的客户端	79
4.6	跨域资源共享和Azure机器学习Web服务	80
4.7	创建一个ASP.NET Azure机器学习Web客户端	80
4.8	让Azure机器学习Web服务的测试变得更简单	83
4.8.1	用户输入验证	84
4.8.2	用ASP.NET Web API创建一个Web服务	87
4.9	启用的CORS 支持	93
4.10	Web API Web 服务的处理逻辑	96
4.11	总结	105

第5章	回归分析	106
5.1	线性回归	106
5.2	Azure机器学习线性回归案例	107
5.2.1	下载汽车数据集	109
5.2.2	上传汽车数据集	110
5.2.3	创建汽车价格的实验	111
5.3	总结	124
第6章	聚类分析	125
6.1	非监督机器学习	125
6.1.1	聚类分析	126
6.1.2	KNN: K最近邻算法	127
6.2	Azure ML Studio聚类模块	127
6.2.1	聚类示例: 批发客户分组	128
6.2.2	发布K-Means聚类实验	135
6.3	总结	142
第7章	Azure 机器学习火柴盒推荐引擎	144
7.1	当今推荐引擎的应用	144
7.2	推荐引擎机制	146
7.3	Azure机器学习火柴盒推荐引擎后台	146
7.4	Azure机器学习火柴盒推荐引擎: 餐馆评分	148
7.5	创建餐馆评分的推荐引擎	149
7.6	创建火柴盒推荐引擎Web服务	156
7.7	总结	159
第8章	Azure 机器学习模型重训练	160
8.1	重训练Azure机器学习模型的工作流程	161
8.2	Azure Maching Learning Studio中的重训练模型	162
8.3	修改初始的训练实验	165
8.4	添加额外的网络节点	168
8.5	批处理服务重新训练模型	172
8.6	总结	174

第1章

数据科学介绍

欢迎来到令人兴奋的微软 Azure 机器学习新世界！不论你是一位专业的数据科学家还是有抱负的新手，都可以通过微软发布的基于云的平台来快速创建、分享、测试、训练、失败、修复、再培训，并以便于使用的 Web 服务形式部署功能强大的机器学习实验。所有的这些都基于最新的算法。利用这一平台，你可以通过在新数据集上不断地“训练”来微调你的实验结果。

比尔·盖茨曾说，“机器学习中的一个突破，其价值抵得上 10 个微软”。新的 Azure 机器学习服务以一种真正的基于云的差异化方式承担起这一令人激动的挑战，它利用一个强大的假设（一些大的数据集、有效的信用卡和浏览器），让今天的机器学习企业家们能学习如何从众多大数据仓库中获利。

1.1 什么是机器学习

机器学习可以被描述为一种利用经验来改进的计算系统。它也可以被描述为一种将数据转换为软件的方法。无论使用什么术语，结果仍然是相同的。数据科学家已经成功地开发了若干创建软件“模型”的方法，这些模型在用海量数据培训后，用于预测某些模式、趋势和结果。

预测分析学是 Azure 机器学习背后的一种基础技术，它可以被简单地定义为一种科学地利用过去预测未来的方式，以帮助得出预期的结果。

在某些情况下，机器学习和预测分析通常会特别有用，因为它们可以远远超过一般人开发的标准的规则引擎或编程逻辑。如果想使用实例或过去的历史经验数据对所需的输出或预测进行优化，机器学习是很好用的手段。描述机器学习的方法之一是把它与现代计算机编程范式进行比较。

传统的编程模式下，程序和数据通过计算机处理以产生所需的输出，例如用程序来处理数据并产生一个报告（见图 1-1）。

传统的编程

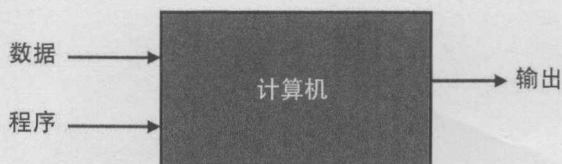


图 1-1 传统的编程范式

而用机器学习进行处理，处理范式便发生巨大的变化。数据和所需的输出是由计算机逆向设计的，来产生一个新的程序，如图 1-2 所示。

机器学习

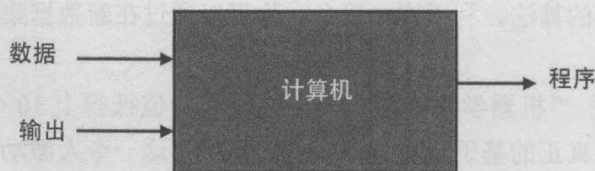


图 1-2 机器学习编程范例

这个新程序的作用是，它可以根据所提供的输入数据有效地“预测”输出，这种方法的主要优点是，所产生的“程序”是利用大量的学习数据和反馈数据训练构建起来的，能够基于提供数据预测期望的输出。从某种意义上说，这相当于有能力创造一只可以下金蛋的鹅！

一个关于预测分析的典型例子在 Amazon 网站上随处可见，你每搜索一件商品，页面会向你呈现一个推荐商品目录，这是因为“买过此类商品的顾客还买了目录里的商品”。这是利用预测分析和人类购买心理建立一个非常有效的营销策略的典型案列。

人类社会本能的需求之一便是不落后，跟随潮流。结合这些深层的心理动机与正确的历史交易数据以及优化的过滤算法，如何实现一个高效的电子商务销售策略便显而易见了。

人类自然本能之一就是害怕置身事外，特别是当其他人都在做的时候。这是社会网络的潜在基础，并且预测分析不会有比结合网络预测分析人类本性更有效用和效率了。结合人类固有的欲望心理与正确的历史交易数据，然后使用优化的过滤算法，便可以实现一种高效的电子商务促销策略。

接下来分析一下这个高度有效的预测算法对数据的要求。基本要求是历史订单，通过

这些订单，系统可以查找与当前浏览物品一起购买的其他物品。然后通过结合用户的个人资料，如年龄、性别、婚姻状况和邮政编码等，对历史数据进行过滤，这样便可以为用户推荐更有针对性的物品列表，甚至满足其他要求。

如果你能够根据他在过去购买的物品类别推断出用户的偏好和购买模式，事情会变得怎样？如果有人购买了弓、箭和野营炉，那么他可能是一位狩猎者。同样，这样的人极有可能喜欢户外活动以及与之相关的野营设备、皮卡车，甚至是棉花糖。

这种联合当前数据来推断额外的数据属性的模式才是数据科学真正起飞的所在，掌握有效利用这项技术的方法将为组织带来巨大的财富。与机器相比，数据科学家们能够做出更具价值的见解与推断，并将其与机器学习过程相结合，来最大化收益。

这也为什么在“快速失败”或确认自定义预测算法的逻辑的过程中，机器学习具备可以快速地测试预言和理论的能力。幸运的是，这是一个 Azure 机器学习真正闪光的领域。在后续章节中，我们将陆续学习如何快速创建、共享、部署和测试 Azure 机器学习实验，并迅速地部署组织中的预测分析系统。

从某种意义上说，Azure 机器学习类似于培养孩子或训练动物，当然它们不需要食物、水，也不需要休息。我们知道，持续地自适应改进是达尔文进化论的主要特征之一。与此类似，Azure 机器学习是计算理论和机器学习能力进展一座重要的里程碑。

机器学习可以与演化本身的许多概念做类比，尤其是在当有充足的时间和数据（现实世界中的经验）的情况下，自然界的有机体如何通过遗传和行为来适应和克服环境的变化这件事情上。

自然法则就是适者生存。

1.2 当下的机器学习风暴

由于以下行业趋势，当下流行的预测分析系统正在以远快于机器进化水平的速度发展。

- 数据以指数级的速度增长。
 - ◇ 我们实实在在地处在价值连城的历史交易数据的“大山”上，并且这些数据通常都被数字存储，容易读取。
 - ◇ 我们可以通过嵌入式系统和链接着各种设备的“物联网”演化系统获得日益丰富的实时数据。
 - ◇ 我们拥有了生成新的合成数据的能力，可以通过对现有历史数据的推导和预测来生成现实的模拟数据。
- 全面数字存储价格低廉。

◇ 如今，我们很容易从网络上获得大量免费的或低成本的、全球范围内可用的数字存储设备。

◇ 从个人设备到私有云和公有云，我们有多种存储机制来容纳永无止境的數據流。

- 普适计算能力。

◇ 如今，云计算服务无处不在，我们随时可以以具有竞争力的价格选择大量的云服务和托管合作伙伴。

◇ 易访问。只要拥有一张信用卡和一个浏览器，你便可以开始以小时或分钟计费的方式来访问自己所需要的所有内容。

- 数据分析的兴起。在许多现实世界的商业应用案例中，预测分析可以带来很好的收益，起到了推动经济的良好作用。

因此我们可以得出，机器学习一个很有趣的方面就是它总是自适应的，并且总是从任何错误或者误算中进行学习。因此，良好的反馈/校正循环系统对于微调预测模型是必不可少的。廉价的云存储设备的出现和不断发展的普适计算使得从数据中快速和高效地挖掘价值变得更加简单。

1.3 预测分析

当今，预测分析随处可见。当你意识到它在日常生活的正常消费中起了多么大的作用时，你甚至会感到恐惧。预测分析系统已经深深地融入了人们生活的方方面面。从保护你的电子邮件、预测你可能喜欢的电影，预测你将支付的保险费，到预测你下一次抵押贷款申请时的贷款利率，这项技术的使用一定程度上会决定收益。

俗话说“差之毫厘，谬以千里”。在当今这个时代下，任何机会都会转瞬即逝。敢于进行分析的人才能创建商业模型并且收获潜在的利益。这再次说明，数据分析预测能力将对我们的社会发挥越来越重要的作用，甚至会达到推动全新的商业模式形成和行业发展的作用，而这一切完全依赖于预测分析的力量和当下正以惊人速率产生的数据。

1.4 无限的机器学习燃料

随着数字时代的发展，当下大多数人时刻都在通过万维网、社交媒体、趣味图片创造着大量新的数字数据。据估计，目前全球数据增长情况是，每两天全世界创造的数据相当