TURING 图灵程序设计丛书

[PACKT] PUBLISHING

[斯洛文尼亚] Boštjan Kaluža 著 武传海 译

# Java 机器学习

## Machine Learning in Java

中国工信出版集团

人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

# Java机器学习

## Machine Learning in Java



[斯洛文尼亚] Boštjan Kaluža 著
武传海 译

人民邮电出版社
北京

**图书在版编目（CIP）数据**

Java机器学习 /（斯洛文）博思蒂安·卡鲁扎著 ；
武传海译. -- 北京 ：人民邮电出版社，2017.9
（图灵程序设计丛书）
ISBN 978-7-115-46680-8

Ⅰ. ①J… Ⅱ. ①博… ②武… Ⅲ. ①JAVA语言－程序设计 Ⅳ. ①TP312.8

中国版本图书馆CIP数据核字(2017)第202589号

## 内 容 提 要

本书介绍如何使用Java创建并实现机器学习算法，既有基础知识，又提供实战案例。主要内容包括：机器学习基本概念、原理，Weka、Mahout、Spark等常见机器学习库的用法，各类机器学习常见任务，包括分类、预测预报、购物篮分析、检测异常、行为识别、图像识别以及文本分析。最后还提供了相关Web资源、各种技术研讨会议以及机器学习挑战赛等进阶所需内容。

本书适合机器学习入门者，尤其是想使用Java机器学习库进行数据分析的读者。

◆ 著　[斯洛文尼亚] Boštjan Kaluža
译　武传海
责任编辑　陈　曦
责任印制　彭志环

◆ 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
三河市海波印务有限公司印刷

◆ 开本：800×1000　1/16
印张：11.5
字数：272千字　2017年9月第1版
印数：1-3 500册　2017年9月河北第1次印刷
著作权合同登记号　图字：01-2017-5590号

定价：49.00元

**读者服务热线：(010)51095186转600　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**

## 作者简介

Boštjan Kaluža

博士，人工智能与机器学习专家，IT运营分析公司Evolven首席数据科学家，主攻机器学习、预测分析、模式挖掘与异常检测，旨在将数据转化为人类可理解的信息与可应用的知识。个人网址：http://bostjankaluza.net。

微信连接

回复“机器学习”“Java”查看相关书单

微博连接

关注@图灵教育 每日分享IT好书

QQ连接

图灵读者官方群I：218139230
图灵读者官方群II：164939616

**图灵社区**
**iTuring.cn**

在线出版，电子书，《码农》杂志，图灵访谈

站在巨人的肩上

Standing on Shoulders of Giants


iTuring.cn

# 前　言

机器学习是人工智能的一个分支，它在算法与数据的协助下，让计算机像人类一样学习和行动。针对给定的数据集，机器学习算法会学习数据的不同属性，并对以后可能遇到的数据属性进行推断。

本书教你如何使用Java创建并实现机器学习算法，既有基础概念的讲解，也有示例供你学习。当然，还会介绍一些常用的机器学习库，如Weka、Apache Mahout、Mallet等。阅读本书后，你将懂得如何为特定问题选择合适的机器学习方法，以及如何比较与评估不同技术的优劣。书中还会讲解性能提升技术，包括输入预处理以及合并不同方法产生的输出。

我们将探讨使用Java库进行机器学习的技术细节，并配有清晰易懂的示例。同时，你还将学习如何准备要分析的数据、如何选择机器学习方法，以及如何衡量流程的有效性。

## 本书内容

**第1章　机器学习应用快速入门**。讲解机器学习的基础知识、常见概念、原理，以及机器学习的应用流程。

**第2章　面向机器学习的Java库与平台**。介绍各种机器学习专用的Java库与平台。你将了解每个库提供的功能，以及可以用于解决哪些问题。涉及的机器学习库有Weka、Java-ML、Apache Mahout、Apache Spark、Deeplearning4j和Mallet。

**第3章　基本算法——分类、回归和聚类**。从最基本的机器学习任务入手，使用小巧又易懂的数据集，介绍分类、回归和聚类的关键算法。

**第4章　利用集成方法预测客户关系**。深入研究一个真实的客户营销数据库，目标是对可能流失以及可进行追加推销与交叉推销的客户进行预测。我们将使用集成方法解决这个问题，并且采用在KDD Cup竞赛中获胜的解决方案。

**第5章　关联分析**。讲解如何使用关联规则挖掘分析共生关系。我们将通过“购物篮分析”了解顾客的购买行为，并讨论如何将这种方法应用到其他领域。

**第6章 使用Apache Mahout制作推荐引擎**。讲解一些基本概念，帮助你了解推荐引擎原理，然后利用Apache Mahout实现两个应用——基于内容的过滤与协同推荐器。

**第7章 欺诈与异常检测**。介绍异常和可疑模式检测的背景，然后讲解两个实际应用——保险索赔欺诈检测与网站流量异常检测。

**第8章 利用Deeplearning4j进行图像识别**。介绍图像识别与基本的神经网络架构，然后讨论如何利用Deeplearning4j库实现各种深度学习架构，以实现对手写体数字的识别。

**第9章 利用手机传感器进行行为识别**。借助传感器数据解决模式识别问题。这一章介绍行为识别过程，讲解如何使用Android设备收集数据，并提出一个分类模型以对日常生活行为进行识别。

**第10章 利用Mallet进行文本挖掘——主题模型与垃圾邮件检测**。讲解文本挖掘的基础知识，介绍文本处理管道，演示如何将其应用于两个实际问题（主题建模与文档分类）。

**第11章 机器学习进阶**。这是全书最后一章，提供关于如何部署模型的实用建议，并进一步给出提示，告诉你去哪里寻找更多资源、资料、场所和技术，以便深入了解机器学习。

## 阅读前提

为了实际运行书中示例，你需要一台安装有JDK的个人计算机。所有能下载的示例与源代码都假定你使用的是支持Maven（一个依赖管理与自动创建工具）与Git（版本控制系统）的Eclipse IDE开发环境。各章示例依赖Weka、Deeplearning4j、Mallet、Apache Mahout等各种库。关于如何获取与安装这些库，会在各章首次用到它们时进行讲解。

## 读者对象

本书为那些想学习如何使用Java机器学习库进行数据分析的人而写。或许你已经对机器学习有了一点了解，但从未用过Java；又或许你懂得一点Java，而在机器学习方面是个新手。不论你属于哪种情况，本书都能让你快速上手，并提供必需的技能，让你能够成功创建、定制，以及在实际生活中部署机器学习应用。如果你懂得一点基本的编程知识以及数据挖掘相关概念会更好，但不要求你必须拥有与数据挖掘程序包相关的开发经验。

## 配套资料

本书专门配有一个在线支持网站（http://machine-learning-in-java.com），从中可以找到所有示例代码、勘误表，以及其他入门资料。

## 排版约定

本书中，你会发现许多不同体例，它们用于区分不同类型的信息。下面给出一些例子，并对其含义进行说明。

正文中的代码、数据库表名、文件夹名、文件名、文件扩展名、路径名、伪URL、用户输入和Twitter用户名显示如下：

“比如，`Bob`拥有`height`、`eye color`、`hobbies`三个属性，对应值依次为`185cm`、`blue`、`climbing`与`sky diving`”。

代码块表示如下：

```
Bob={
height: 185cm,
eye color: blue,
hobbies: climbing, sky diving
}
```

所有命令行输入或输出写成如下形式：

```
12,3,7,2,0,1,8,9,13,4,11,5,15,10,6,14,16
```

**新术语**与**重要词语**使用黑体显示。你在屏幕上看到的词，比如菜单或对话框中的词，在正文中显示如下：“在项目属性上点击鼠标右键，选择**Java Build Path**，单击**Libraries**选项卡，选择**Add External JARs**。”

警告或重要的注意事项。

提示和技巧。

## 读者反馈

欢迎提供反馈，请将你对本书的看法告诉我们：哪些方面是你喜欢的，哪些方面你不喜欢。读者反馈对我们来说很重要，因为这可以帮助我们推出更符合读者需求的著作。

要给我们提供反馈，只需向feedback@packtpub.com发送电子邮件，并在邮件主题中指出书名。

如果你有擅长的主题，并有志于写书或撰稿，请参阅www.packtpub.com/authors的撰稿指南。

## 读者支持

购买本社图书后，你将获得各种帮助，让手中图书最大限度地发挥功效。

### 下载示例代码

你可以从本书配套网站（http://machine-learning-in-java.com）下载书中示例代码。导航到Downloads版块，点击到Git仓库链接。

当然，你也可以使用自己的账号登录http://www.packpub.com网站下载本书示例代码。不论你在哪里购买本书，都可以访问http://www.packtpub.com/support。注册之后，我们会使用电子邮件将示例代码直接发送给你。

下载示例代码时，请按照如下步骤进行：

(1) 使用你的电子邮件地址与密码登录我们的网站，如果尚未加入会员，请先注册加入；
(2) 移动鼠标到SUPPORT菜单之上；
(3) 点击Code Downloads & Errata；
(4) 在Search文本框中输入书名；
(5) 选择你想下载代码文件的图书；
(6) 从下拉菜单中选择你购买本书的地点；
(7) 点击Code Download。

示例文件下载完成后，请使用如下最新版本的解压缩软件对文件进行解压。

- WinRAR/7-Zip for Windows
- Zipeg/iZip/UnRarX for Mac
- 7-Zip/PeaZip for Linux

## 勘误

虽然我们力图让图书内容准确无误，但错误仍不可避免。如果你在本社图书中发现错误（包括正文和代码），请告诉我们，我们将感激不尽。你这样做不仅可以让其他读者免遭同样的挫折，还可帮助我们改进该书的后续版本。无论你发现什么错误，都请告诉我们。为此，可以访问http://www.packtpub.com/submit-errata，输入书名，单击链接Errata Submission Form，再输入错误详情。提交的勘误得到确认后，将被上传到我们的网站或添加到既有的勘误列表。

要查看已提交的勘误，请访问https://www.packtpub.com/books/content/support，并在搜索框中输入书名，Errata栏将列出你搜索的信息。

## 打击盗版

网上散布的盗版材料是各类媒体屡禁不绝的问题。在保护版权和许可方面，本社的态度非常严肃。如果你在网上看到本社作品的非法复制品，请马上把网址或网站名告诉我们，以便我们能够采取补救措施。

请通过copyright@packtpub.com与我们取得联系，并提供可疑的盗版材料链接。

感谢你为保护我们的作者提供的帮助，也十分感激对于我们提供有价值内容的能力给予的保护。

## 答疑

只要有与本书相关的问题，都可通过questions@packtpub.com与我们联系，我们将尽力解决。

## 电子书

扫描如下二维码，即可购买本书电子版。

# 目　录


第 1 章　机器学习应用快速入门 …… 1

1.1　机器学习与数据科学 …… 1

1.1.1　机器学习能够解决的问题 …… 2

1.1.2　机器学习应用流程 …… 3

1.2　数据与问题定义 …… 4

1.3　数据收集 …… 5

1.3.1　发现或观察数据 …… 5

1.3.2　生成数据 …… 6

1.3.3　采样陷阱 …… 7

1.4　数据预处理 …… 7

1.4.1　数据清洗 …… 8

1.4.2　填充缺失值 …… 8

1.4.3　剔除异常值 …… 8

1.4.4　数据转换 …… 9

1.4.5　数据归约 …… 10

1.5　无监督学习 …… 10

1.5.1　查找相似项目 …… 10

1.5.2　聚类 …… 12

1.6　监督学习 …… 13

1.6.1　分类 …… 14

1.6.2　回归 …… 16

1.7　泛化与评估 …… 18

1.8　小结 …… 21

第 2 章　面向机器学习的 Java 库与平台 …… 22

2.1　Java 环境 …… 22

2.2　机器学习库 …… 23

2.2.1　Weka …… 23

2.2.2　Java 机器学习 …… 25

2.2.3　Apache Mahout …… 26

2.2.4　Apache Spark …… 27

2.2.5　Deeplearning4j …… 28

2.2.6　MALLET …… 29

2.2.7　比较各个库 …… 30

2.3　创建机器学习应用 …… 31

2.4　处理大数据 …… 31

2.5　小结 …… 33

第 3 章　基本算法——分类、回归和聚类 …… 34

3.1　开始之前 …… 34

3.2　分类 …… 35

3.2.1　数据 …… 35

3.2.2　加载数据 …… 36

3.2.3　特征选择 …… 37

3.2.4　学习算法 …… 38

3.2.5　对新数据分类 …… 40

3.2.6　评估与预测误差度量 …… 41

3.2.7　混淆矩阵 …… 41

3.2.8　选择分类算法 …… 42

3.3　回归 …… 43

3.3.1　加载数据 …… 43

3.3.2　分析属性 …… 44

3.3.3　创建与评估回归模型 …… 45

3.3.4　避免常见回归问题的小技巧 …… 48

3.4　聚类 …… 49

3.4.1　聚类算法 …… 49

3.4.2　评估 …… 50

3.5　小结 …… 51

第 4 章 利用集成方法预测客户关系……52
4.1 客户关系数据库……52
4.1.1 挑战……53
4.1.2 数据集……53
4.1.3 评估……54
4.2 最基本的朴素贝叶斯分类器基准……55
4.2.1 获取数据……55
4.2.2 加载数据……56
4.3 基准模型……58
4.3.1 评估模型……58
4.3.2 实现朴素贝叶斯基准线……59
4.4 使用集成方法进行高级建模……60
4.4.1 开始之前……60
4.4.2 数据预处理……61
4.4.3 属性选择……62
4.4.4 模型选择……63
4.4.5 性能评估……66
4.5 小结……66
第 5 章 关联分析……67
5.1 购物篮分析……67
5.2 关联规则学习……69
5.2.1 基本概念……69
5.2.2 Apriori 算法……71
5.2.3 FP-增长算法……71
5.2.4 超市数据集……72
5.3 发现模式……73
5.3.1 Apriori 算法……73
5.3.2 FP-增长算法……74
5.4 在其他领域中的应用……75
5.4.1 医疗诊断……75
5.4.2 蛋白质序列……75
5.4.3 人口普查数据……76
5.4.4 客户关系管理……76
5.4.5 IT 运营分析……76
5.5 小结……77
第 6 章 使用 Apache Mahout 制作推荐引擎……78
6.1 基本概念……78
6.1.1 关键概念……79
6.1.2 基于用户与基于项目的分析……79
6.1.3 计算相似度的方法……80
6.1.4 利用与探索……81
6.2 获取 Apache Mahout……81
6.3 创建一个推荐引擎……84
6.3.1 图书评分数据集……84
6.3.2 加载数据……84
6.3.3 协同过滤……89
6.4 基于内容的过滤……97
6.5 小结……97
第 7 章 欺诈与异常检测……98
7.1 可疑与异常行为检测……98
7.2 可疑模式检测……99
7.3 异常模式检测……100
7.3.1 分析类型……100
7.3.2 事务分析……101
7.3.3 规划识别……101
7.4 保险理赔欺诈检测……101
7.4.1 数据集……102
7.4.2 为可疑模式建模……103
7.5 网站流量异常检测……107
7.5.1 数据集……107
7.5.2 时序数据中的异常检测……108
7.6 小结……113
第 8 章 利用 Deeplearning4j 进行图像识别……114
8.1 图像识别简介……114
8.2 图像分类……120
8.2.1 Deeplearning4j……120
8.2.2 MNIST 数据集……121
8.2.3 加载数据……121
8.2.4 创建模型……122
8.3 小结……128

第 9 章　利用手机传感器进行行为识别 ……129
9.1　行为识别简介 ……129
9.1.1　手机传感器 ……130
9.1.2　行为识别流水线 ……131
9.1.3　计划 ……132
9.2　从手机收集数据 ……133
9.2.1　安装 Android Studio ……133
9.2.2　加载数据采集器 ……133
9.2.3　收集训练数据 ……136
9.3　创建分类器 ……138
9.3.1　减少假性转换 ……140
9.3.2　将分类器嵌入移动应用 ……142
9.4　小结 ……143
第 10 章　利用 Mallet 进行文本挖掘——主题模型与垃圾邮件检测 ……144
10.1　文本挖掘简介 ……144
10.1.1　主题模型 ……145
10.1.2　文本分类 ……145
10.2　安装 Mallet ……146
10.3　使用文本数据 ……147
10.3.1　导入数据 ……149
10.3.2　对文本数据做预处理 ……150
10.4　为 BBC 新闻做主题模型 ……152
10.4.1　BBC 数据集 ……152
10.4.2　建模 ……153
10.4.3　评估模型 ……155
10.4.4　重用模型 ……156
10.5　垃圾邮件检测 ……157
10.5.1　垃圾邮件数据集 ……158
10.5.2　特征生成 ……159
10.5.3　训练与测试模型 ……160
10.6　小结 ……161
第 11 章　机器学习进阶 ……162
11.1　现实生活中的机器学习 ……162
11.1.1　噪声数据 ……162
11.1.2　类不平衡 ……162
11.1.3　特征选择困难 ……163
11.1.4　模型链 ……163
11.1.5　评价的重要性 ……163
11.1.6　从模型到产品 ……164
11.1.7　模型维护 ……164
11.2　标准与标记语言 ……165
11.2.1　CRISP-DM ……165
11.2.2　SEMMA 方法 ……166
11.2.3　预测模型标记语言 ……166
11.3　云端机器学习 ……167
11.4　Web 资源与比赛 ……168
11.4.1　数据集 ……168
11.4.2　在线课程 ……169
11.4.3　比赛 ……170
11.4.4　网站与博客 ……170
11.4.5　场馆与会议 ……171
11.5　小结 ……171

# 第1章 机器学习应用快速入门

本章介绍机器学习的基础知识，包括常见主题与概念，这些内容将让你更容易理解相关逻辑以及所讲主题。本章的目标是让你快速了解**应用机器学习**的详细步骤，掌握机器学习的主要原理。本章涵盖以下内容：

- 介绍机器学习及其与数据科学的关系
- 讨论机器学习应用的基本步骤
- 讨论所处理数据的类型及其重要性
- 讨论收集数据以及对数据进行预处理的方法
- 使用机器学习理解数据
- 使用机器学习从数据获取有用信息并创建预测器

如果你已经熟悉机器学习，并急于开始编写代码，请跳过本章内容，直接阅读其他章节。然而，如果你想重温这些内容或者搞清一些概念，强烈建议你认真学习本章。

## 1.1 机器学习与数据科学

如今，每个人都在谈论机器学习与数据科学。那么，机器学习究竟是什么？它与数据科学有着怎样的联系呢？这两个术语常被混淆，因为它们经常使用相同的方法，有着明显的重叠。因此，下面先区分这两个术语。

Josh Wills在Twitter上说：

> “所谓的数据科学家指这样一类人，他们比软件工程师更懂统计学，比统计学家更懂软件工程。”

更具体地说，数据科学包含从数据获取知识的整个过程，它综合运用统计学、计算机科学以及其他领域的各种方法，帮助人们从数据中获取有用的知识与信息。事实上，数据科学是一个不断反复的过程，包括数据的采集、清洗、分析、可视化和部署。

另一方面，机器学习主要涉及数据科学的分析与建模阶段使用的通用算法与技术。对于机器学习，Arthur Samuel在1959年提出如下定义：

> “机器学习是指研究、设计与开发某些算法，让计算机获得学习的能力，而不需要明确的编程。”

### 1.1.1　机器学习能够解决的问题

机器学习方法主要有如下三种：

- 监督学习
- 无监督学习
- 强化学习

给定一组样本输入$X$与它们的结果$Y$，监督学习的目标是产生一个通用的映射函数$f$，使得每一个输入都有对应的确定输出，即$f: X Y$。

监督学习的一个应用例子是检测信用卡欺诈。学习算法会学习所有带有“正常”或“可疑”标记（向量$Y$）的信用卡交易（矩阵$X$），并最终产生一个决策模型（即$f$函数），对未见过的交易打标记（“正常”或“可疑”）。

相反，无监督学习算法所学的数据没有给定的结果标签Y，它主要学习数据的结构，比如将相似的输入数据归入某个聚类。因此，使用无监督学习能够发现隐藏在数据中的模式。无监督学习的一个应用例子是基于物品（Item-based）的推荐系统，其学习算法会发现购物者一同购买的相似商品，比如购买了书A的人也购买了书B。

强化学习从完全不同的角度处理学习过程。它假设有一个智能体（agent，可以是机器人、自动程序或计算机程序）与动态环境进行交互，以实现某个特定目标。环境由一组状态描述，智能体可以做出不同行为，以从一种状态变为另一种状态。有些状态被标为目标状态，如果智能体实现了这种状态，就会获得很大的奖励；而在其他状态中，智能体得到的奖励很少或没有，甚至还会被“惩罚”。强化学习的目标是找到最优策略，即映射函数，指定每个状态要采取的行为动作，而没有指导者（teacher）明确告知这样做是否会实现目标状态。强化学习的一个例子是汽车自动驾驶程序，其中“状态”与“驾驶条件”（比如当前速度、路况信息、周围交通状况、速度限制、路障）相对应，行为会驱动汽车做出诸如左转、右转、停止、加速、前行等动作。学习算法会产生一个策略，指定汽车在特定驾驶条件下要采取的动作。

本书中，我们把学习重点放在监督学习与无监督学习上，因为它们有许多相同的概念。如果强化学习激起了你的兴趣，建议阅读Richard S. Sutton与Andrew Barto二人合写的*Reinforcement Learning: An Introduction*，它是一本很好的入门书。

### 1.1.2 机器学习应用流程

本书讲解的重点是机器学习的应用。我们将提供切实可用的技能，帮助你顺利地将学习算法应用于各种不同的情景设置。相比于机器学习相关的理论与数学知识，我们将把更多时间用来学习如何将机器学习技术应用于具体实践。借助这些实际应用技术，你可以让机器学习在具体应用中发挥强大作用。我们把讲解重点放在监督学习与无监督学习上，涵盖从数据科学到创建机器学习应用流程的所有必需步骤。

标准的机器学习应用流程就是回答一系列问题，可概括成如下5个步骤。

(1) **数据与问题定义**：第一步是要问一些有趣的问题。你试图解决的问题是什么？它为何重要？哪种形式的结果能够回答你的问题？回答是简单的“是与否”吗？你需要从现有问题中挑选吗？

(2) **数据收集**：一旦有问题要解决，你就需要使用相关数据。问一问自己，哪种数据能够帮助你回答问题。你能从现有可用来源获取所需数据吗？你必须对多个来源进行合并吗？你必须生产数据吗？存在取样偏差吗？你需要多少数据？

(3) **数据预处理**：数据预处理的第一项任务是数据清洗，比如填补缺失值、平整噪声数据、移除异常值、解决数据一致性。通常，随后会有对多个数据源的整合以及数据转换，包括把数据转换到特定范围（数据标准化）、将数据分成一系列值段（数据离散化）、降低数据维数。

(4) **利用无监督学习与监督学习进行数据分析与建模**：数据分析与建模包括无监督机器学习与监督机器学习、统计推断和预测。这个阶段有多种机器学习算法可供选用，比如k最近邻算法、朴素贝叶斯算法、决策树、支持向量机、逻辑回归、K均值算法等。至于选用哪种算法，取决于第一步中提到的问题定义以及所收集的数据类型。经过这一步，我们最终会从数据推导出模型。

(5) **模型评价**：最后一步是对模型进行评价。通过机器学习创建模型后，接下来检查模型对源数据的拟合程度。如果模型的针对性太强，即对训练数据过拟合，那么它对新数据的预测效果就很有可能比较差；如果模型太通用，这意味着模型对训练数据欠拟合。比如，向欠拟合的天气预测模型询问加利福尼亚州的天气时，得到的回答总是“晴朗”。大多数时候这个回答是对的，但就有效预测天气来说，这个模型真的没什么用。这一步的目标是准确评价模型，确保模型面对新数据也能正常工作。进行模型评价时，常用的方法有独立测试、训练集、交叉验证、留一法交叉验证。

接下来，我们将详细讲解每个步骤，并且尝试理解机器学习应用流程过程中必须回答的问题类型，还要了解与数据分析、评估相关的概念。