

第一章 绪论

1.1 背景及应用

1.2 基本概念

1.3 数据挖掘主要任务

1.4 本课程教学目标和安排

1.5 课外资源

第一章 绪论

1.1 背景及应用

1.1.1 相关领域

1.1.2 为什么要进行数据挖掘

1.1.3 主要应用

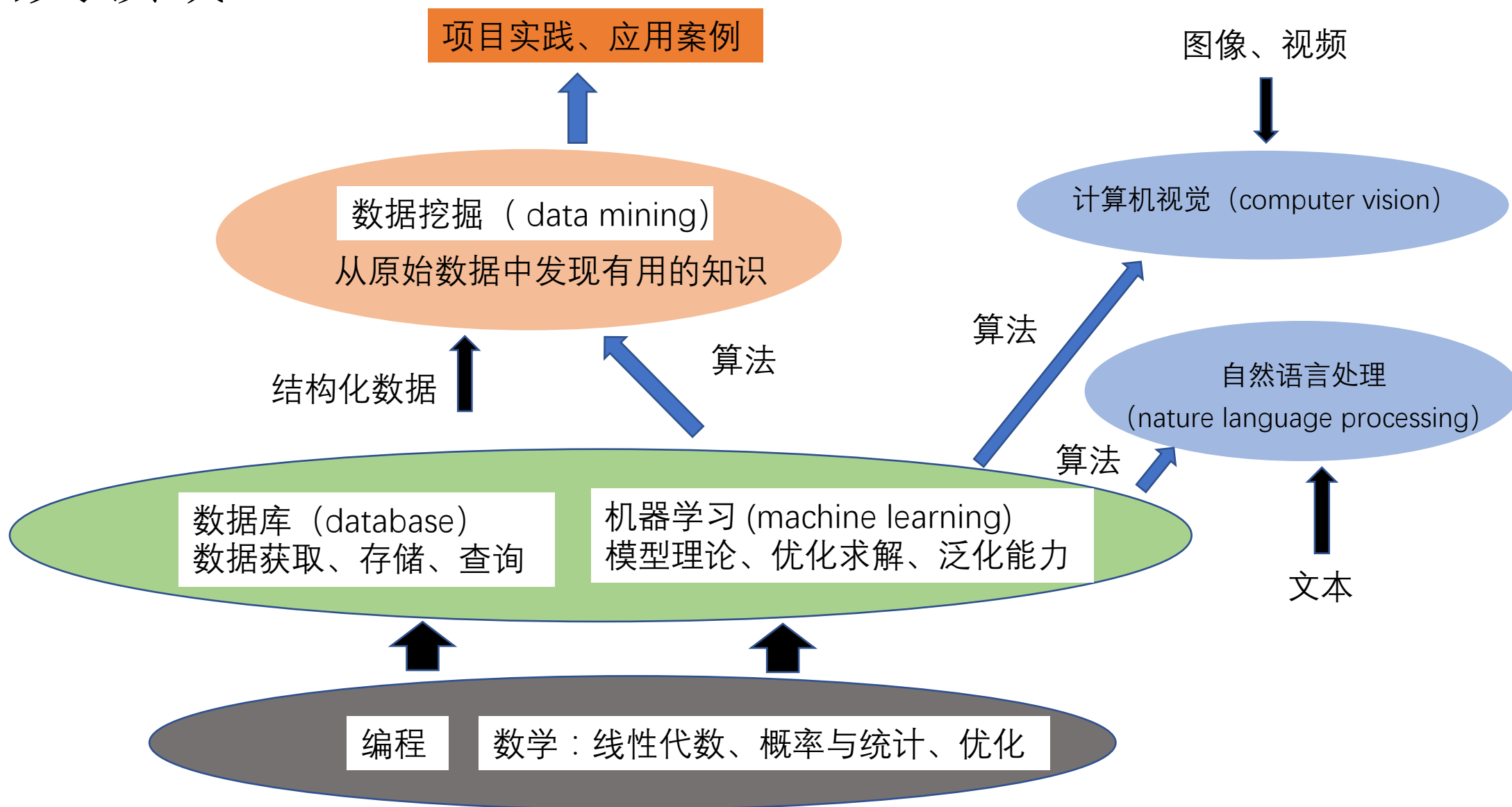
1.2 基本概念

1.3 数据挖掘主要任务

1.4 本课程教学目标和安排

1.5 课外资源

相关领域



为什么要进行数据挖掘？

数据是信息时代的“油田”。谁掌握了数据，谁就有能源！



互联网线上数据



“大数据”
你贡献了多少？

用户产生数据



科学实验



2019-2020中国人工智能算力发展评估报告

175ZB大概相当于**70000亿**部4K版哪吒之魔童降世。

技术融合将带动数据快速增长

Data growing 27.2% CAGR



WHAT IS A ZETTABYTE?

1,000,000,000,000gigabyte

1,000,000,000,000terabyte

1,000,000,000,000petabyte

1,000,000,000,000exabyte

1,000,000,000,000zettabyte

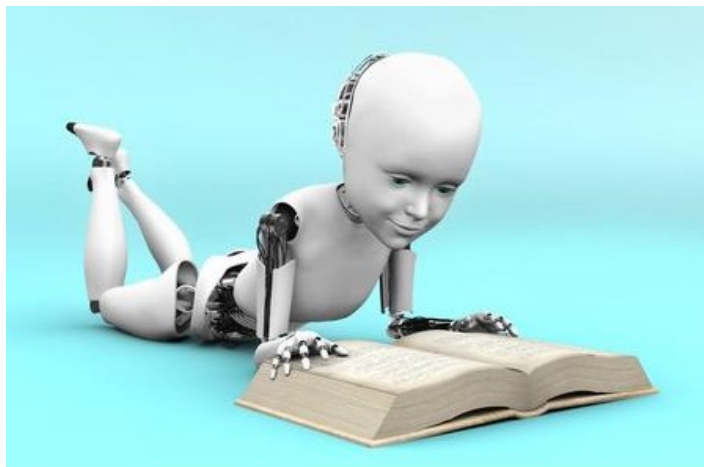
为什么要进行数据挖掘？

油田不等于汽油，数据 \neq 有用信息

数量大、结构复杂、产生快 --> 人脑不够用、太累、成本太高



怎么办？

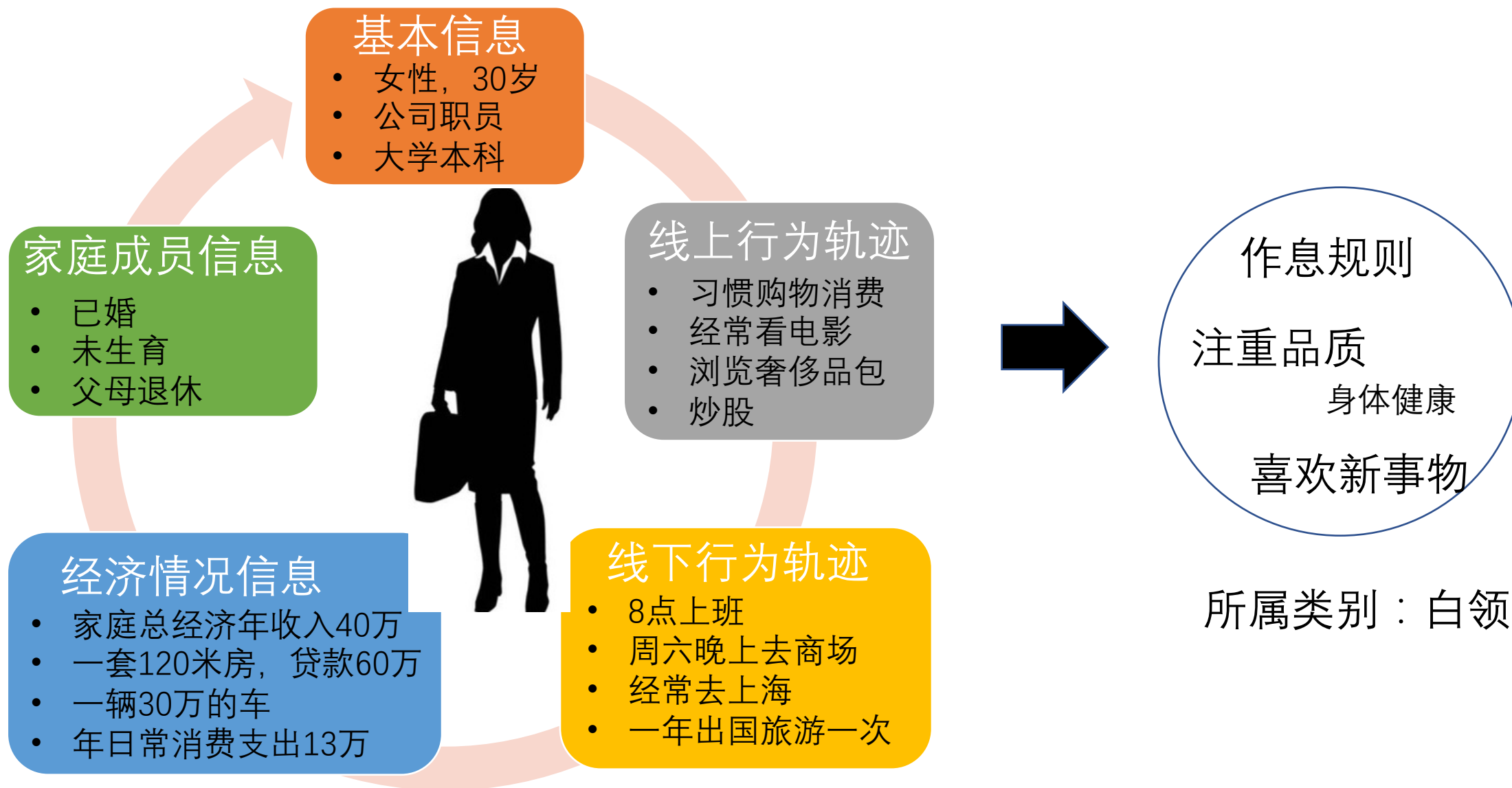


数据挖掘

数据挖掘的主要应用

- 用户历史行为数据挖掘，用于精准营销
- 文本挖掘，用于舆情分析
- 社交网络数据挖掘，用于社区检测
- 交通、出行轨迹分析和预测疫情传播情况
- 传统领域如工业制造流水线产生的数据
- 其他...

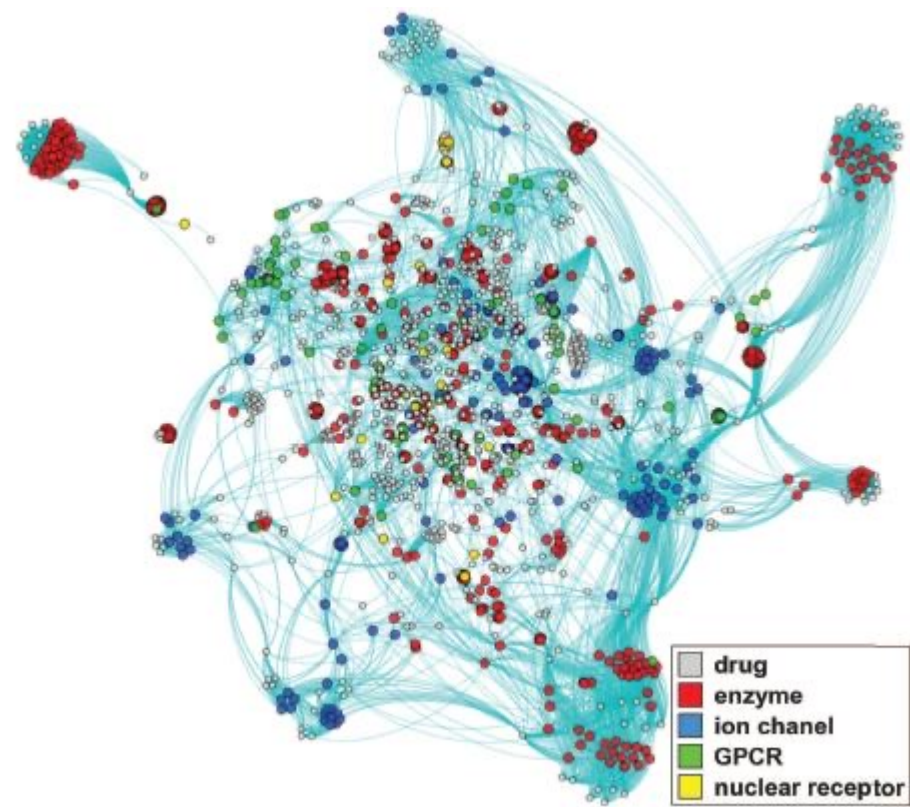
用户归类



网络挖掘

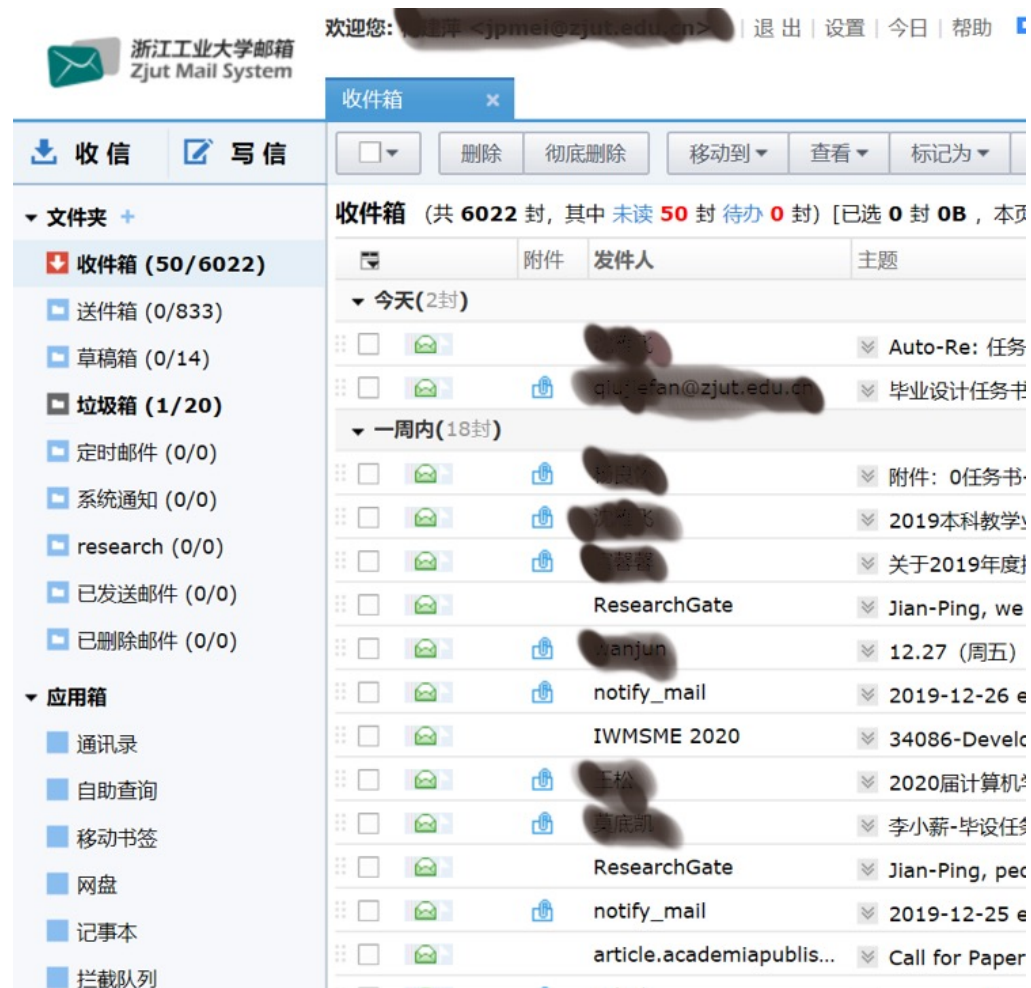


社交网络



蛋白质-药物作用网络

文本挖掘



电子邮箱-邮件分类

美丽人生的短评 ····· (全部 171157 条)

热门 / 最新 / 好友

老鸡 | 扶立 看过 ★★★★★ 2008-01-10

如果谎言可以这样美丽, 我也情愿生活在谎言之中

林愈静 看过 ★★★★★ 2006-04-21

看了这个不要看《辛德勒名单》或者看了《辛德勒名单》不要看这个。时间: 2005

寂地 看过 ★★★★★ 2006-01-05

即使是悲惨世界,也要大大的笑着.

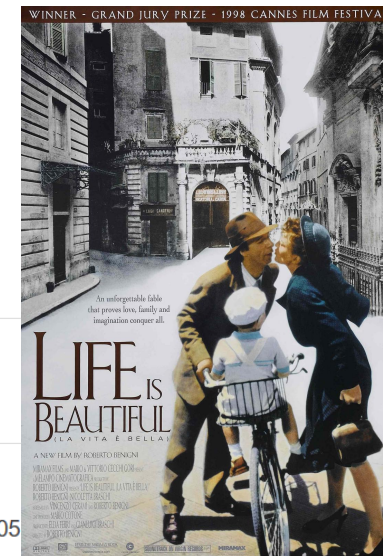
喜欢吗~ 看过 ★★★★★ 2017-02-08

三刷了。记得上学时跟父母吵过, 中国的父母经常想让孩子相信这世界的丑陋, 而国外的父母即便身在地狱也要让孩子相信在天堂, 让他们开心的活着, 你们为什么要让我的童年的这么痛苦。哈哈, 大概就是受这剧影响吧。

Lan~die 看过 ★★★★★ 2007-04-04

关于父爱的伟大电影。以非凡的想象力和诙谐幽默演绎一场不堪回首的历史惨剧, 那怦动的热情和对人生充满希望的美丽震撼人心。“为了看到阳光, 我们来到世上。为了成为阳光, 我们存于世上。”在Guido身上, 你看不到那些痛苦、隐忍、挣扎和艰难。这位一直用荒谬的态度对待人生的荒谬、以达观的态度对

> 更多短评 171157条



3963 有用

907 有用

499 有用

评论情感分析

第一章：绪论

1.1 背景及应用

1.2 基本概念

1.2.1 数据表示和类型

1.2.2 数据挖掘基本流程

1.3 数据挖掘主要任务

1.4 本课程教学目标和安排

1.5 课外资源

数据表示和类型

基本表示形式：

- 数据集中每个对象由一系列特征来描述：对象-属性

对象	属性					
	序号	色泽	根蒂	重量	甜度	敲声
	1	青绿	稍卷	3.3	高	清脆
	2	浅白	卷曲	3.5	一般	浑浊
	3	浅白	稍卷	2.9	高	清脆

对西瓜数据集，每个西瓜为一个对象（对应行），由色泽、根蒂等属性（对应列）来描述。

- 数据集中对象之间的关联关系：“对象-对象”

包括：图表示的数据（社交网络）、对象之间相似度

	对象1	对象2	对象3
对象1	1	0.8	0.3
对象2	0.8	1	0.6
对象3	0.3	0.6	1

注意：
“对象”又叫“样本”或“样例”，
“属性”又叫“特征”。

数据表示和类型

特征（或属性）主要分为以下几种类型：

连续特征(continuous)：取连续值

如房屋面积、价格。

你能举几个例子吗？

等级特征(ordinal)：取离散值但有大小

如收入等级取高、中、低；评价分1-5颗星

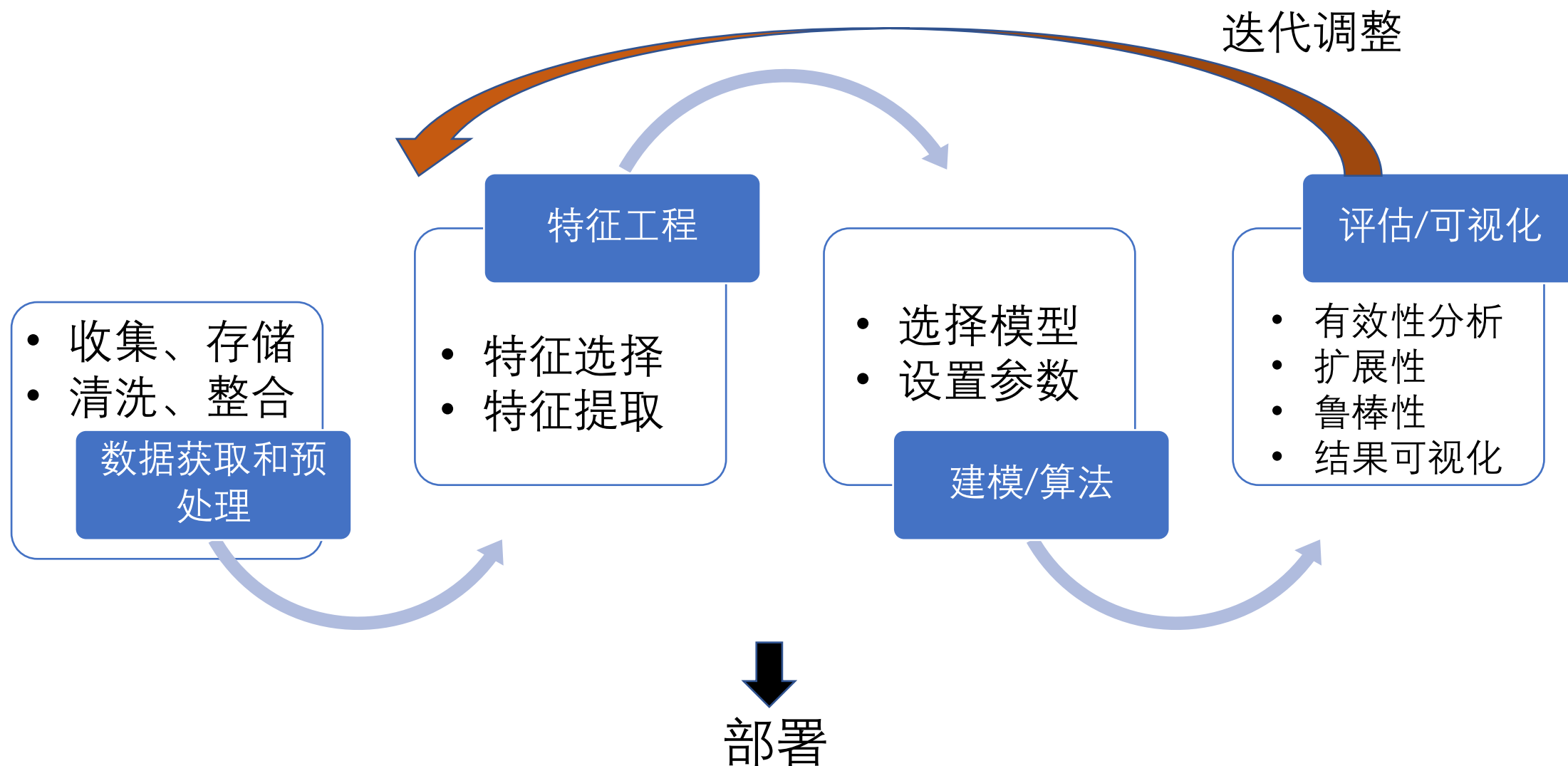
类别特征(categorical)：取离散值且没有大小

如性别、颜色。

有大小

离散

数据挖掘基本流程



数据挖掘基本流程-以电影票房预测为例

获取数据：

从m1095、票房网、豆瓣网等获取电影票房、质量、属性等数据

特征工程：

分析最重要的信息，最后选区客观衡量导演、演员水平，根据历史电影评分、导演信息、演员信息、票房信息、电影类型信息、评价信息等特征进行组合最终共有74个特征。

挖掘算法：因为是预测连续值，用回归

误差分析：与真实值的最小均方误差（Mean Square Error）

模型训练好之后进行部署应用。

流浪地球2 (2023)



导演: 郭帆

编剧: 杨治学 / 龚格尔 / 郭帆 / 叶濡畅

主演: 吴京 / 刘德华 / 李雪健 / 沙溢 / 宁理 / 更多...

类型: 科幻 / 冒险 / 灾难

制片国家/地区: 中国大陆

语言: 汉语普通话 / 俄语 / 英语 / 印地语 / 法语

上映日期: 2023-01-22(中国大陆)

片长: 173分钟






又名: The Wandering Earth II / The Wandering Earth 2 /

《流浪地球》前传

IMDb: tt13539646

豆瓣评分

8.2  999686人评价

5星  41.5%
4星  36.0%
3星  17.2%
2星  3.7%
1星  1.5%

好于 95% 科幻片

好于 97% 灾难片

想看

看过

评价: ☆☆☆☆☆

 写短评  写影评  分享到 ▼

推荐

流浪地球2的剧情介绍 ·····

太阳即将毁灭，人类在地球表面建造出巨大的推进器，寻找新的家园。然而宇宙之路危机四伏，为了拯救地球，流浪地球时代的年轻人再次挺身而出，展开争分夺秒的生死之战。

发展趋势

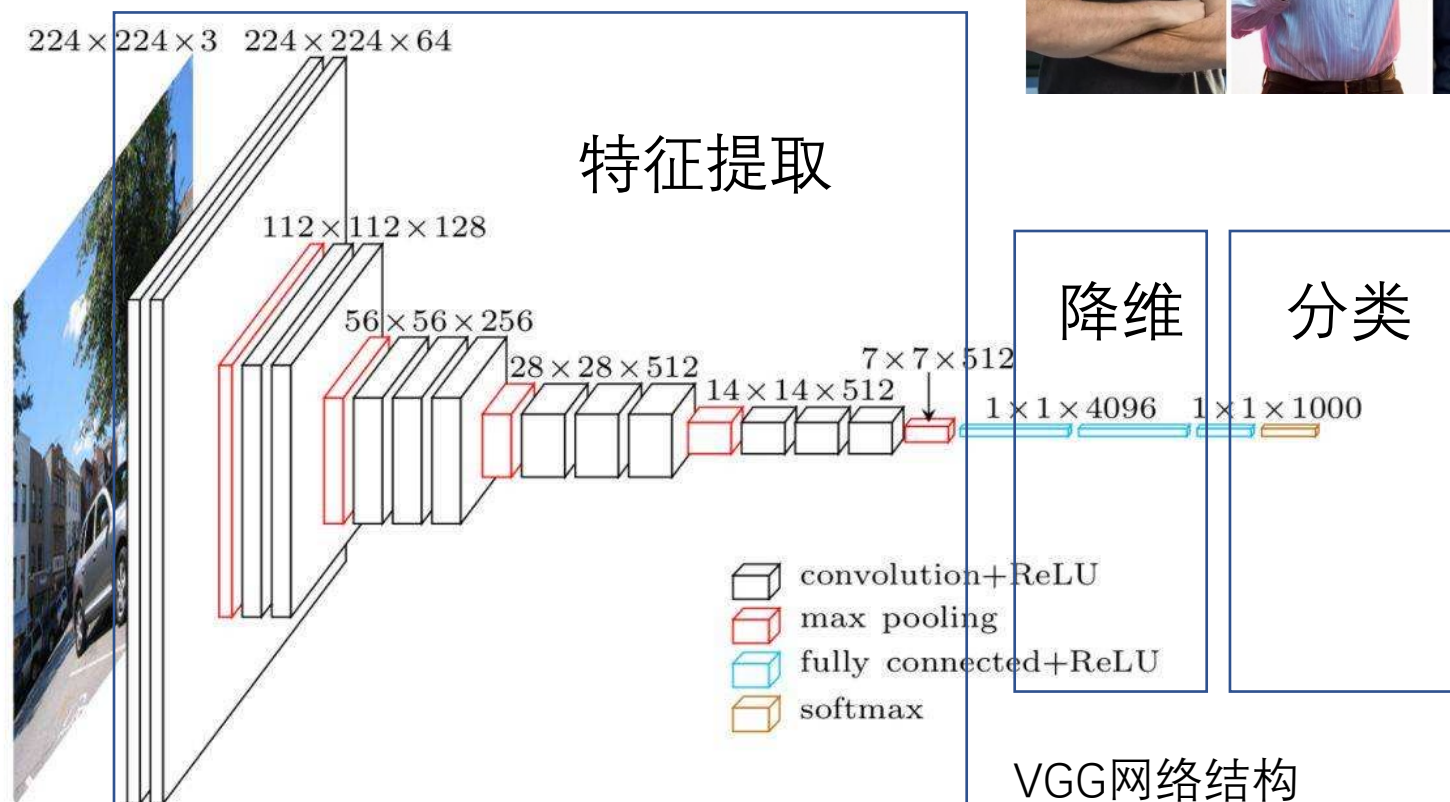
特征工程
组合式

传统方法

→ 特征学习
→ 端到端

➔ 深度神经网络

2018年度图灵奖



第一章：绪论

1.1 背景及应用

1.2 基本概念

1.3 数据挖掘主要任务

1.3.1 分类

1.3.2 聚类

1.3.3 关联规则挖掘

1.4 本课程教学目标和安排

1.5 课外资源

分类

监督学习

过程：基于已知类标签的样本训练一个分类器或模型。

目标：训练好的模型对未知样本的分类尽可能准确。

特征或属性

类标签

面积(m ²)	房间数目	是否学区房	离地铁站距离(km)	交付房价(万/m ²)	一年后房价是否涨
120	3	是	1.5	2.5	是
90	2	否	1.0	2.0	否
90	3	是	2.0	1.8	是

面积(m ²)	房间数目	是否学区房	离地铁站距离(km)	交付房价(万/m ²)	一年后房价是否涨
120	4	否	2.5	2.7	?



Model

Learn Classifier

如果希望预测第二年的房价呢？

回归 (regression)

监督学习



应用：电影票房、股票价格预测等。

给定一个训练集，其中每个样本的标签为连续值；用该训练集学习一个模型用于预测新样本的输出值。

回归模型的学习可以理解为对一个连续函数的拟合过程。

回归用的标签（或输出）是连续值，分类的标签是离散值（类别）。

面积(m ²)	房间数目	是否学区房	离地铁站距离(km)	交付房价(万/m ²)	一年后房价(万)
120	3	是	1.5	2.5	3.0
90	2	否	1.0	2.0	2.2
90	3	是	2.0	1.8	2.4

测试样本

面积(m ²)	房间数目	是否学区房	离地铁站距离(km)	交付房价(万/m ²)	一年后房价(万)
120	4	否	2.5	2.7	?

聚类 无监督学习

过程：给定一个无标签的数据集，把数据集中的样本分组，又叫簇。

目的：同一个簇的样本之间的相似度大于不同簇的样本间的相似度。

数据集中的样本没有标签

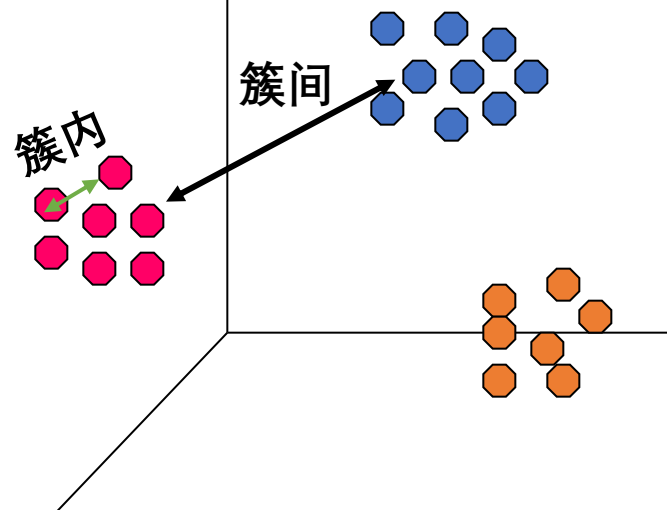
大小	房间数目	是否学区房	离地铁站距离	交付房价 (万/平米)
120	3	是	1.5	2.5
90	2	否	1.0	2.0
90	3	是	2.0	1.8



聚类是一种无监督学习方法

最小化簇内距离

最大化簇间距离



分类和聚类：共同点与差异

共同点

找出数据集中样本之间的分组/类别关系

差异

分类前已经知道几个类，以及每个类分别代表什么；一般需要标记好类别的样本作为训练集；

聚类前不清楚簇的数目以及每个簇表示什么；一般不需要标签而直接基于样本的特征或样本之间的关系进行分组。

如何选择：如果由足够多标记数据，则考虑分类，否在考虑聚类。

关联规则挖掘

对象：记录/交易集，每条记录为多个商品的集合；

目的：挖掘重要的商品共现 (co-occurrence) 关系。

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

发现的规则:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$

应用：商品捆绑营销、超市货品摆放

第一章：绪论

1.1 背景及应用

1.2 基本概念

1.3 数据挖掘主要任务

1.4 本课程教学目标和安排

1.5 课外资源

本课程教学目标

■理解基本的数据挖掘流程：数据预处理、算法、评估

■掌握常用方法和算法：

- 数据预处理：缺失数据处理、噪声处理、规范化
- 相似度和相异度衡量：距离度量、余弦相似度等
- 降维方法：主成分分析
- 分类算法：K最近邻、决策树、朴素贝叶斯
- 聚类算法：层次聚类、K均值
- 关联规则挖掘：Apriori

■理解相关基本概念：监督与无监督、过拟合等

■了解基于优化进行建模的基本方法

注意：本课程是一门入门课程，不包含高级机器学习算法，比如核函数、半监督学习、深度学习。

基本算法：线性回归、逻辑回归、SVM、神经网络将在《机器学习》课程中学习。

教学计划-理论课（根据具体情况可能会略有调整）

课次	章节	主要内容
1	绪论	背景、相关领域、基本概念、应用案例、课程内容、课外资源
2	数据认知与预处理	特征类型及转换、相似度衡量、缺失值处理、数据规范化
3	降维	降维的作用、主成分分析原理和算法
4	分类：最近邻、模型评估	KNN方法、模型评估：验证方法和度量
5	分类：决策树	基于决策树的预测、信息熵、构造决策树算法（C4.5）、剪枝
6	分类：朴素贝叶斯	生成式方法、贝叶斯公式、朴素贝叶斯算法（朴素的含义、算法步骤）
7	组合分类	Boosting和Bagging框架及各自代表性算法
8	分类案例学习	基于jupyter notebook的分类案例实现和结果分析
9	聚类：k-均值	目标函数、算法、k的值、初始化
10	聚类：层次聚类	基本方法、不同Linkage
11	聚类案例学习	聚类案例实现与分析
12	关联规则挖掘	常用应用、主要问题、Apriori算法
课外自学与拓展		
自学	基于python的数据分析	Pandas, scikit learn, matplotlib
拓展	案例：文本分析	文本表示、文本归类、情感分类

教学计划-上机实验课（根据具体情况可能会略有调整）

课次	章节	主要内容
	课前自学	基于python的数据分析，熟悉numpy, pandas, sklearn, matplotlib
1	预处理、主成分分析	规约前后的影响、降维及可视化；
2	模型评估、k最近邻	计算分类器性能度量、比较不同评估方法、实现交叉验证调参
3	决策树、朴素贝叶斯	决策树分类、朴素贝叶斯分类
4	K均值、层次聚类	K均值聚类、层次聚类(不同linkage)

目的：实现基本算法，了解各个算法的基本特点；

课后作业（加深对概念、算法的理解）

包括：向量之间距离、相似度计算、决策树构建、朴素贝叶斯分类、聚类

第一章：绪论

1.1 背景及应用

1.2 基本概念

1.3 数据挖掘主要任务

1.4 本课程教学目标和安排

1.5 课外资源

课外资源

• 参考书

1. 《机器学习实战》 Peter Harrington（著）-对每个算法的python实现
注意，该书中代码是基于Python 2, 直接在Python 3下运行可能会出现问题。
2. 《Python数据挖掘入门与实践》 Robert Layton（著）-基于scikit-learn

• 在线课程

1. Andrew Ng（斯坦福）：Machine Learning。主要内容：线性回归、过拟合、前馈神经网络、梯度下降等基本概念。
2. Andrew Ng：Deep Learning。深度神经网络基础、主要结构、热门应用。

课外资源

- 数据集

1. UCI Machine Learning Repository [website](#)
2. Kaggle 竞赛、数据集 [website](#)

- 比赛平台

1. 阿里云天池大赛 [website](#)
2. CCF大数据与计算智能大赛 [website](#)
3. 国际数据挖掘顶级会议相关竞赛KDDCUP [website](#)

课后作业（截止时间：下次上课前一天）

自学相关资料，熟悉 jupyter notebook 的基本操作和基于python的数据分析，尤其是pandas、sklearn以及matplotlib库的基本用法：

在学习通作业->自学-基于python的数据分析**上传对应的html文件**：

注意：不同版本在具体方法的调用上可能稍有不同，按照版本学习对应文档。