



数据挖掘概论

17/10/14

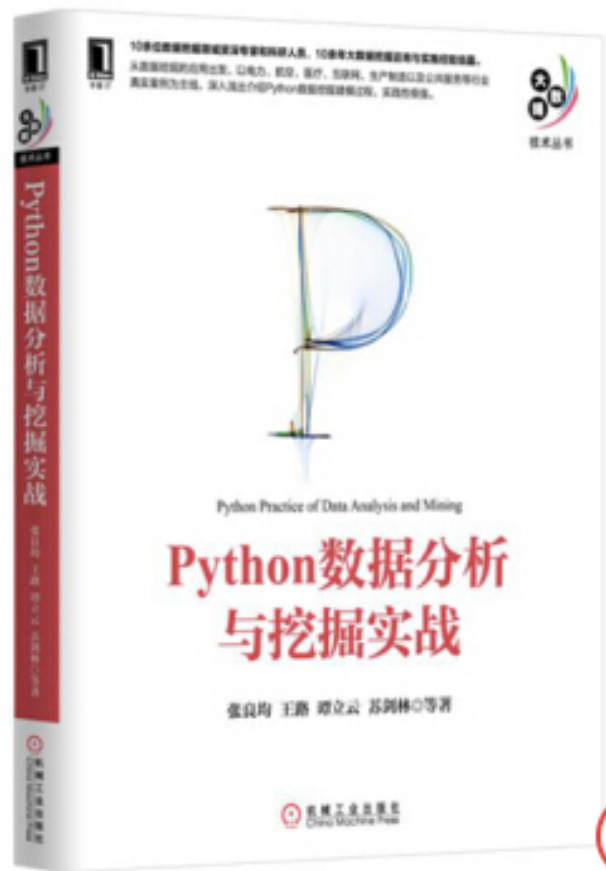


目录



数据挖掘企业项目介绍

Python数据分析与挖掘实战



数据挖掘企业项目介绍

Python数据分析与挖掘实战

- 第6章 电力窃漏电用户自动识别
- 第7章 航空公司客户价值分析
- 第8章 中医证型的关联规则挖掘
- 第9章 基于水色图像的水质评价
- 第10章 家用电器用户行为分析及事件识别
- 第11章 应用系统负载分析与容量预测
- 第12章 电子商务网站用户行为分析及服务推荐
- 第13章 财政收入影响因素分析及预测模型
- 第14章 基于基站定位数据的商圈分析
- 第15章 电商产品评论数据情感分析



数据挖掘企业项目介绍

第6章 电力窃漏电用户自动识别

背景

- 全国每年因窃电造成的损失在200亿元左右，被查获的不足30%
- 深圳龙岗工业区一家只有两条生产线的小塑料包装厂，一年窃电折价就30 - 40万元
- 某市06年因窃电损失达4亿元
- 传统打击手段：突击检查

目标

- 如何通过监测数据自动识别偷漏电行为？



数据挖掘企业项目介绍

第7章 航空公司客户价值分析

背景

- 竞争激烈：同行业、高铁

目标

- 客户价值分析，精准营销



数据挖掘企业项目介绍

第12章 电子商务网站用户行为分析及服务推荐

背景

- 某法律网站是一家大型的法律资讯信息网站，它一直致力于为用户提供丰富的法律资讯信息与专业法律咨询服务，并为律师与律师事务所提供卓有成效的互联网整合营销解决方案。
- 大量的访问用户，每天上千万次的点击量，为其带来发展也带来瓶颈。
- 如何留住需要帮助用户，快速找到其感兴趣的页面？并进一步为其推荐律师？

目标

- 客户行为分析
- 用户精准画像
- 智能推荐



数据挖掘企业项目介绍

第15章 电商产品评论数据情感分析

背景

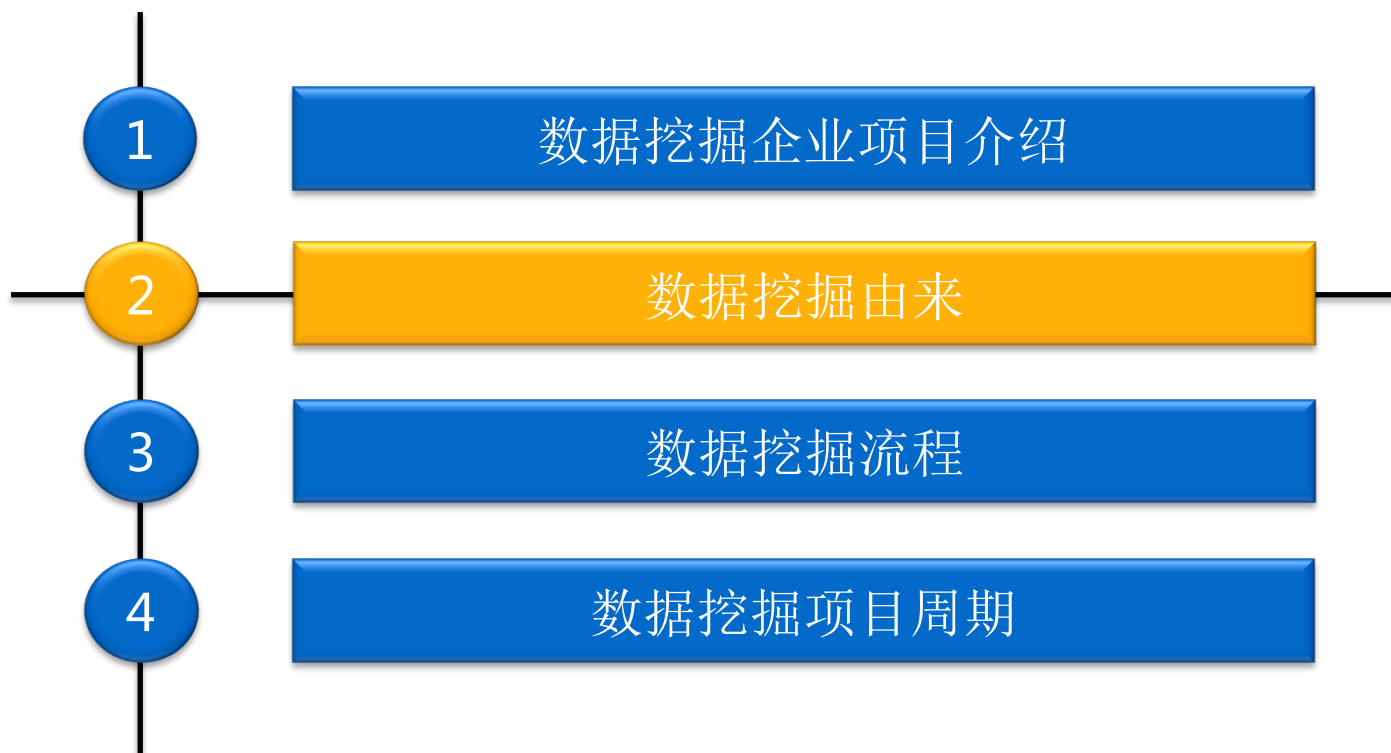
- 网购盛行，竞争激烈
- 消费者对商家反馈更多更直接
- 反馈多以文字评论为主，难以有效利用

目标

- 电商用户评论数据情感分析
- 商品优缺点挖掘



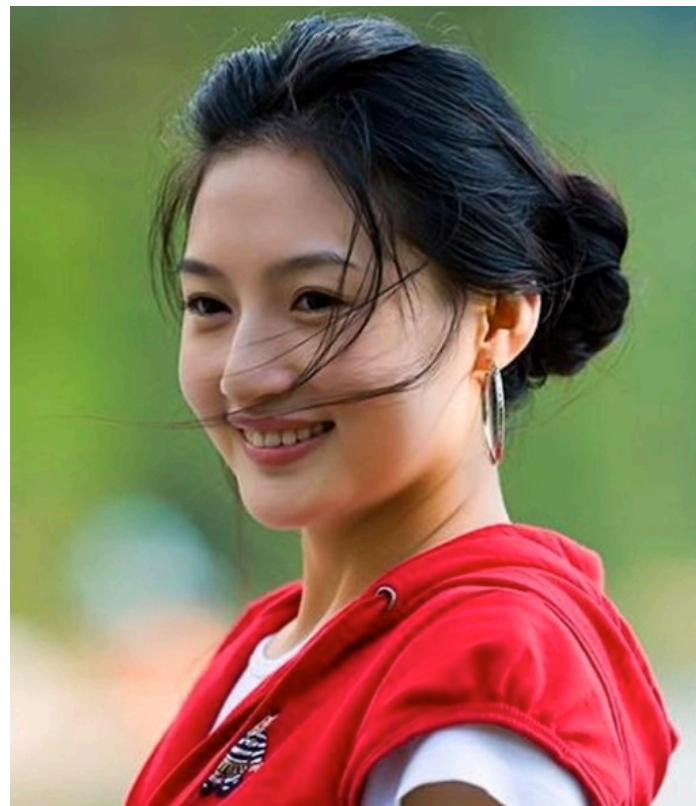
目录



数据挖掘由来

引例

男、女？



数据挖掘由来

引例

请分别讨论下列各组数据的内部关系，并填空。

x_1	3	1	7	2	4
y_1	4.5	2.5	8.5	3.5	?

$$y_1 = x_1 + 1.5$$

x_2	3	6	8	1	2
y_2	10.5	37.5	65.5	2.5	?

$$y_2 = x_2^2 + 1.5$$

x_1	2.0	6.0	5.0	1.0	4.0
x_2	7.0	9.0	3.0	2.0	5.0
y	52.8	96.7	21.2	6.0	?

$$y = x_1^{3/2} + x_2^2 + 1$$



数据挖掘由来

引例

商品推荐：



推荐系统：技术、评估及高效算法

京东悦读节！全场自营图书音像每满150减50！共15波200减100优惠券限时抢！叠券可享300减200，相当于折上3.3折！[立即查看活动](#)

[美] 弗朗西斯科·里奇 (Francesco Ricci), [美] 利奥·罗卡奇 (Lior Rokach), [美] 布拉哈·夏皮拉 (Bracha Shapira), [美] 保罗 B. 坎特 (Paul Kantor) 编；李艳民, 胡聪, 吴宾, 王雪丽 丁彬利 译

京 东 价：**¥114.70** [8.3折] [定价：¥139.00] (降价通知)

促销信息：以下促销可在购物车任选其一

满减 每满150.00元，可减50.00元现金 [详情 >>](#)

累计评价
300+

共3项促销

配 送 至：

广东广州市萝岗区

 有货，支持 99元免基础运费 | 货到付款

服 务：由 京东 发货，并提供售后服务。23:00前下单,预计明天(04月21日)送达

1

+

加入购物车

一键购

温馨提示：1. 支持7天无理由退货

41个卖家在售 **¥89.60** 起

企业批量购书

分享 关注商品

举报

数据挖掘由来

引例

商品推荐：

人气单品

深入理解机器学习：从原理到算法



用户网络行为画像 大数据中的用户网络行为画

¥49.40

集体智慧编程



集体智慧编程

¥75.10

计算广告：互联网商业变现的市场与技术



计算广告：互联网商业变现的市场与技术

¥57.80

深入理解机器学习：从原理到算法



深入理解机器学习：从原理到算法

¥65.20

算法导论 (原书第3版) /计算机科学丛书



算法导论 (原书第3版) /计算机科学丛书

¥104.70

数据挖掘 概念与技术 (原书第3版)



数据挖掘 概念与技术 (原书第3版)

¥64.60

您可能还需要

机器学习



机器学习

¥77.00

加入购物车

计算机程序设计艺术 卷2 半数值算法 (第3版)




计算机程序设计艺术 卷2 半数值算法 (第3版)

¥156.40

加入购物车

TensorFlow实战



TensorFlow实战

¥61.80

加入购物车

终极算法：机器学习和人工智能如何重塑世界



终极算法：机器学习和人工智能如何重塑世界

¥51.00

加入购物车

数据科学与大数据分析 数据的发现 分析 可视化与表示



数据科学与大数据分析 数据的发现 分析 可视化与表示

¥65.60

加入购物车

第一行代码 Android 第2版



第一行代码 Android 第2版

¥75.10

加入购物车

Spring实战 (第4版)



Spring实战 (第4版)

¥84.60

加入购物车

深入浅出MyBatis技术原理与实战



深入浅出MyBatis技术原理与实战

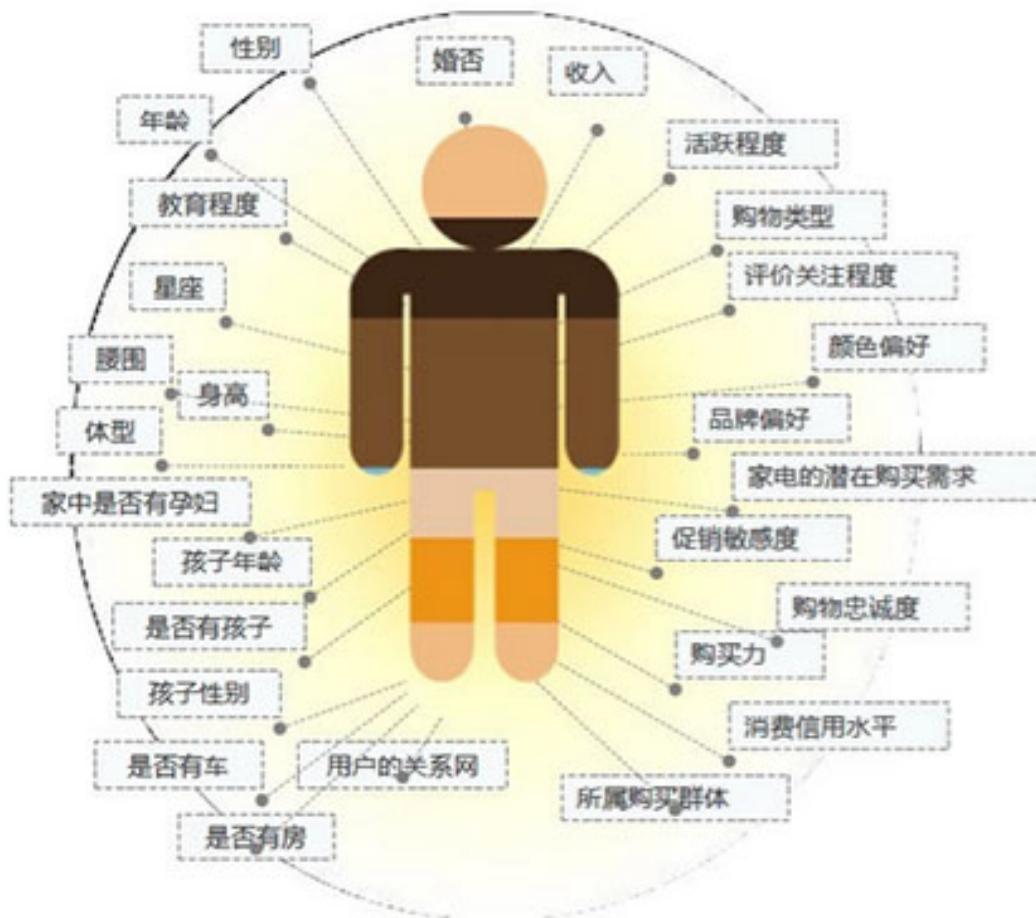
¥62.70

加入购物车

数据挖掘由来

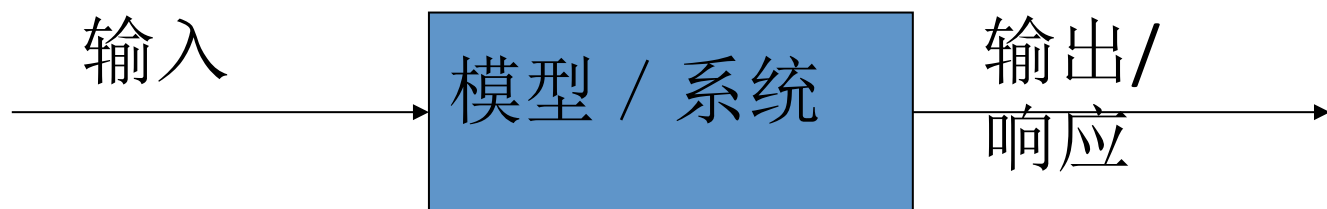
引例

电商网站用户画像



数据挖掘由来

引例



输入：发型、喉结、胡须，已知数据对，历史购物记录

输出 / 响应：男 / 女，对应 y 值，推荐结果



数据挖掘由来

人工智能简介

人工智能 [维基百科]

- 人工智能 (Artificial Intelligence) 亦称机器智能，是指由人工制造出来的系统所表现出来的智能。
- MIT的约翰·麦卡锡在1956年的达特茅斯会议上提出的：人工智能就是要让机器的行为看起来就像是人所表现出的智能行为一样。
- “像人一样思考”、“像人一样行动”、“理性地思考”和“理性地行动”
- <https://zh.wikipedia.org/zh-cn/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD#.E5.BC.B7.E4.BA.BA.E5.B7.A5.E6.99.BA.E8.83.BD>



数据挖掘由来

人工智能简介

人工智能：强弱之分 [维基百科]

- **强人工智能**观点认为有可能制造出真正能推理（ Reasoning ）和解决问题（ 解决问题 ）的智能机器，并且，这样的机器能将被认为是有知觉的，有自我意识的。
- **弱人工智能**观点认为不可能制造出能真正地推理和解决问题的智能机器，这些机器只不过看起来像是智能的，但是并不真正拥有智能，也不会有自主意识。
- 机器是否能思考，与潜水艇是否能游泳的问题很像 [人工智能时代]。
- 帮你寻找约会对象的网站和帮你割草的机器人，它们的做法是否和你一样并不重要，却会以你永远都无法到达的速度、准确度以及更低的成本来完成这些工作。



数据挖掘由来

人工智能简介

图灵测试 [百度百科]

- 测试者在与被测试者（一个人和一台机器）隔开
- 通过一些装置（如键盘）向被测试者随意提问
- 30%
- 人类智能



数据挖掘由来

人工智能简介

人工智能的发展历史 [维基百科]

- Buchanan, Bruce G. A (Very) Brief History of Artificial Intelligence (PDF). AI Magazine. 2005: 53–60 [30 August 2007]
- 1950年 阿兰·图灵出版《计算机与智能》
- 1956年 约翰·麦卡锡在美国达特茅斯电脑大会上“创造”“人工智能”一词。 1956年 美国卡内基·梅隆大学展示世界上第一个人工智能软件的工作。
- 1958年 约翰·麦卡锡在麻省理工学院发明Lisp语言——一种A . I . 语言。
- 1964年 麻省理工学院的丹尼·巴洛向世人展示，电脑能掌握足够的自然语言从而解决了开发计算机代数词汇程序的难题。
- 1965年 约瑟夫·魏岑堡建造了ELIZA——一种互动程序，它能以英语与人就任意话题展开对话



数据挖掘由来

人工智能简介

- 1969年 斯坦福大学研制出Shakey：一种集运动、理解和解决问题能力于一身的机器人。
- 1979年 第一台电脑控制的自动行走器“斯坦福车”诞生。
- 1983年 世界第一家批量生产统一规格电脑的公司“思考机器”诞生。
- 1985年 哈罗德·科岑编写的绘图软件Aaron在A . I . 大会亮相。
- 1997年 IBM（国际商用机械公司）制造的电脑“深蓝”击败了国际象棋冠。
- 2011年2月17日，IBM、德克萨斯大学联合研制的超级电脑“沃森” (Watson)在《危险边缘》夺冠。
- 2016年，阿法狗4:1战胜李世石。
- 2017年，阿法狗3:0战胜柯洁。



数据挖掘由来

人工智能简介

人工智能研究子领域 [维基百科]

- 演绎、推理和解决问题
- 知识表示法：知识表示和常识知识库
- 学习：机器学习
- 自然语言处理
- 运动和控制
- 知觉：机器感知，计算机视觉和语音识别
- 社交：情感计算
- 创造力：计算机创造力（人工直觉和人工想像）
- 多元智能：强人工智能（结合以上所有的技能并且超越大部分人类的能力）
- 伦理管理：机器超越人类，机器反客为主
- 经济冲击：即将被机器人取代的职业



数据挖掘由来

人工智能简介

人工智能应用领域 [维基百科]

- 机器视觉
- 指纹识别
- 人脸识别
- 视网膜识别
- 虹膜识别
- 掌纹识别
- 专家系统
- 自动规划等



数据挖掘由来

人工智能简介

当今人工智能主要两个方向〔人工智能时代〕

1、合成智能（又称机器学习、神经网络、大数据、认知系统或者遗传算法）

- 从经验中学习
- 模式识别、语音识别

2、人造劳动者（又称“机器人”）

- 传感器与执行器相结合
- 手机GPS导航（收集交通信息、发出路线）



数据挖掘由来

数据挖掘的起源 [数据挖掘导论]

数据挖掘是在大型数据存贮库中，自动发现有用信息的过程。数据挖掘技术用来探查大型数据库，发现先前未知的有用模式。



数据挖掘由来

数据挖掘的起源 [数据挖掘导论]

传统数据分析技术解决不了的新问题 - 新数据集带来的问题：

- 可伸缩性 - 量大、搜索空间呈指数集增长
- 高维性
- 异种数据和复杂数据 - web网页、DNA序列、地理气象数据
- 数据的所有权与分布 - 数据分布式存放
- 份传统分析



数据挖掘由来

数据挖掘的起源 [数据挖掘导论]

多学科汇集

- 统计学 - 抽样、估计和假设检验
- 人工智能、模式识别和机器学习
- 其他领域 - 最优化、进化计算、信息论、信号处理、可视化和信息检索



数据挖掘由来

数据挖掘与机器学习算法

数据挖掘算法 [数据挖掘十大算法]

- C4.5、K-means、SVM、Apriori、EM、PageRank、AdaBoost、kNN、Naive Bayes、CART

机器学习算法 [机器学习]

- 线性回归、决策树、神经网络、支持向量机、贝叶斯分类器、集成学习、聚类、降维与度量学习、特征选择与稀疏学习、计算学习理论、半监督学习、概率图模型、规则学习、强化学习



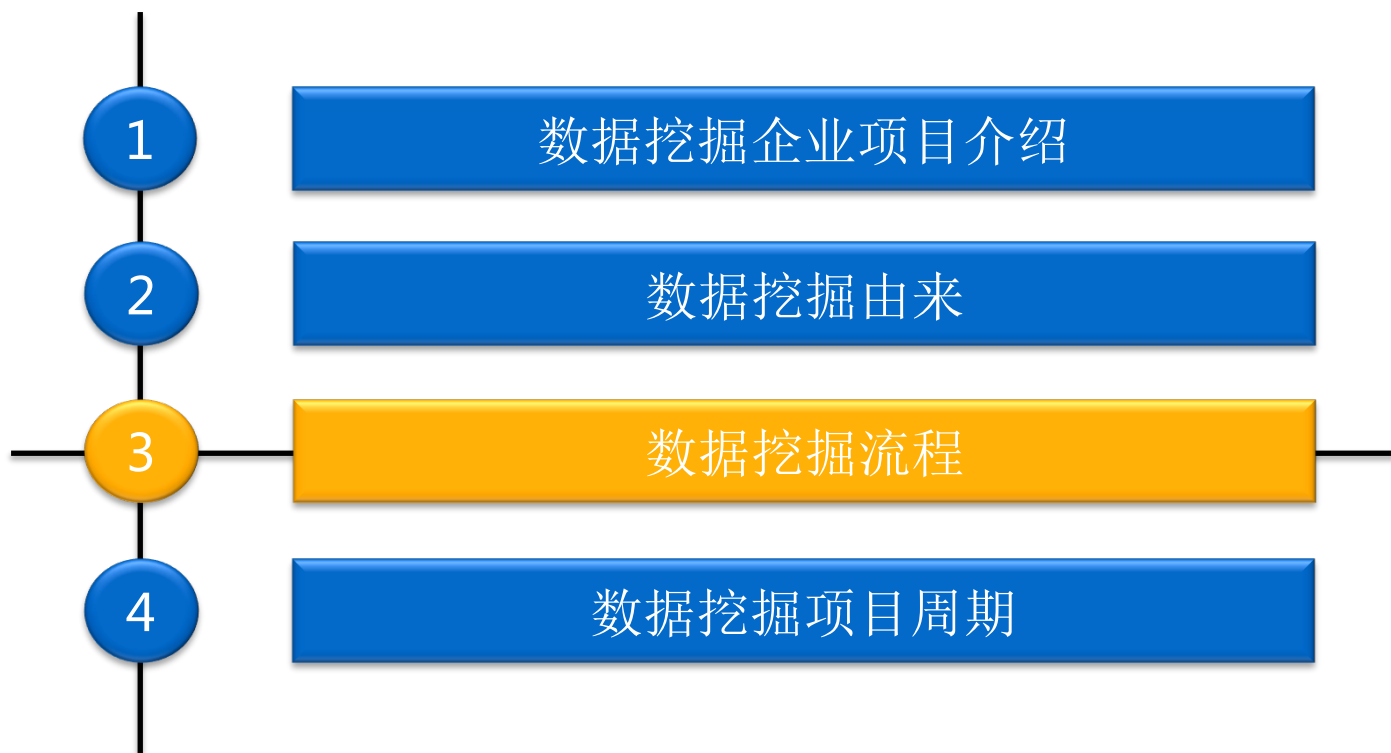
数据挖掘由来

数据挖掘常见任务 [R语言与数据挖掘]

- 分类与回归
- 聚类分析
- 关联规则
- 偏差检测
- 智能推荐
- 自然语言处理
- 时间序列



目录



数据挖掘流程

数据挖掘建模过程：CRISP-DM方法论 - 跨行业数据挖掘标准流程

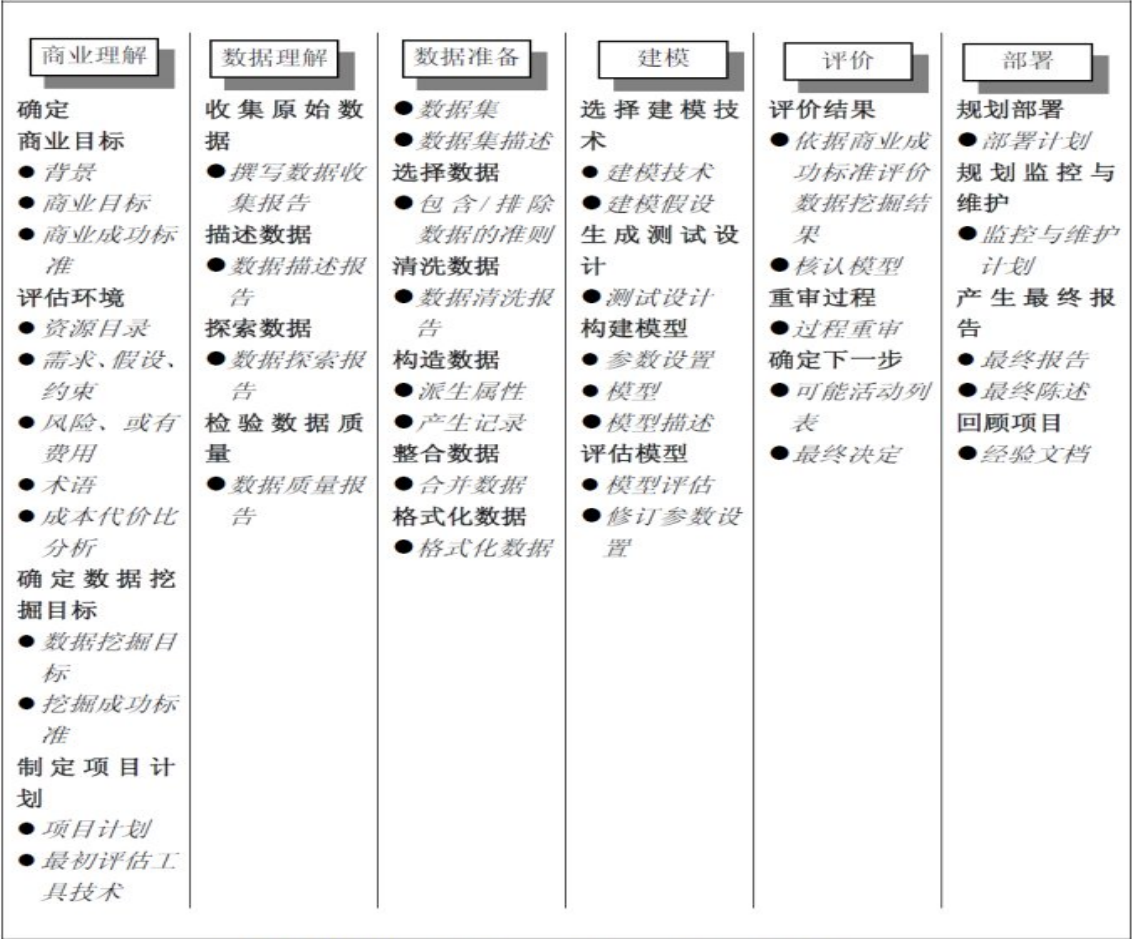


图 3 CRISP-DM 参考模型的一般任务（粗体）和其输出（斜体）



数据挖掘流程

第1步：定义挖掘目标

要想充分发挥数据挖掘的价值，必须要对目标有一个清晰明确的定义，即决定到底想干什么。

- 本次的挖掘目标是什么？
- 系统完成后能达到什么样的效果？
- 可明确评估的实施效果



数据挖掘流程

第2步：数据取样

实现数据挖掘目标需要哪些数据？抽取数据的标准：

- 相关性：与数据挖掘任务相关
- 代表性：样本数据能合理代表目标数据（例：男女分配符合目标样本）
- 可靠性：数据能反应样本的真实状况（相反：假数据）
- 有效性：满足挖掘任务的需要（例：时间长度足够长）



第3步：数据探索

- 检查数据的异常值或意外状态（“脏”数据）
- 探索数据里明显的规律和趋势
 - “垃圾进，垃圾出”（Junk in, junk out）
 - 目的：是为了保证样本数据的质量，从而为保证模型质量打下基础
 - 主要包括：数据质量分析、数据特征分析

数据挖掘流程

第3步：数据探索

- 数据质量分析
- 缺失值分析
- 异常值分析
- 重复数据



数据挖掘流程

缺失值分析

缺失值的影响：

- 数据挖掘建模将丢失一定量的信息
- 可能会导致数据规律难以被发现
- 可能导致模型效果低

处理方式：

- 通过简单的统计得出有缺失值的属性个数，以及单个属性内值的缺失数和缺失率
- 根据具体情况，对缺失值进行删除、插补或不处理



数据挖掘流程

日期	用户A用电量	用户B用电量	用户C用电量
2014/9/1	235.8333	350.8333	478.3231
2014/9/2	236.2708	351.2708	515.4564
2014/9/3	238.0521	353.0521	517.0909
2014/9/4	235.9063	350.9063	514.89
2014/9/5	236.7604	351.7604	
2014/9/8		352.4167	486.0912
2014/9/9	237.4167	353.6563	516.233
2014/9/10	238.6563		
2014/9/11	237.6042	352.6042	435.3508
2014/9/12	238.0313	353.0313	487.675
2014/9/15	235.0729	350.0729	
2014/9/16	235.5313	350.5313	660.2347
2014/9/17		349.4688	621.2346
2014/9/18	234.4688		611.3408
2014/9/19	235.5	350.5	643.0863
2014/9/22	235.6354	350.6354	642.3482



缺失值处理 - 插补法

插补方法	方法描述
均值/中位数/众数插补	根据属性值的类型，用该属性取值的平均数/中位数/众数进行插补。
使用固定值	将缺失的属性值用一个常量替换。如广州一个工厂普通外来务工人员的“基本工资”属性的空缺值可以用2016年广州市普通外来务工人员工资标准1895元/月该方法就是使用固定值。
最近临插补	在记录中找到与缺失样本最接近的样本的该属性值插补
回归方法	对带有缺失值的变量，根据已有数据和与其有关的其他变量（因变量）的数据建立拟合模型来预测缺失的属性值。
插值法（自建函数）	插值法是利用已知点建立合适的插值函数，未知值由对应点求出的函数值近似代替。



数据挖掘流程

异常值分析

检验数据是否含有录入错误以及含有不合常理的数据，其中包括：

- 离群点
- 不符合数据集背景的值
- 不一致的值

异常值分析方法：

- 简单统计量分析（最大值、最小值）
- 3σ 原则
- 箱型图分析离群点



异常值分析

客户id	姓名	性别	联系方式	消费金额
01	name1	M	151888888812	100
02	name2	F	151888888813	213
03	name3	M	179888888814	401
04	name4	F	1518888	98
05	name5	M	152888888813	113
06		F	151888888814	67
07	name7	ABC	134888888812	109
08	name8	M	135888888813	134
09	name9	M	189888888814	10089
10	name10	F	178888888812	93



数据挖掘流程

数据特征分析

- 分布分析（高斯分布、泊松分布、二项分布...）
- 对比分析（横向、纵向...）
- 统计量分析（均值、中位数、方差、标准差...）
- 周期性分析（周期、频率、幅度...）
- 贡献度分析（帕累托原则、主成分分析）
- 相关性分析（相关系数、相似度...）

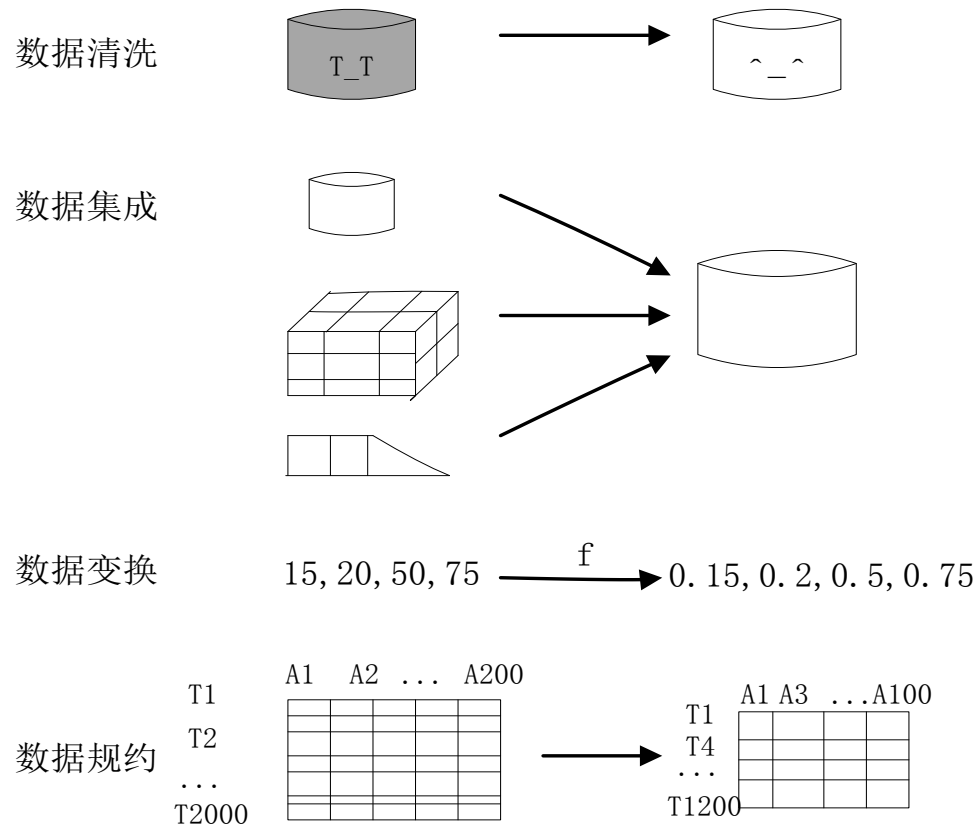


数据挖掘流程

第4步：数据预处理

占到了整个过程的60%，主要包括：

- 数据清洗（处理“脏”数据）
- 数据集成（数据筛选、属性选择.....）
- 数据变换（变量转换、标准化.....）
- 数据规约（属性降维）



数据挖掘流程

数据变换

对数据规范化，转换成“适当的”格式，以适用于挖掘任务及算法的需要：

- 简单函数变换：平方、开方、对数、差分
- 归一化（标准化）：消除量纲的影响
- 连续属性离散化
- 属性构造
- 小波变换



数据挖掘流程

归一化

- 最小-最大规范化：使结果值映射到[0,1]之间
- 零-均值规范化：也叫标准差标准化，平均数为0，标准差为1。
- 小数定标规范化：通过移动属性值的小数位数，将属性值映射到[-1, 1]之间，移动的小数位数取决于属性值绝对值的最大值

$$x^* = \frac{x - \min}{\max - \min}$$

$$x^* = \frac{x - \bar{x}}{\sigma}$$

$$x^* = \frac{x}{10^k}$$



数据挖掘流程

第5步：挖掘建模

常用模型：

- 决策树：ID3、C4.5、C5.0、CART
- 回归分析：线性回归、非线性回归
- 支持向量机、神经网络
- 贝叶斯网络
- K-means
- IBCF、UBCF
- Apriori
-



第6步：模型评价

模型，哪家强？

- 在评价过程中，不断在模型中调试参数，发现不同的评价结果
- 使用多种评价方法进行评价
- 结合业务，对模型评价结果进行解读



数据挖掘流程

第6步：模型评价

混淆矩阵

- 准确率： $\text{precision} = \text{TP}/(\text{TP}+\text{FP})$
- 召回率（灵敏度）： $\text{recall} = \text{TP}/(\text{TP}+\text{FN})$
- 特异性： $\text{TN}/(\text{TN}+\text{FP})$
- F系数

		预测值	
		1	0
实际值	1	47	3
	0	1	49

		预测值	
		1	0
实际值	1	TP	FN
	0	FP	TN



数据挖掘流程

第6步：模型评价

模型评价可视化

- ROC曲线 (Receiver Operating Characteristic , 受试者工作特征)
- 描述灵敏度 (真阳性) 和特异性 (真阴性)
- 直虚线表示没有预测价值的分类器
- 曲线越倾向左上方 , 说明模型能越好地识别阳性值
- 可用ROC曲线下面积 (Area Under the ROC, AUC) 作为评价指标
- 风险图 (Risk Chart , 又称累计增益图)

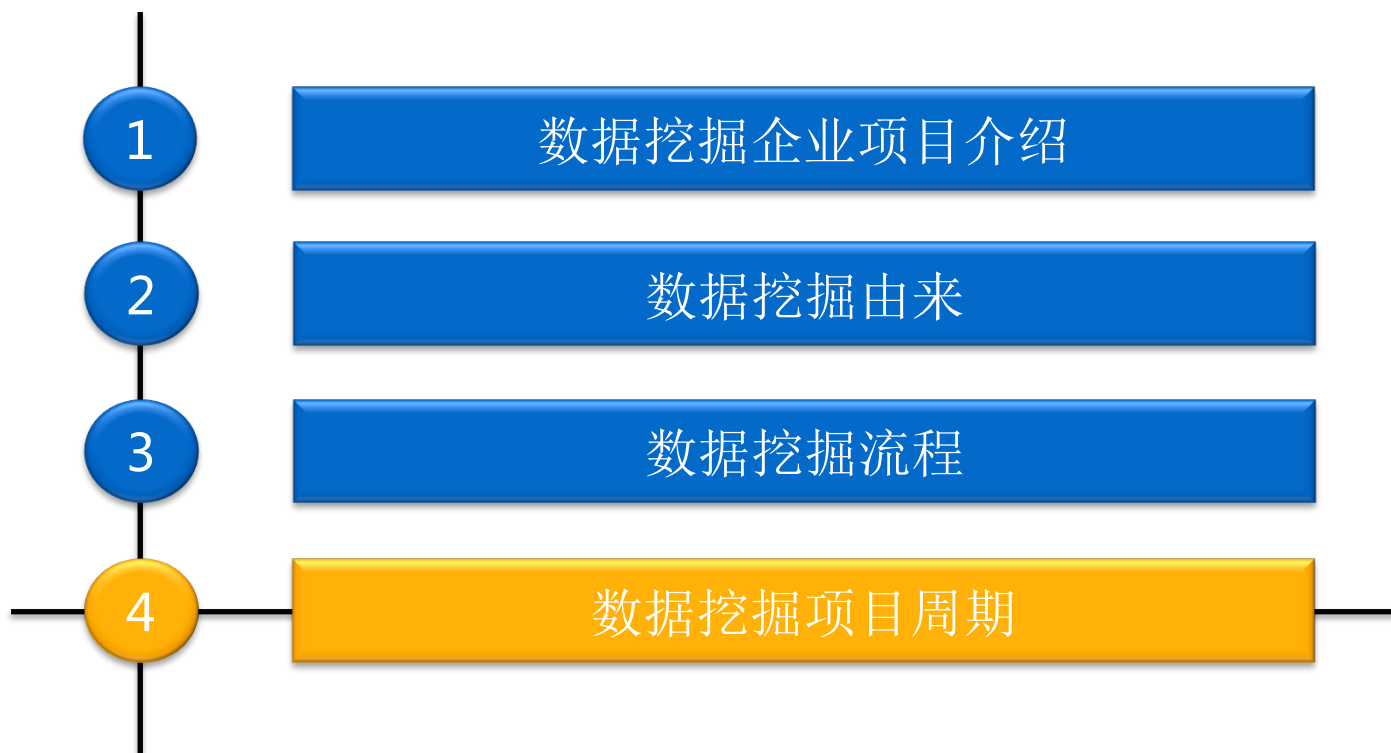


第7步：模型部署

- 自动外汇交易系统
- 车站人流实时监测与预测
- 天气预报
- 行为数据监测与犯罪者识别
- 窃漏电用户识别分析系统



目录



数据挖掘项目周期

- 项目预研
- 项目立项
- 需求确认
- 系统原型及可视化展示
- 合同签订
- 数据挖掘-数据抽取-数据探索预处理-挖掘建模-模型评价
- 项目部署
- 系统开发
- 项目评估



常用数据挖掘建模工具

- Python
- R
- MATLAB
- TipDM
- SAS Enterprise Miner
- IBM SPSS Modeler
- WEKA
- SQL Server





大数据成就未来



Thank you!

泰迪科技 : www.tipdm.com
热线电话 : 40068-40020

