

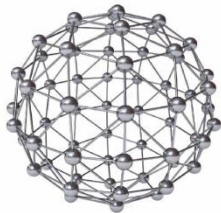


教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目  
数据科学与大数据技术系列规划教材

华为信息与网络  
技术学院指定教材

# 机器学习

赵卫东 董亮 编著



系统完整数据科学与大数据技术专业解决方案

名校名师打造大数据领域精品力作

强调基本概念和机器学习算法

兼顾机器学习经典内容，突出深度学习前沿

中国工信出版集团 人民邮电出版社  
POSTS & TELECOM PRESS

# 实践驱动的机器学习课程建设

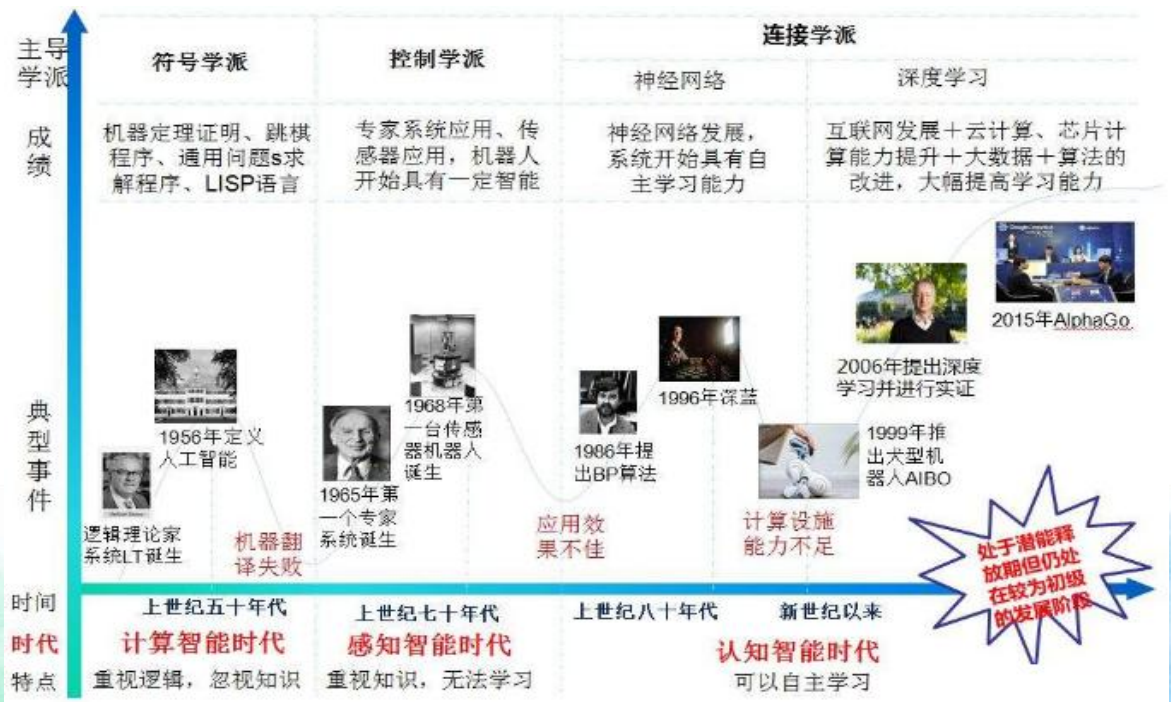
赵卫东 复旦大学

wdzhao@fudan.edu.cn



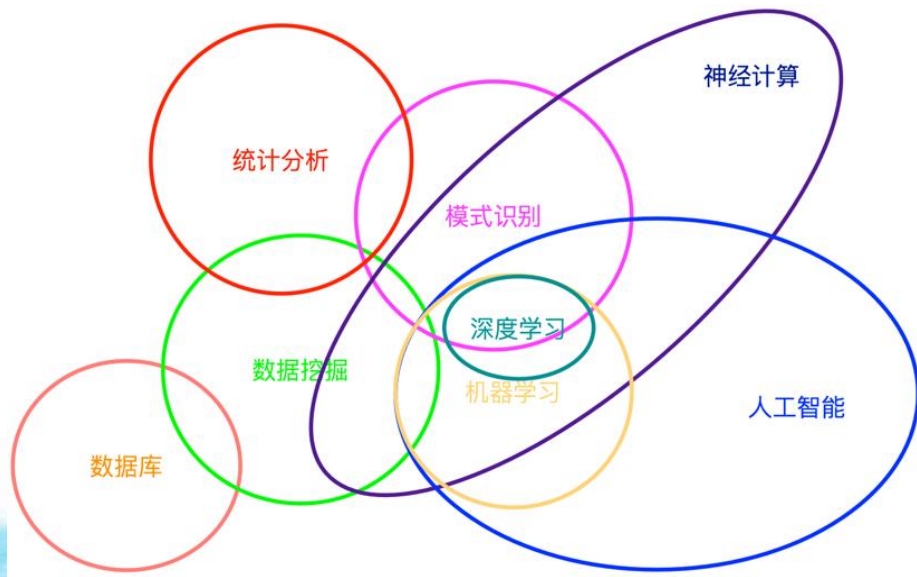
# 机器学习在人工智能发展中的作用

- 机器学习是人工（机器）智能的重要技术基础，也是智能科学的核心。
- 人工智能的4次浪潮基本与机器学习的方法有关。
- 人工智能的突破在于机器学习算法的进步。



# 机器学习的复杂性

- 机器学习主要的理论基础涉及到高等数学、线性代数、概率论与数理统计、数值逼近、最优化理论、数据库、计算复杂理论等，核心要素是数据、算法和模型。
- 机器学习的应用范围非常广泛，涉及各行各业。
- 机器学习技术的不成熟性（过高的期望）。



# 机器学习的演变（1）

- 机器学习的起源可以追溯到上世纪50年代早期人工智能的符号演算、逻辑推理、自动机模型、启发式搜索、模糊数学、专家系统、神经网络等，乃至最近的深度学习算法等。
- 强烈的概率逻辑和归纳逻辑，对大量、高质量数据的依赖过高。

机器学习阶段	年份	主要成果	代表人物
人工智能起源	1936	自动机模型理论	Alan Turing
	1943	MP模型	Warren McCulloch、Walter Pitts
	1951	符号演算	John von Neumann
	1950	逻辑主义	Claude Shannon
	1956	人工智能	John McCarthy、Marvin Minsky、Claude Shannon
人工智能初期	1958	LISP	John McCarthy
	1962	感知器收敛理论	Frank Roseblatt
	1972	通用问题求解(GPS)	Allen Newell、Herbert Simon
	1975	框架知识表示	Marvin Minsky
进化计算	1965	进化策略	Ingo Rechenberg
	1975	遗传算法	John Henry Holland
	1992	基因计算	John Koza
专家系统和知识工程	1965	模糊逻辑、模糊集	Lotfi Zadeh
	1969	DENDRA、MYCIN	Feigenbaum、Buchanan、Lederberg
	1979	ROSPECTOR	Duda

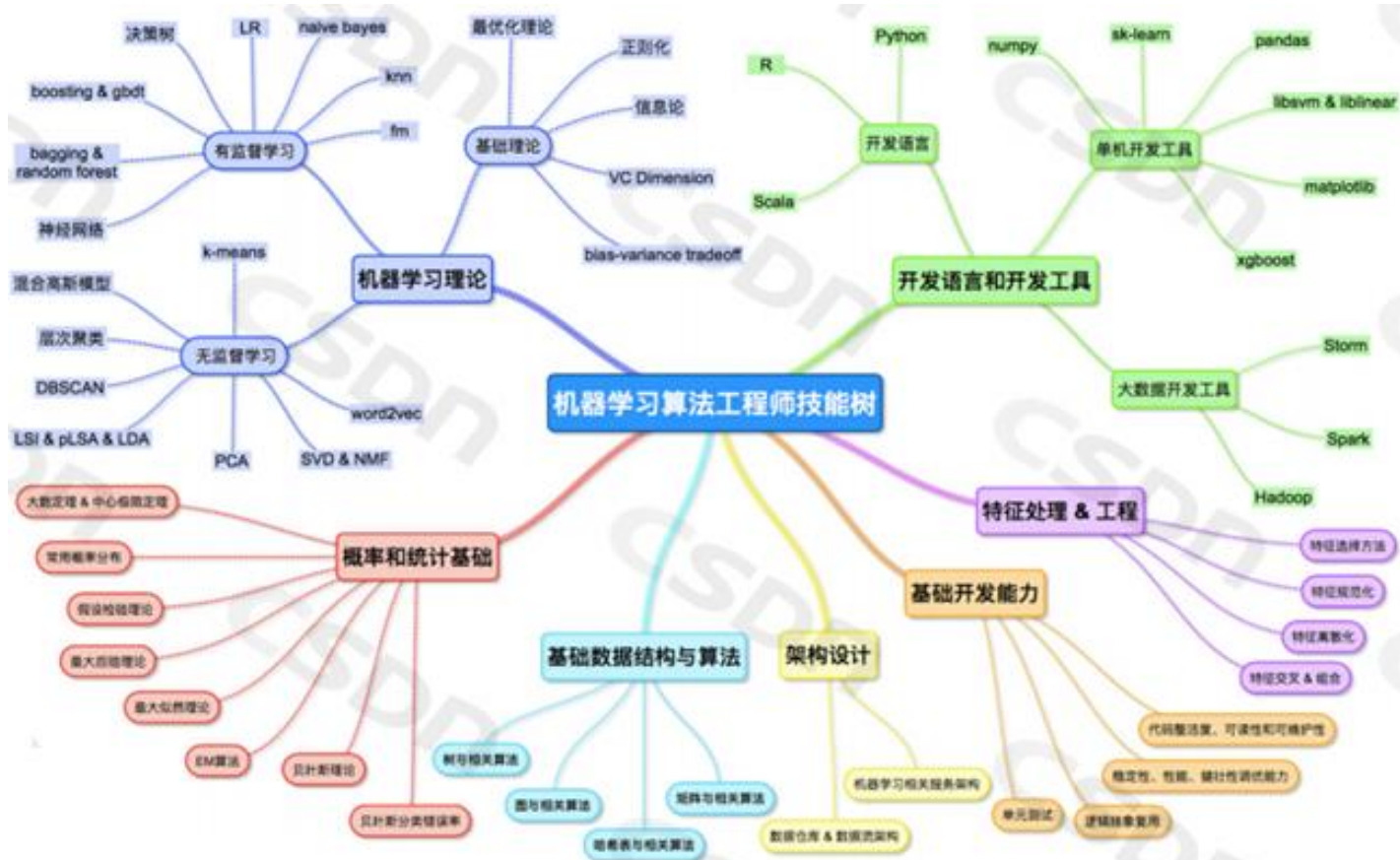
# 机器学习的演变（2）

神经网络	1982	Hopfield网络	Hopfield
	1982	自组织网络	Kohonen
	1986	BP算法	Rumelhart、McClelland
	1989	卷积神经网络	LeCun
	1998	LeNet	LeCun
	1997	循环神经网络RNN	Sepp Hochreiter、Jurgen Schmidhuber
分类算法	1986	决策树ID3算法	J. Ross Quinlan
	1988	Boosting算法	Freund、Michael Kearns
	1993	C4.5算法	J. Ross Quinlan
	1995	AdaBoost算法	Yoav Freund、Robert Schapire
	1995	支持向量机	Corinna Cortes、Vapnik
	2001	随机森林	Leo Breiman、Adele Cutler
深度学习	2006	深层神经网络训练方法	Geoffrey Hinton
	2012	谷歌大脑	Andrew Ng
	2014	生成对抗网络GAN	Ian Goodfellow

# 机器学习人才培养的难题

- 数理要求高
- 学习成本高（算法、编程语言、平台、应用领域知识等）
- 跨学科综合能力
- 实践机会少

- Python
- R
- TensorFlow
- Caffe
- 开源社区Github



# 机器学习平台多样 选择困境？

华为云 GPU 加速服务解决方案

行业	
基础应用	3D渲染, 视频处理, 机器学习, 图像处理, 基因测序, 金融计算
实例	G1 虚拟图形桌面 NVIDIA M60 + GRID G2 虚拟图形桌面 NVIDIA M60 + GRID
计算服务	EC2, ECS, EMR, EKS, Fargate
硬件	Fu



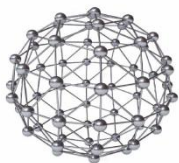
# 《机器学习》教材的特点

- 《机器学习》教材相对国内主流教材的特点：
- 强调基本概念和常用机器学习算法，简明易懂，剪系统性强
- 兼顾机器学习经典内容，突出深度学习前沿
- 重视案例和实践教学

教育部高等学校计算机类专业教学指导委员会、华为ICT产教合作单位  
数据科学与大数据技术专业系列规划教材 | 华为信息与网络技术学院指定教材

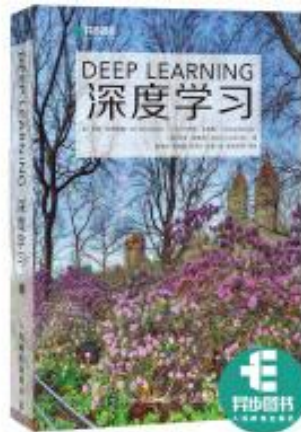
## 机器学习

赵其东 原亮 编著



系列完整 数据科学与大数据技术专业解决方案  
名校名师 7位大数据领域领军人物  
循序渐进 深入浅出机器学习算法  
案例驱动学习 经典内容，突出深度学习前沿

中国工信出版集团 | 人民邮电出版社  
POST & TELECOM PRESS





# 《机器学习》教材的内容 (1) 程

## 目录 CONTENTS

### 第 1 章 机器学习概述 ..... 1

- 1.1 机器学习简介 ..... 2
  - 1.1.1 机器学习简史 ..... 2
  - 1.1.2 机器学习主要流派 ..... 3
- 1.2 机器学习、人工智能和数据挖掘 ..... 5
  - 1.2.1 什么是人工智能 ..... 5
  - 1.2.2 什么是数据挖掘 ..... 6
  - 1.2.3 机器学习、人工智能与数据挖掘的关系 ..... 6
- 1.3 典型机器学习应用领域 ..... 7
- 1.4 机器学习算法 ..... 13
- 1.5 机器学习的一般流程 ..... 20

### 习题 ..... 21

### 第 2 章 机器学习基本方法 ..... 23

- 2.1 统计分析 ..... 24
  - 2.1.1 统计基础 ..... 24
  - 2.1.2 常见概率分布 ..... 29
  - 2.1.3 参数估计 ..... 30
  - 2.1.4 假设检验 ..... 32
  - 2.1.5 线性回归 ..... 33
  - 2.1.6 逻辑回归 ..... 36
  - 2.1.7 判别分析 ..... 37
  - 2.1.8 非线性模型 ..... 38
- 2.2 高维数据降维 ..... 39
  - 2.2.1 主成分分析 ..... 39
  - 2.2.2 奇异值分解 ..... 42
  - 2.2.3 线性判别分析 ..... 43
  - 2.2.4 局部线性嵌入 ..... 46
  - 2.2.5 拉普拉斯特征映射 ..... 47
- 2.3 特征工程 ..... 49

### 2.3.1 特征构建 ..... 49

### 2.3.2 特征选择 ..... 50

### 2.3.3 特征提取 ..... 51

### 2.4 模型训练 ..... 51

### 2.4.1 模型训练常见术语 ..... 51

### 2.4.2 训练数据收集 ..... 51

### 2.5 可视化分析 ..... 52

### 2.5.1 可视化分析的作用 ..... 53

### 2.5.2 可视化分析方法 ..... 53

### 2.5.3 可视化分析常用工具 ..... 54

### 2.5.4 常见的可视化图表 ..... 56

### 2.5.5 可视化分析面临的挑战 ..... 66

### 习题 ..... 66

### 第 3 章 决策树与分类算法 ..... 68

### 3.1 决策树算法 ..... 69

### 3.1.1 分支处理 ..... 70

### 3.1.2 连续属性离散化 ..... 76

### 3.1.3 过拟合问题 ..... 78

### 3.1.4 分类效果评价 ..... 83

### 3.2 集成学习 ..... 87

### 3.2.1 装袋法 ..... 87

### 3.2.2 提升法 ..... 88

### 3.2.3 GBDT ..... 90

### 3.2.4 随机森林 ..... 91

### 3.3 决策树应用 ..... 93

### 习题 ..... 96

### 第 4 章 聚类分析 ..... 97

### 4.1 聚类分析概念 ..... 98

### 4.1.1 聚类方法分类 ..... 98

### 4.1.2 良好聚类算法的特征 ..... 99

### 机器学习

### 4.2 聚类分析的质量 ..... 100

### 4.2.1 外部指标 ..... 100

### 4.2.2 内部指标 ..... 101

### 4.3 基于划分的聚类 ..... 103

### 4.3.1 $k$ -均值算法 ..... 103

### 4.3.2 $k$ -medoids 算法 ..... 108

### 4.3.3 $k$ -prototype 算法 ..... 108

### 4.4 基于密度的聚类 ..... 109

### 4.4.1 DBSCAN 算法 ..... 109

### 4.4.2 OPTICS 算法 ..... 111

### 4.4.3 DENCLUE 算法 ..... 112

### 4.5 基于层次的聚类 ..... 115

### 4.5.1 BIRCH 聚类 ..... 115

### 4.5.2 CURE 算法 ..... 118

### 4.6 基于网格的聚类 ..... 121

### 4.7 基于模型的聚类 ..... 121

### 4.7.1 概率模型聚类 ..... 121

### 4.7.2 模糊聚类 ..... 126

### 4.7.3 Kohonen 神经网络聚类 ..... 126

### 习题 ..... 132

### 第 5 章 文本分析 ..... 134

### 5.1 文本分析介绍 ..... 135

### 5.2 文本特征提取及表示 ..... 135

### 5.2.1 TF-IDF ..... 136

### 5.2.2 信息增益 ..... 136

### 5.2.3 互信息 ..... 137

### 5.2.4 卡方统计量 ..... 138

### 5.2.5 词嵌入 ..... 138

### 5.2.6 语言模型 ..... 139

### 5.2.7 向量空间模型 ..... 141

### 5.3 知识图谱 ..... 142

### 5.3.1 知识图谱相关概念 ..... 143

### 5.3.2 知识图谱的存储 ..... 144

### 5.3.3 知识图谱挖掘与计算 ..... 145

### 5.3.4 知识图谱的构建过程 ..... 146

### 5.4 词法分析 ..... 151

### 5.4.1 文本分词 ..... 151

### 5.4.2 命名实体识别 ..... 154

### 5.4.3 词义消歧 ..... 155

### 5.5 句法分析 ..... 155

### 5.6 语义分析 ..... 157

### 5.7 文本分析应用 ..... 158

### 5.7.1 文本分类 ..... 159

### 5.7.2 信息抽取 ..... 161

### 5.7.3 问答系统 ..... 162

### 5.7.4 情感分析 ..... 163

### 5.7.5 自动摘要 ..... 164

### 习题 ..... 165

### 第 6 章 神经网络 ..... 166

### 6.1 神经网络介绍 ..... 167

### 6.1.1 前馈神经网络 ..... 167

### 6.1.2 反馈神经网络 ..... 169

### 6.1.3 自组织神经网络 ..... 172

### 6.2 神经网络相关概念 ..... 173

### 6.2.1 激活函数 ..... 173

### 6.2.2 损失函数 ..... 176

### 6.2.3 学习率 ..... 178

### 6.2.4 过拟合 ..... 180

### 6.2.5 模型训练中的问题 ..... 181

### 6.2.6 神经网络效果评价 ..... 184

### 6.3 神经网络应用 ..... 184

### 习题 ..... 188

### 第 7 章 贝叶斯网络 ..... 189

### 7.1 贝叶斯理论概述 ..... 190

### 7.2 贝叶斯概率基础 ..... 190

### 7.2.1 概率论 ..... 190

### 7.2.2 贝叶斯概率 ..... 191

### 7.3 朴素贝叶斯分类模型 ..... 192

### 7.4 贝叶斯网络推理 ..... 195

### 7.5 贝叶斯网络的应用 ..... 200

### 7.5.1 中文分词 ..... 200

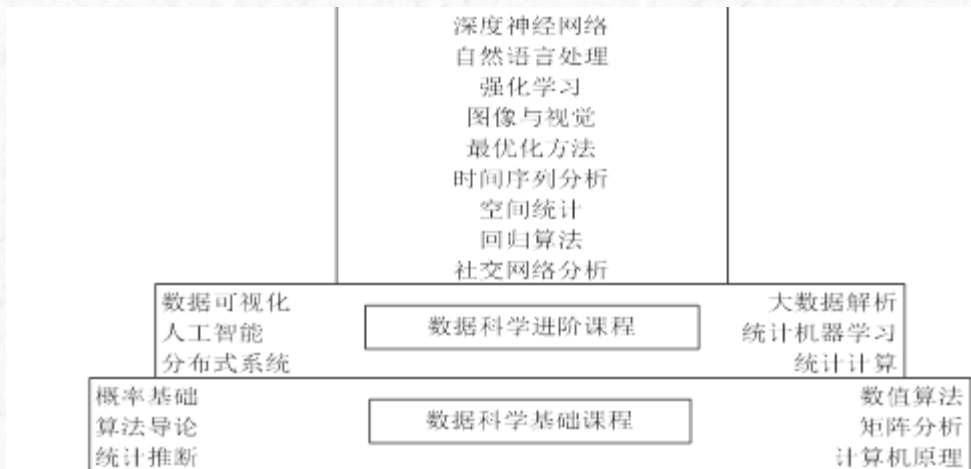
### 7.5.2 机器翻译 ..... 201

# 《机器学习》教材的内容(1) 程

7.5.3 故障诊断.....201	11.3 深度学习流行框架.....264	<b>第14章 实验</b> .....319	14.3 客户分群.....333
7.5.4 疾病诊断.....202	习题.....265		14.3.1 分析业务需求.....333
习题.....204	<b>第12章 高级深度学习</b> .....266	14.3.2 上传客户信息数据.....335	14.3.3 准备客户分群工作区.....336
<b>第8章 支持向量机</b> .....205	12.1 高级卷积神经网络.....267	14.1 华为 FusionInsight 产品平台介绍.....320	14.3.4 创建数据挖掘流程.....337
8.1 支持向量机模型.....206	12.1.1 目标检测与跟踪.....267	14.2 银行定期存款业务预测.....321	14.3.5 客户分群模型保存和应用.....344
8.1.1 核函数.....206	12.1.2 目标分割.....270	14.2.1 上传银行客户及存贷款数据.....322	<b>●参考文献</b> .....347
8.1.2 模型原理分析.....207	12.2 高级循环神经网络应用.....272	14.2.2 准备存款业务分析工作区.....322	
8.2 支持向量机应用.....210	12.2.1 Encoder-Decoder 模型.....272	14.2.3 创建数据挖掘流程.....323	
习题.....215	12.2.2 注意力模型.....273	14.2.4 定期存款业务模型保存和应用.....330	
<b>第9章 进化计算</b> .....216	12.2.3 LSTM 高级应用.....274		
9.1 遗传算法的基础.....217	12.3 无监督式深度学习.....275		
9.1.1 基因重组与基因突变.....217	12.3.1 深度信念网络.....275		
9.1.2 遗传算法实现技术.....218	12.3.2 生成对抗网络模型.....277		
9.1.3 遗传算法应用案例.....222	12.4 强化学习.....277		
9.2 蚁群算法.....223	12.5 迁移学习.....279		
9.3 蜂群算法.....225	12.6 对偶学习.....282		
习题.....227	习题.....283		
<b>第10章 分布式机器学习</b> .....229	<b>第13章 推荐系统</b> .....284		
10.1 分布式机器学习基础.....230	13.1 推荐系统基础.....285		
10.1.1 参数服务器.....230	13.1.1 推荐系统的应用场景.....285		
10.1.2 分布式并行计算类型.....231	13.1.2 相似度计算.....286		
10.2 分布式机器学习框架.....232	13.2 推荐系统通用模型.....288		
10.3 并行决策树.....238	13.2.1 推荐系统结构.....288		
10.4 并行 K-均值算法.....238	13.2.2 基于人口统计学的推荐.....288		
习题.....240	13.2.3 基于内容的推荐.....289		
<b>第11章 深度学习</b> .....242	13.2.4 基于协同过滤的推荐算法.....290		
11.1 卷积神经网络.....243	13.2.5 基于图的模型.....292		
11.1.1 卷积神经网络简介.....243	13.2.6 基于关联规则的推荐.....293		
11.1.2 卷积神经网络的结构.....244	13.2.7 基于知识的推荐.....299		
11.1.3 常见卷积神经网络.....246	13.2.8 基于标签的推荐.....300		
11.2 循环神经网络.....254	13.3 推荐系统评测.....301		
11.2.1 RNN 基本原理.....254	13.3.1 评测方法.....301		
11.2.2 长短期记忆网络.....260	13.3.2 评测指标.....302		
11.2.3 门限循环单元.....263	13.4 推荐系统常见问题.....306		
	13.5 推荐系统实例.....309		
	习题.....318		

# 《机器学习》预修课程

- 扎实的高等数学、统计学、线性代数等数学基础
- 掌握Python编程语言
- 分布式计算理论（建议），针对大数据机器学习
- 相关应用领域的基本了解



# 《机器学习》实践教学

## 资讯类推荐系统实例

中国移动 10:50 AM 100% 中国移动 11:05 AM 100% 中国移动 11:07 AM 100%

### 股权投资 (成熟期)

苏州工业园区元禾控股投资管理有限公司 (下称“元禾资本”) 是苏州元禾控股股份有限公司 (下称“元禾控股”) 的控股企业。元禾控股专业化并集创业平台孵化运营与管理、承担基金与资源对接和产业发展的功能。成为元禾控股投资价值最大化和增值实现的运作。

### 新一轮科技革命和产业变革将由工业互联网引发?

Ascp作为中国软交会的一个重要行业会议, 6月15日举行的2017工业互联网(大陆)峰会吸引了众多“大咖”, 围绕构建“一带一路”工业互联网生态体系、大数据驱动工业智能化转型升级、工业互联网与智能制造等主题展开了热烈的观点交锋。工业互联网正在引发新一轮科技。

### 中国四大超级平台共同助推中国梦!!!

### 云布——纺织企业管理系统 & 交易平台

云布, 是针对纺织行业开发的一套能满足企业资产管理和内部业务所需的多项工作。云布应用, 覆盖了生产、供应、成本、质量、生产、在库、物流等纺织企业最关注的业务环节, 同时还涵盖了业务流程相关的产品、订单、报价、逻辑存、客户、供应商等系统管理功能, 旨在为...

### 手术机器人走进医疗领域

随着“智慧医疗”建设不断深入, 机器人等前沿科技与医疗领域结合将更加紧密。手术机器人有望走进医院, 成为国内医疗领域, 目前, 由中国科学家研发的世界首台自主式微创手术机器人成功进行了第一次手术, 宣告了这一关键科技装备的隆重诞生。

### 机器人检测认证产业将迎来爆发

中国的高质量的机器人产业发展, 机器人是制造业转型升级的利器, 其研发、制造、应用将是未来一个国家和国际制造业水平的重要标志。机器人是抢占智能社会先发优势的战略性新兴产业, 机器人科技创新和产业发展将为全球经济注入新动力。

AI产业成智能产业发展核心 或将引发行业

### 云布纺织企业管理平台好不好

传统纺织管理软件 VS 云布

传统纺织管理软件	云布
“一对一”人工服务	互联网+自助服务
数据孤岛	数据互通
信息滞后	实时数据
操作复杂	简单易用
部署困难	随时随地

### 纺织行业的先进生产力

### 这篇文章, 你觉得怎么样?

### CRM客户关系管理

Salesforce CRM是利用现代信息技术手段, 在企业与客户之间建立一种双向的、互动的交易管理系统。它结合企业现实管理, 综合利用客户信息, 进行数据挖掘与分析, 从而提升公司的业务效率, 主要功能: 客户管理; 主要实现对客户信息的维护、沟通记录的查看、客户等级管理。

### 企业管理软件系统ERP如何帮助提高企业管理水平

一、什么是管理软件系统? 一套优秀的管理系统重点不在于软件技术, 而在于管理理念。一套优秀的管理理念重点不在于理念, 而在于实践, 更在于理念与实践的结合; 将先进的管理理念与有形的管理行为相结合, 这就是企业管理模式; 二、中小企业中存在的主要问题则到中小企业管理...

### 【观麦】对不起, 食材配送ERP领域没有“一招鲜 吃遍天”!

### 公交集团生产运营部组织召开ERP系统建设交流会

6月7日上午, 公交集团生产运营部组织召开ERP系统建设交流会, 青岛海信网络科技公司相关人员参加了会议, 会议首先由海信网络科技公司负责人对ERP系统进行总体介绍, 同时对运营、机务、安全、人员、财务、移动办公等11个分子系统功能模块以及业务操作流程进行了展示...

### 零售行业 | 通达零售ERP

零售行业管理软件方案功能: 通过综合计划、预测、分配和补货来有效地加强您的供应链; 使用多功能电子商务与供应商合作, 与顾客密切交流; 使用强大的数据挖掘工具为报表和跟踪情况提供决策信息, 为企业从头部至整体运营提供完整的视图; 通过多渠道销售和量身定制的渠道营销...

### ERP实施中13个最常见的错误

企业资源计划(ERP)系统需要企业投入大量的资金、资源和时间, 其每一个环节的影响都上千万至几亿美元, 那将需要耗费数月之久, 浪费一个成功的机会, 那么能够导致失败的公司工作流程, 构成成本, 但是一个轻微与部署不合理的ERP产品可能会导致成本上升, 生产效率低下, 为了看...

【高端访谈】GSK中国IT的创新之路

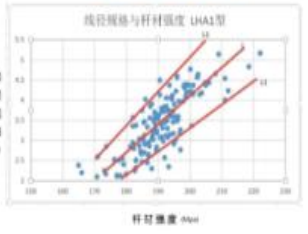
项目沉浸式教学的实施难点:  
项目从哪里来? 企业内部创新科创项目以及与合作横向课题。

## 我们近2年与企业的典型合作项目

京东公司: 发现好货单品素材写作  
天呈医流: 客服机器人  
江苏中天: 电缆质量检测数据分析  
杨浦政府采购中心: 串标检测系统



### 发现好货



这是一个需要多年积累沉淀的过程!

# 特别致谢

- 基于华为MLS产品平台的实践过程，通过分析银行定期存款业务和对客户进行分群两个案例说明算法应用过程。其中银行存款业务分析中运用了逻辑回归、随机森林等算法，在客户分群的案例中主要基于聚类算法对客户进行分类。

## 华为FusionInsight产品平台

