第一讲 - 商务数据分析概论

张建章

阿里巴巴商学院

杭州师范大学

2022-09-01

杭州師範大學
Hangzhou Normal University

新大陆
Newland

目录

目录

课程考核说明

根据教学大纲要求，本课程的考核办法为：

$$\begin{aligned} \text{总成绩} = \text{期末成绩} \times 50\% \ &+ \text{日常作业} \times 30\% \\ + \text{日常考勤} \times 10\% \ &+ \text{课堂表现} \times 10\% \end{aligned}$$

其中，期末考试采用**上机考试**形式。

课程简介

课程名称：《商务数据分析实战》

课程目标：

① 掌握经典的数据分析方法；

② 培养数据驱动的商务计算思维；

③ 通过编程高效解决商务分析问题。

授课方式： 课堂讲授 + 实践案例

课程主旨：

联合**新大陆科技集团有限公司**开展产学研融合深度合作，推动高校商科人才培养改革。

实验环境

编程语言： Python 3.X

开发环境： Pycharm + Anaconda

交互环境： Jupyter-lab (Anaconda 已内置)

常用软件包： NLTK，scikit-learn，pandas，numpy，matplotlib，上述软件包 Anaconda 均已内置，MLxtend，huggingface，需要通过 pip 命令自行安装。

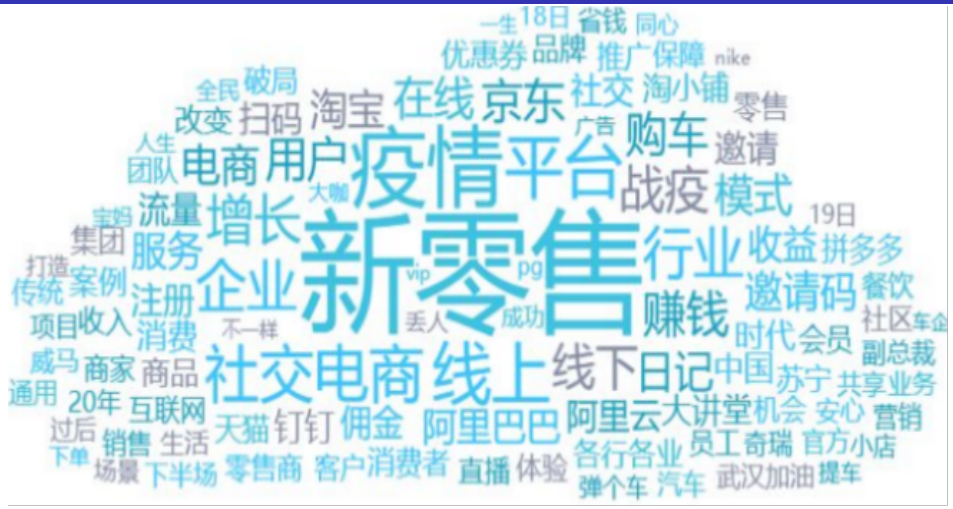
操作系统： Linux 桌面版 (推荐)，Windows，Mac OS (推荐)

在线环境： Kaggle (推荐)，Google colab

学习资源： Kaggle (推荐)，Towards Data Science (推荐)，Stack Overflow (推荐)，Github，CSDN，阿里云天池

应用场景

数据分析



文本分析

可视化

量化投资

目录

引入案例 – 扫码免费使用电子秤

这种电子秤一般都是医疗器械公司在药房或者人流量大的商业区放置，客户扫码免费用，可能要关注公众号，客户的 ID 信息以及测量信息就会被收集，通过分析，就可以策划不同类型保健品的重点推广区域，即**精细化营销**，若该企业购买微信广告服务，则不同粉丝会收到不同的广告，即**个性化营销**。

数据挖掘的发展历史

数据挖掘是指从数据集合中自动抽取隐藏在数据中的那些有用信息的非平凡过程，这些信息的表现形式为规则、概念、规律及模式等。

20 世纪下半叶，数据库中数据不断膨胀，简单查询和统计已无法满足企业的商业需求，同时，计算机领域的机器学习也取得巨大进展，人们将两者结合起来，用**数据库**管理系统存储数据，用**智能算法**分析数据，尝试挖掘数据背后的**知识**。

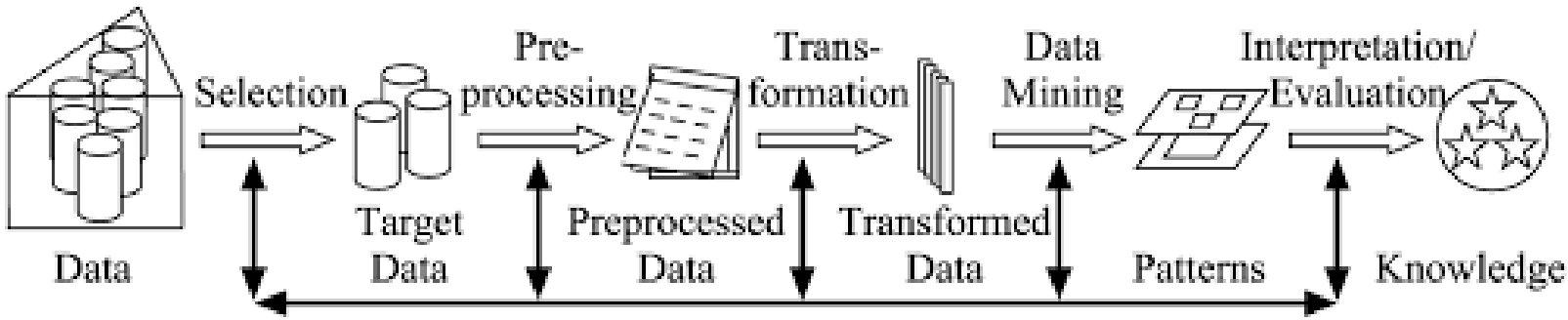


图 1: 数据库中的知识发现流程

统计分析 V.S. 数据挖掘

数据挖掘融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的理论和技术，**统计分析为数据挖掘提供重要的理论和技术支撑**。

图 2: 统计分析与数据挖掘，抓住老鼠就是好猫

商业实战中的数据分析思路：针对具体的业务，分析需求，确定分析思路，根据分析思路挑选和匹配合适的分析算法、分析技术，根据验证的效果和资源匹配等一系列因素进行综合权衡，从而决定最终的思路、算法和解决方案。

数据挖掘典型技术及其在商业中的应用

- **决策树**：不需要任何领域的知识，很适合探索式的知识发掘，并且可以处理高维度的数据，可用于用户划分、行为预测、规则梳理等；
- **神经网络**：就是通过输入多个非线性模型以及不同模型之间的加权互联，最终得到一个输出模型，是目前人工智能技术的主流，可用于用户划分、行为预测、营销响应等；
- **回归**：主要是指多元线性回归和逻辑斯蒂回归，在数据化运营中更多使用的是逻辑斯蒂回归，可用于响应预测、分类划分等；
- **关联规则**：找出数据集中的频繁模式，即多次重复出现的模式和并发关系，可用于挖掘顾客的购物习惯，制定有针对性的营销策略；
- **聚类**：将观察对象的群体按照相似性和相异性进行不同群组的划分，为业务方的精细化运营提供具体的细分依据和相应的运营方案建议；
- **其他方法**：贝叶斯分类、支持向量机、主成分分析等。

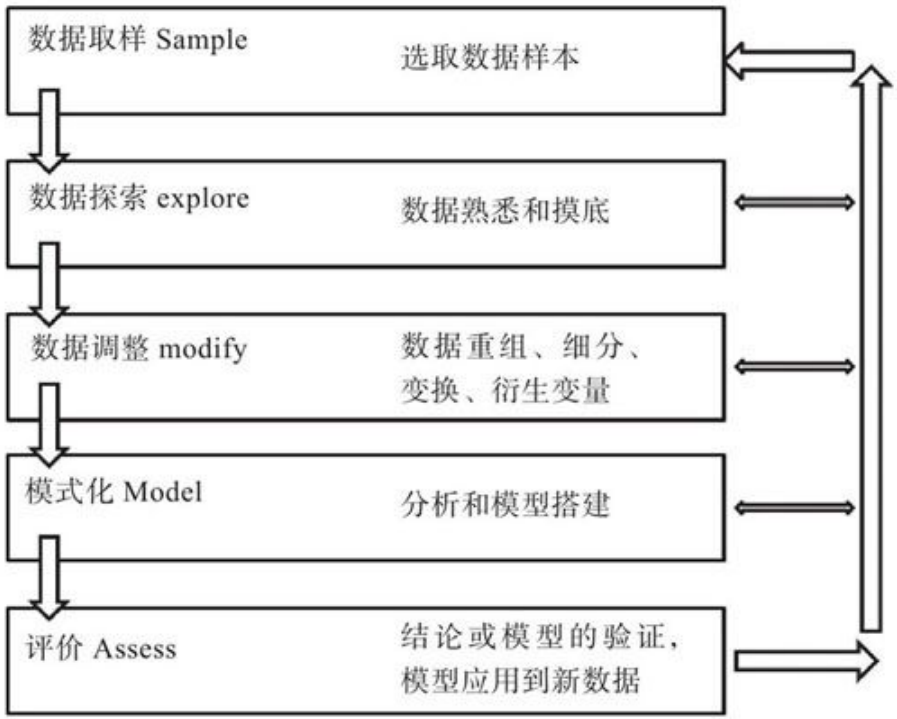
互联网行业数据挖掘应用的特点

- 数据的海量性；
- 数据分析（挖掘）的周期短；
- 数据分析（挖掘）成果的时效性明显变短；
- 互联网行业新技术、新应用、新模式的更新换代相比于传统行业而言更加迅速、周期更短、更加具有颠覆性，相应地对数据分析挖掘的应用需求也更为苛刻，且要多样化。

目录

SEMMA 方法论

SEMMA 是全球领先的商业分析软件与服务供应商 SAS 所提出的数据挖掘商业应用方法论，这 5 个英文字母分别代表 Sample（数据取样）、Explore（数据探索）、Modify（数据调整）、Model（模式化）、Assess（评价与评估）这 5 个核心环节，这 5 个环节可以按照 SEMMA 的顺序流转，在适当情况下各环节之间也可以相互流转。

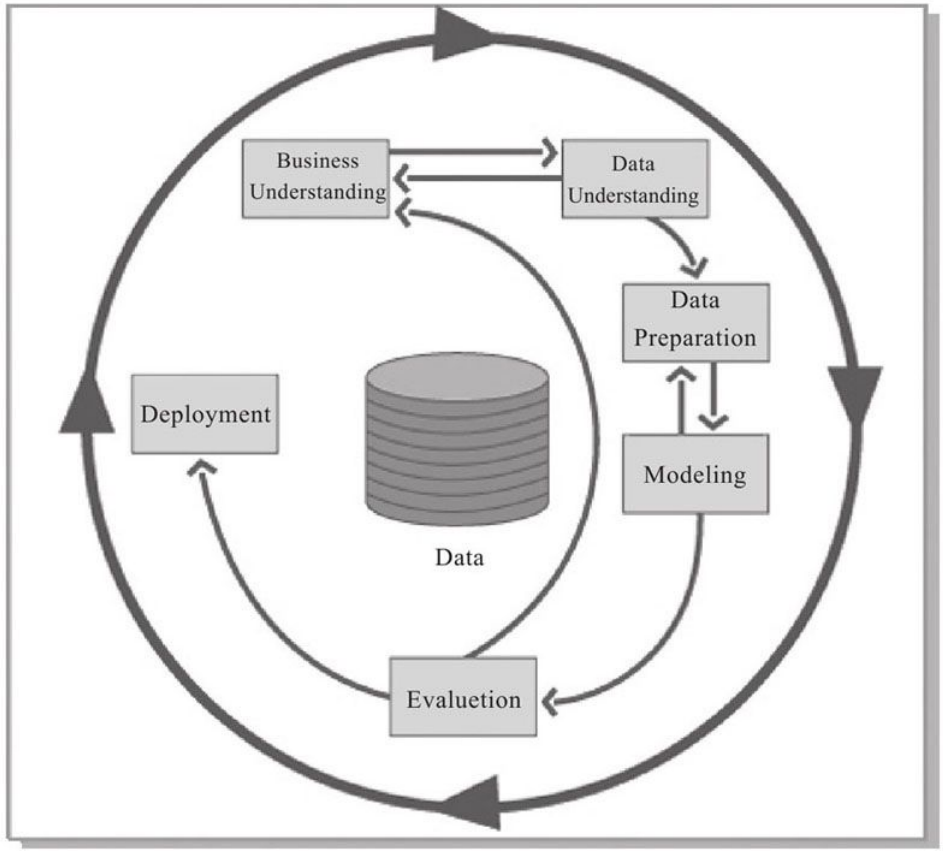


SEMMA 方法论

1. **数据取样**：从数据仓库海量数据中取出足够的有代表性的数据，同时又能有效节约计算资源；
2. **数据探索**：对数据进行深入摸底和熟悉的过程，如数据间的相关性、数据缺失情况和程度等；
3. **数据调整**：把之前所抽取的原始数据进行调整和转换，以使得数据能更加容易地反映出事物的内在规律和联系，模型的建立更加容易、更加有效；
4. **模式化**：搭建模型，发现知识；
5. **评价**：对模型和发现的知识进行综合评价和介绍，是最终体现数据挖掘和数据分析商业价值的环节。

CRISP-DM 方法论

CRISP-DM 方法论全称为 Cross-Industry Standard Process for Data Mining，即跨行业的数据挖掘标准流程, 一个数据挖掘商业实践的完整过程包括 6 个阶段，分别为业务理解、数据理解、数据准备、模型搭建、模型评估和模型发布。



CRISP-DM 方法论

1. **业务理解**：正确理解业务背景和业务需求，同时能把业务需求有效转化成合理的分析建模需求；
2. **数据理解**：从数据收集开始，通过一系列的数据探索和熟悉，识别数据质量问题，发现数据的内部属性；
3. **数据准备**：数据清洗、重组、转换及衍生等；
4. **模型搭建**：搭建模型，发现知识；
5. **模型评估**：彻底评估备选模型，挑选冠军模型，评价模型的稳定性，确保模型（或结论）正确回答了当初的业务需求；
6. **模型发布**：模型投入业务应用，产生商业价值，并且应用效果要及时跟踪和反馈，以便后期的优化和更新。

Tom Khabaza 的挖掘 9 律 I

Tom Khabaza 是 20 世纪 90 年代著名的数据挖掘工具平台 Clementine 的早期核心开发者之一，他总结的挖掘 9 律在数据挖掘业界产生了广泛的反响和认同。

1. **业务目标律**：业务目标是所有数据挖掘解决方案的本源，数据挖掘不是为了挖掘而挖掘，所有的数据挖掘都必须而且应该服务于特定的商业（业务）目的，离开了业务目的和业务应用，就没有数据挖掘的价值；
2. **业务知识律**：业务知识是数据挖掘每一步的核心，数据挖掘的本质就是将业务知识、经验和洞察力与数据挖掘方法相结合，从数据中发现有价值的东西；
3. **数据准备律**：数据准备能让数据挖掘流程事半功倍，其目的主要是让数据变动更干净，更能真实体现业务背景，更加容易被模型发现其隐含的有价值的商业信息和商业规律；

Tom Khabaza 的挖掘 9 律 II

4. **天下没有免费的午餐**：只有通过实际验证才能发现给定应用的正确模型，一个模型无论搭建过程如何完美，如果没有在实际数据中经过验证，就没有任何价值和意义；
5. **沃特金斯定律**：总会有模式存在，只要有数据，一定是可以从中发现有价值的信息的；
6. **数据挖掘将业务领域的感知放大**：得益于数据挖掘的技术和流程，使得数据中隐藏的知识和有价值的信息能被发现；
7. **预测定律**：预测将信息从局部扩展到整体，数据挖掘使得我们可以透过已知的去发现（某些）未知的；
8. **价值定律**：模型的价值只能由其所满足的业务需求和商业应用价值来决定，而不是由模型本身的精度和稳定性决定；
9. **变化定律**：任何模型或者分析结论都是有时间限制的，今天还是非常有价值的模型，或许明天就过时了，所有模型的维护和优化都非常重要；

目录

安装所需软件环境

1. 在你的机器上自行安装本课程所需的软件（Anaconda，Pycharm），并自行熟悉 Jupyter-lab 和 Pycharm 的基本用法；

2. 观看大数据时代纪录片第四集商业之变，了解大数据的开发和利用对商业生态的改变。

THE END