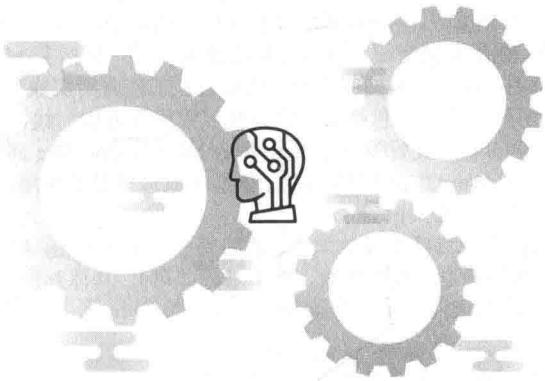
Broadview®
www.broadview.com.cn

机器学习基础

从入门到求职

胡欢武◎编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



机器学习基础

从入门到求职

胡欢武◎编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书是一本机器学习算法方面的理论+实践读物，主要包含机器学习基础理论、线性回归模型、分类模型、聚类模型、降维模型和深度神经网络模型六大部分。机器学习基础理论部分包含第 1、2 章，主要介绍机器学习的理论基础和工程实践基础。第 3 章是线性回归模型部分，主要包括模型的建立、学习策略的确定和优化算法的求解过程，最后结合三种常见的线性回归模型实现了一个房价预测的案例。第 4 至 11 章详细介绍了几种常见的分类模型，包括朴素贝叶斯模型、K 近邻模型、决策树模型、Logistic 回归模型、支持向量机模型、随机森林模型、AdaBoost 模型和提升树模型，每一个模型都给出了较为详细的推导过程和实际应用案例。第 12 章系统介绍了五种常见的聚类模型，包括 K-Means 聚类、层次聚类、密度聚类、谱聚类和高斯混合聚类，每一个模型的原理、优缺点和工程应用实践都给出了较为详细的说明。第 13 章系统介绍了四种常用的降维方式，包括奇异值分解、主成分分析、线性判别分析和局部线性嵌入，同样给出了详细的理论推导和分析。最后两章分别是 Word2Vec 和 Doc2Vec 词向量模型和深度神经网络模型，其中，第 14 章详细介绍了 Word2Vec 和 Doc2Vec 模型的原理推导和应用；第 15 章深度神经网络模型系统介绍了深度学习相关的各类基础知识。

本书适合对人工智能和机器学习感兴趣的学生、求职者和已工作人士，以及想要使用机器学习这一工具的跨行业者（有最基本的高等数学、线性代数、概率基础即可），具体判别方法建议您阅读本书的前言。

图书在版编目（CIP）数据

机器学习基础：从入门到求职 / 胡欢武编著. —北京：电子工业出版社，2019.4
ISBN 978-7-121-35521-9

Ⅰ. ①机… Ⅱ. ①胡… Ⅲ. ①机器学习 Ⅳ.①TP181

中国版本图书馆CIP数据核字(2018)第252505号

责任编辑：安　娜
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：24　字数：417.7 千字
版　　次：2019 年 4 月第 1 版
印　　次：2019 年 4 月第 1 次印刷
定　　价：89.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

前言

首先解答读者可能产生的一个疑问：本书的书名是《机器学习基础：从入门到求职》，但本书几乎通篇都在讲机器学习各种模型的原理推导和应用实例，这是为什么呢？其实本书的定位是帮助求职者快速入门并掌握机器学习相关的基础核心知识，降低学习成本，节省更多的时间。

为什么要这样做呢？原因也很简单。机器学习算法相关的岗位待遇比一般的开发岗位要好一些，但要求也变得更多。从目前的行情来看，站在公司招聘的角度，是一个既要、又要、还要的过程，即：既要掌握比较扎实的机器学习理论基础，又要有实践经验、懂业务场景，还要能编码、会计算机算法题。

对求职者来说，要求确实是太高了些。但这个岗位待遇好，有前途，也有“钱途”，因而很多人都报以极高的热情涌入，导致这个行业的招聘水涨船高，毕竟企业永远都是择优而选，优中取优！亲历过这几年求职或招聘的人可能会比较有感触：

2015 年，机器学习在国内市场刚兴起的时候，懂机器学习算法的人不多，那时候企业招聘，只要是懂些皮毛的，可能都有机会去试一试。

2016 年，市场开始火热，只懂些皮毛就不行了，必须还要懂得比较系统一点，要求求职者能够“手推”模型原理，再附加一些业务实践经验和计算机基础知识。

2017 年，招聘的人不仅会问“手推”算法原理，还会细问项目内容及对业务的理解，再附加两道算法题。

2018 年，招聘方希望你既要像算法工程师一样能“手推”模型原理，又要像传统程序员一样会写代码，还要像有工作经验的员工一样，有一些比较拿得出手的项目。

看到这里，如果你被这么多的要求吓到了，那么恭喜你，借这个心理转换过程重新定位自己，你可以学习本书；如果你决定迎难而上，那么也恭喜你，借这个机会赶紧查漏补缺，同样可以阅读本书。但是，如果你是计算机科班出身，已经信心满满，手握重点高校学历，拥有重大科研项目经历及各种大厂实习经验，还有多篇“顶会”论文，那么这本书真的不适合你。

回到关于本书的定位问题上。上面说了既要、又要、还要的过程，也就是理论基础+业务能力+工程实践能力的过程。理论基础就是我们一直所说的机器学习算法理论，业务能力是指相关的项目或者工作经验，工程实践能力就是动手写代码的能力。对于一个想求职机器学习相关岗位的应届生，或者是想将机器学习应用到自己专业领域的人士，再或者是一个有一定编程经验想要转算法岗位的人来说，机器学习理论可能都是第一拦路虎。本书希望可以帮助读者用最短的时间、最少的精力，攻克这最难的一关。所以，再次提醒大家，本书并没有讲述如何面试求职，而是可以带你快速入门并应用机器学习，带你走近机器学习求职的起点，帮你节省一些学习和摸索的时间，本书并不是一本机器学习岗位求职大全，也绝非是你求职准备的终点。

如果看到这里，还不确定是否适合学习本书，那么看看本书的“机器学习求职 60 问”吧，这些都是求职过程中可能遇到的高频问题，也是机器学习需要掌握的核心理论基础，而这些问题，在本书中都有较为详细的推导和解答。如果你看了这些问题以后觉得都已经掌握了，那么本书不适合你。如果对一半以上问题觉得没什么概念或者似懂非懂，那么建议你看一看本书，相信你会有所收获！

机器学习求职 60 问

类型一：基础概念类

问题 1：过拟合与欠拟合（定义、产生的原因、解决的方法各是什么）。

问题 2：L1 正则与 L2 正则（有哪些常见的正则化方法？作用各是什么？区别是什么？为什么加正则化项能防止模型过拟合）。

问题 3：模型方差和偏差（能解释一下机器学习中的方差和偏差吗？哪些模型是降低模型方差的？哪些模型是降低模型偏差的？举例说明一下）。

问题 4：奥卡姆剃刀（说一说机器学习中的奥卡姆梯刀原理）。

问题 5：模型评估指标（回归模型和分类模型各有哪些常见的评估指标？各自的含义是什么？解释一下 AUC？你在平时的实践过程中用到过哪些评估指标？为什么要选择这些指标）。

问题 6：风险函数（说一下经验风险和结构风险的含义和异同点）。

问题 7：优化算法（机器学习中常见的优化算法有哪些？梯度下降法和牛顿法的原理推导）。

问题 8：激活函数（神经网络模型中常用的激活函数有哪些？说一下各自的特点）。

问题 9：核函数（核函数的定义和作用是什么？常用的核函数有哪些？你用过哪些核函数？说一下高斯核函数中的参数作用）。

问题 10：梯度消失与梯度爆炸（解释一下梯度消失与梯度爆炸问题，各自有什么解决方案）。

问题 11：有监督学习和无监督学习（说一下有监督学习和无监督学习的特点，

举例说明一下）。

问题 12：生成模型与判别模型（你知道生成模型和判别模型吗？各自的特点是什么？哪些模型是生成模型，哪些模型是判别模型）。

类型二：模型原理类

问题 13：线性回归（线性回归模型的原理、损失函数、正则化项）。

问题 14：KNN 模型（KNN 模型的原理、三要素、优化方案以及模型的优/缺点）。

问题 15：朴素贝叶斯（朴素贝叶斯模型的原理推导，拉普拉斯平滑，后验概率最大化的含义以及模型的优/缺点）。

问题 16：决策树（决策树模型的原理、特征评价指标、剪枝过程和原理、几种常见的决策树模型、各自的优/缺点）。

问题 17：随机森林模型（RF 模型的基本原理，RF 模型的两个“随机”。从偏差和方差角度说一下 RF 模型的优/缺点，以及 RF 模型和梯度提升树模型的区别）。

问题 18：AdaBoost（AdaBoost 模型的原理推导、从偏差和方差角度说一下 AdaBoost、AdaBoost 模型的优/缺点）。

问题 19：梯度提升树模型（GBDT 模型的原理推导、使用 GBDT 模型进行特征组合的过程、GBDT 模型的优/缺点）。

问题 20：XGBoost（XGBoost 模型的基本原理、XGBoost 模型和 GBDT 模型的异同点、XGBoost 模型的优/缺点）。

问题 21：Logistic 回归模型（LR 模型的原理、本质，LR 模型的损失函数，能否使用均方损失、为什么）。

问题 22：支持向量机模型（SVM 模型的原理，什么是“支持向量”？为什么使用拉格朗日对偶性？说一下 KKT 条件、软间隔 SVM 和硬间隔 SVM 的异同点。SVM 怎样实现非线性分类？SVM 常用的核函数有哪些？SVM 模型的优/缺点各是什么）。

问题 23：K-Means 聚类（K-Means 聚类的过程和原理是什么？优化方案有哪些？各自优/缺点是什么）。

问题 24：层次聚类（层次聚类的过程、原理和优/缺点）。

问题 25：密度聚类（密度聚类的基本原理和优/缺点）。

问题 26：谱聚类（谱聚类的基本原理和优/缺点）。

问题 27：高斯混合聚类（高斯混合聚类的原理和优/缺点）。

问题 28：EM 算法（EM 算法的推导过程和应用场景）。

问题 29：特征分解与奇异值分解（特征分解与奇异值分解的原理、异同点、应用场景）。

问题 30：主成分分析（PCA 模型的原理、过程、应用场景）。

问题 31：线性判别分析（LDA 模型的原理、过程、应用场景）。

问题 32：局部线性嵌入（LLE 模型的原理、过程、应用场景）。

问题 33：词向量（Word2Vec 模型和 Doc2Vec 模型的类别，各自原理推导、应用和参数调节）。

问题 34：深度神经网络（深度神经网络模型的原理，反向传播的推导过程，常用的激活函数，梯度消失与梯度爆炸问题怎么解决？说一下深度神经网络中的 Dropout、早停、正则化）。

类型三：模型比较类

问题 35：LR 模型与 SVM 模型的异同点。

问题 36：LR 模型与朴素贝叶斯模型的异同点。

问题 37：K 近邻模型与 K-Means 模型的异同点。

问题 38：ID3 决策树、C4.5 决策树、CART 决策树的异同点。

问题 39：PCA 模型与 LDA 模型的异同点。

问题 40：Bagging 模型与 Boosting 模型的异同点。

问题 41：GBDT 模型与 XGBoost 模型的异同点。

问题 42：Word2Vec 模型中 CWOB 模式与 Skip 模式的异同点。

问题 43：Word2Vec 模型和 Doc2Vec 模型的异同点。

类型四：模型技巧类

问题 44：模型调参（随便选一个上述涉及的模型，说一下它的调参方式与过程）。

问题 45：特征组合（常见的特征组合方式有哪些？各自特点是什么）。

问题 46：特征工程（结合实践解释一下你所理解的特征工程）。

问题 47：缺失值问题（说一下你遇到的缺失值处理问题，你知道哪些缺失值处理方式？你使用过哪些，效果怎样）。

问题 48：样本不平衡问题（你知道样本不平衡问题吗？你是怎样处理的？效果怎么样？除上采样和下采样外，你还能自己设计什么比较新颖的方式吗）。

问题 49：特征筛选（特征筛选有哪几种常见的方式？结合自己的实践经验说一下各自的原理和特点。）

问题 50：模型选择（你一般怎样挑选合适的模型？有实际的例子吗？）

问题 51：模型组合（你知道哪些模型组合方式？除了运用 AdaBoost 和 RF，你自己有使用过 Bagging 和 Embedding 方式组合模型吗？结合实际例子说明一下）。

问题 52：A/B 测试（了解 A/B 测试吗？为什么要使用 A/B 测试）。

问题 53：降维（为什么要使用降维？你知道哪些降维方法？你用过哪些降维方式？结合实际使用说明一下）。

问题 54：项目（你做过哪些相关的项目？挑一个你觉得印象最深刻的说明一下）。

问题 55：踩过的坑（你在使用机器学习模型中踩过哪些坑？最后你是如何解决的）。

类型五：求职技巧类

问题 56：机器学习求职要准备哪些项？各项对应如何准备？

问题 57：机器学习相关的学习内容有哪些？学习路线应该怎么定？有什么推荐的学习资料？

问题 58：机器学习岗位求职的投递方式有哪些？什么时间投递最合适？投递目标应该怎样选择？

问题 59：机器学习岗位求职的简历最好写哪些内容？所做的项目应该如何描述？

问题 60：面试过程中自我介绍如何说比较合适？求职心态应该如何摆正？如果遇到压力该如何面对？面试过程中如何掌握主导权？怎样回答面试官最后的“你还有什么要问我的”问题？怎样面对最后的人力资源面试？

致谢

首先，我要感谢每一位为此书做出贡献的人和每一位读者，你们的认可与鼓励是我坚持写作的源动力，希望本书的内容可以给你们带来一份惊喜！

其次，我要感谢我的妻子彭璐。这些年我们一路从校园恋爱走到今天，过程真的十分不易。谢谢你一路对我的陪伴与付出，你就是我人生中最好的伯乐！

最后，我要感谢我的父母和兄弟。谢谢你们这么多年来对我的付出与支持，不管遇到什么困难，你们总是默默地站在我身后，给了我无穷的动力！

胡欢武

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/35521

目录