

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

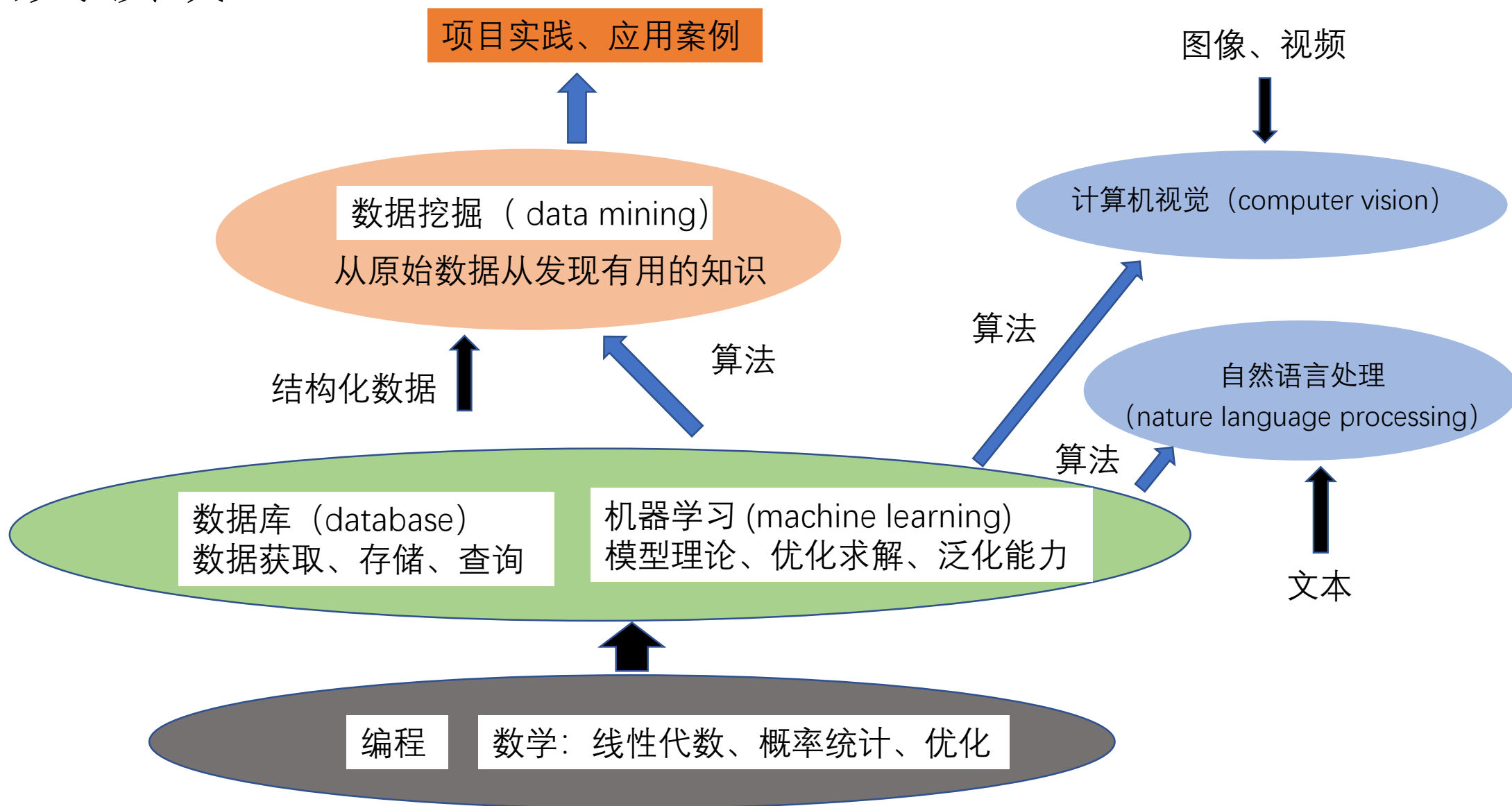
1.2 基本概念

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

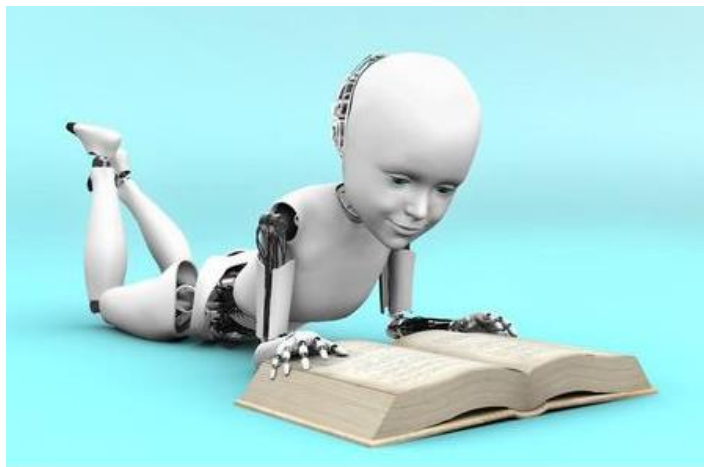
为什么要进行数据挖掘？

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？

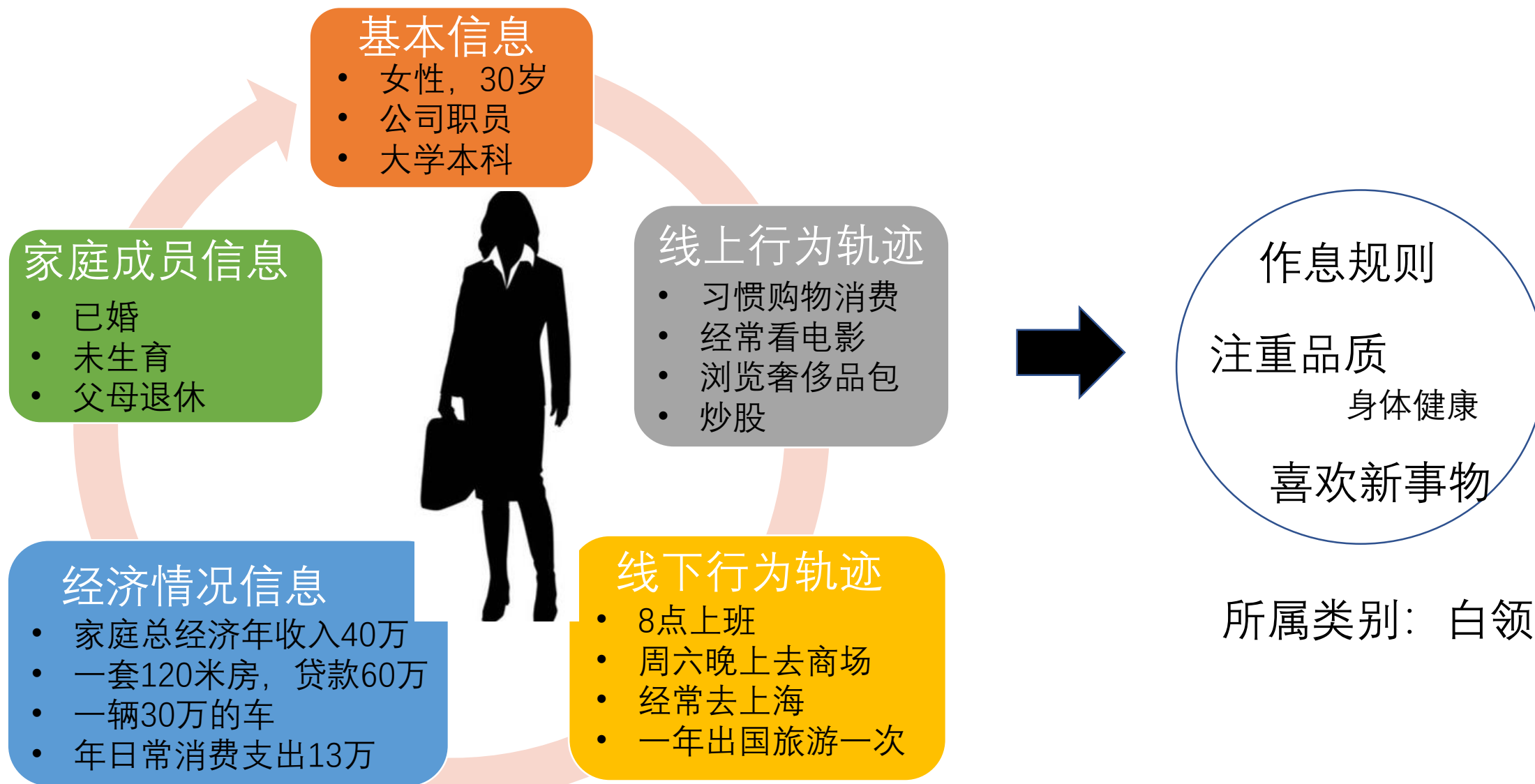


数据挖掘

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...

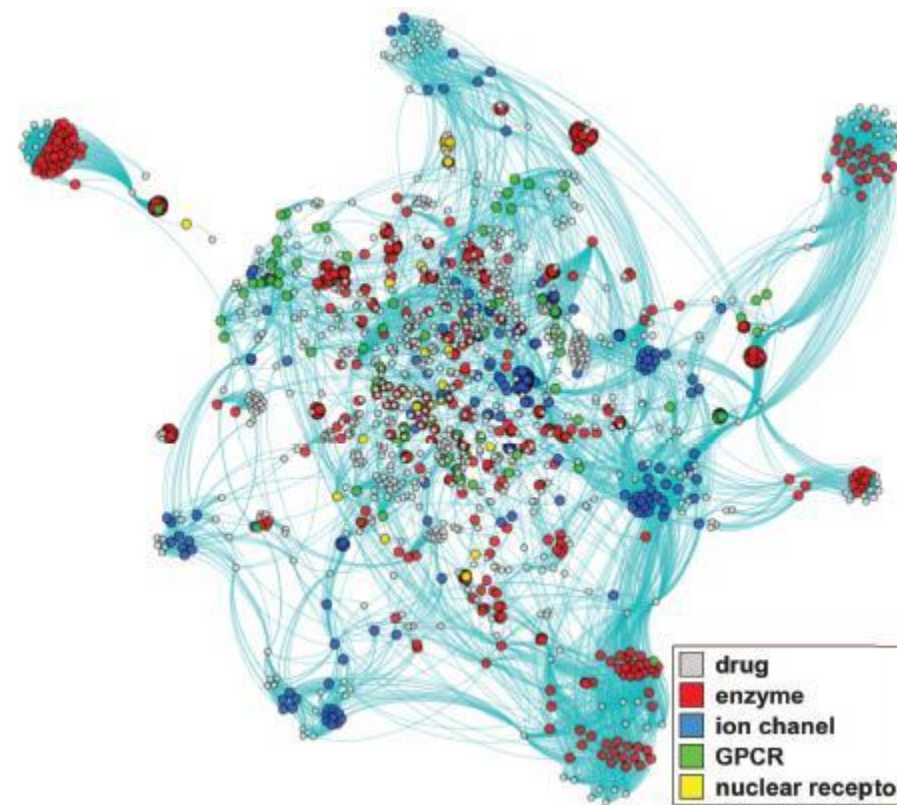
用户归类



网络挖掘

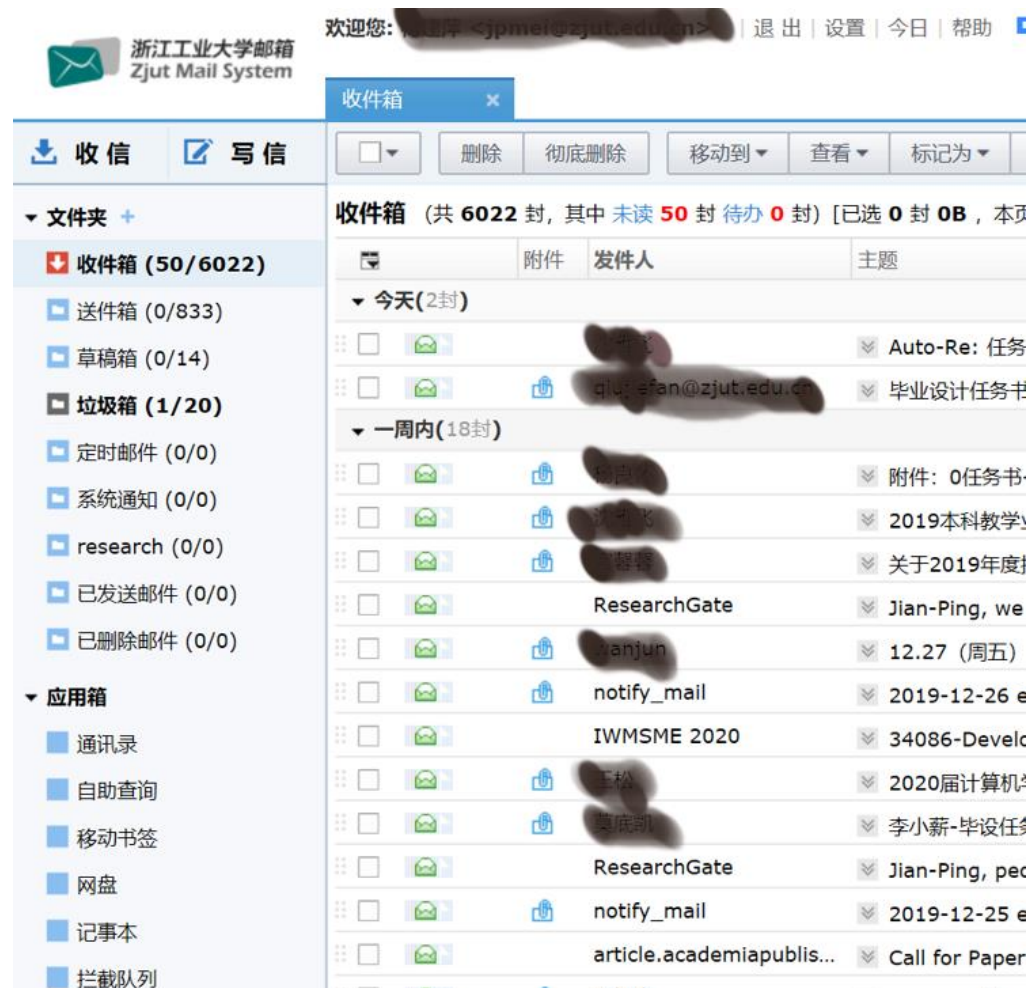


社交网络



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.

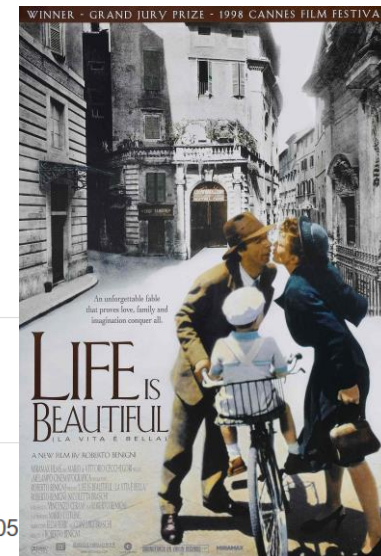
喜欢吗~ 看过 ★★★★★ 2017-02-08

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, 那悸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



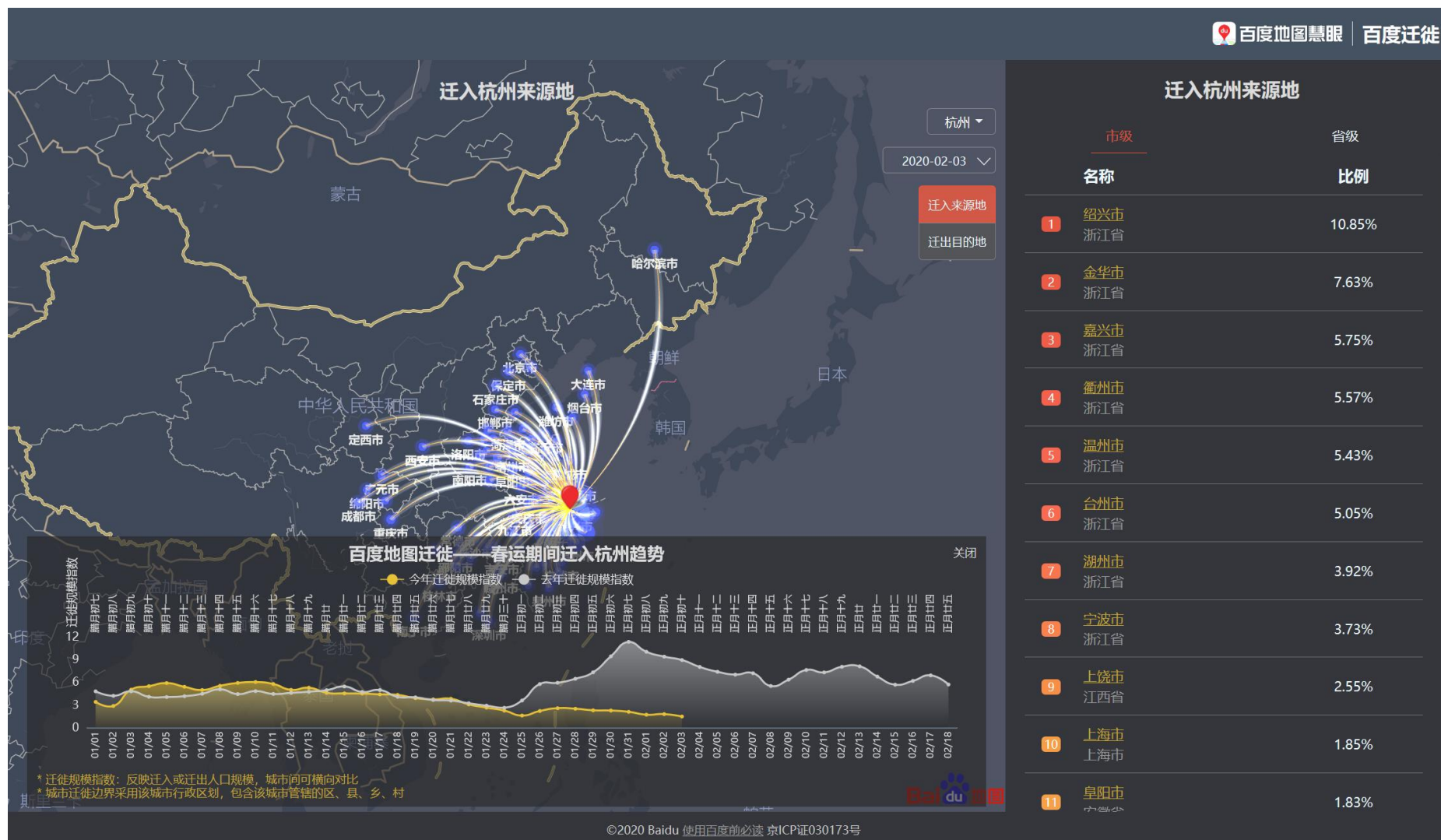
3963 有用

907 有用

499 有用

评论情感分析

疫情分析



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：

有大小

连续特征(continuous)：取连续值

如房屋面积、价格。

等级特征(ordinal)：取离散值但有大小

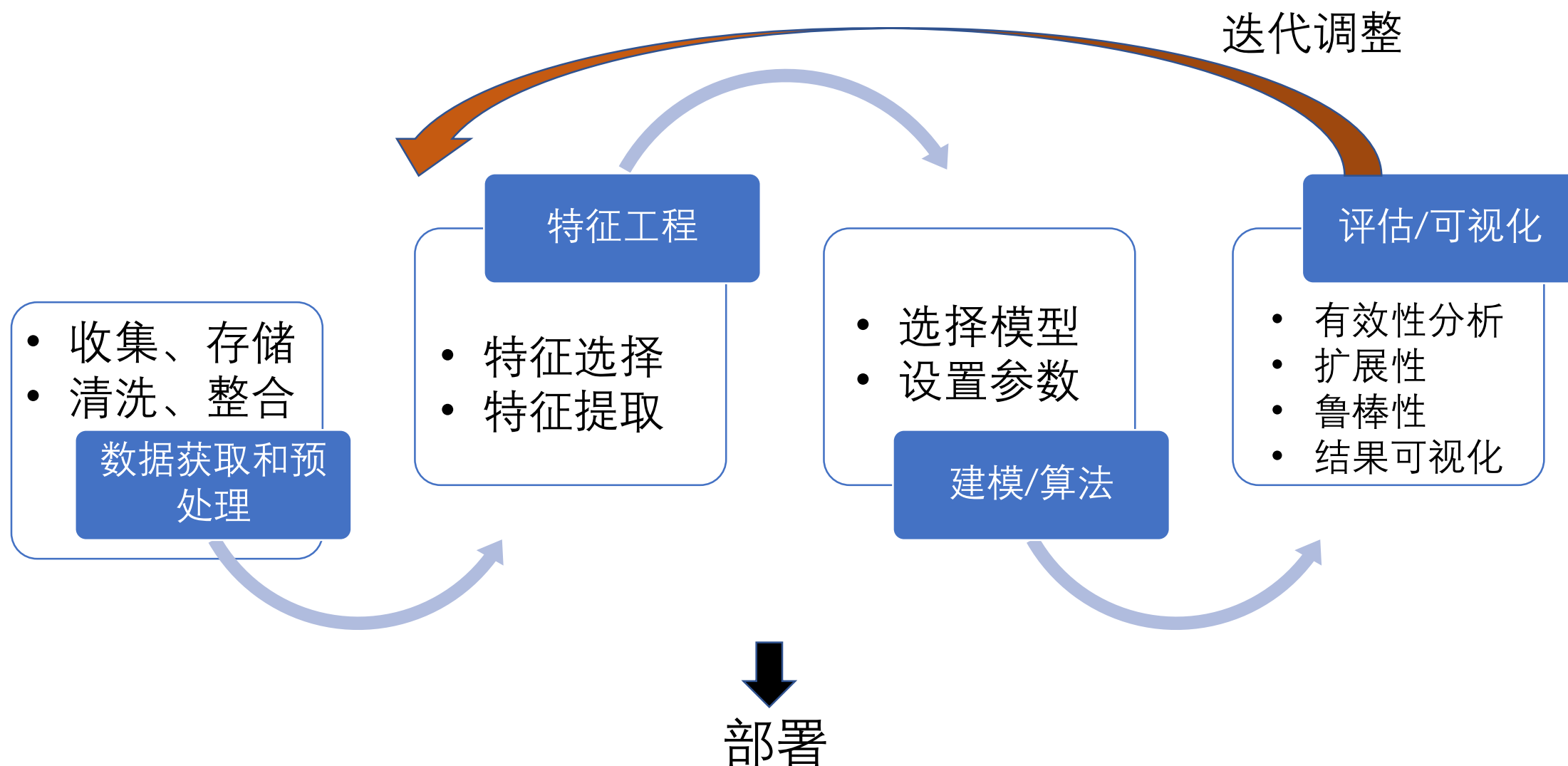
如收入等级取高、中、低；评价分1-5颗星

离散

类别特征(categorical)：取离散值且没有大小

如性别、颜色。

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 (2019)



导演: 郭帆

编剧: 龚格尔 / 严东旭 / 郭帆 / 叶俊策 / 杨治学 / 吴荑 / 叶濡畅 / 沈晶晶 / 刘慈欣

主演: 屈楚萧 / 吴京 / 李光洁 / 吴孟达 / 赵今麦 / 更多...

类型: 科幻 / 灾难

制片国家/地区: 中国大陆

语言: 汉语普通话 / 英语 / 俄语 / 法语 / 日语 / 韩语 / 印尼语


上映日期: 2019-02-05(中国大陆)




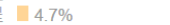

片长: 125分钟

又名: The Wandering Earth

IMDb链接: [tt7605074](https://www.imdb.com/title/tt7605074)

豆瓣评分

7.9  1378670人评价

5星  32.9%
4星  37.9%
3星  22.1%
2星  4.7%
1星  2.4%

好于 87% 科幻片

好于 88% 灾难片

想看

看过

评价: ☆☆☆☆☆

发展趋势

特征工程
组合式

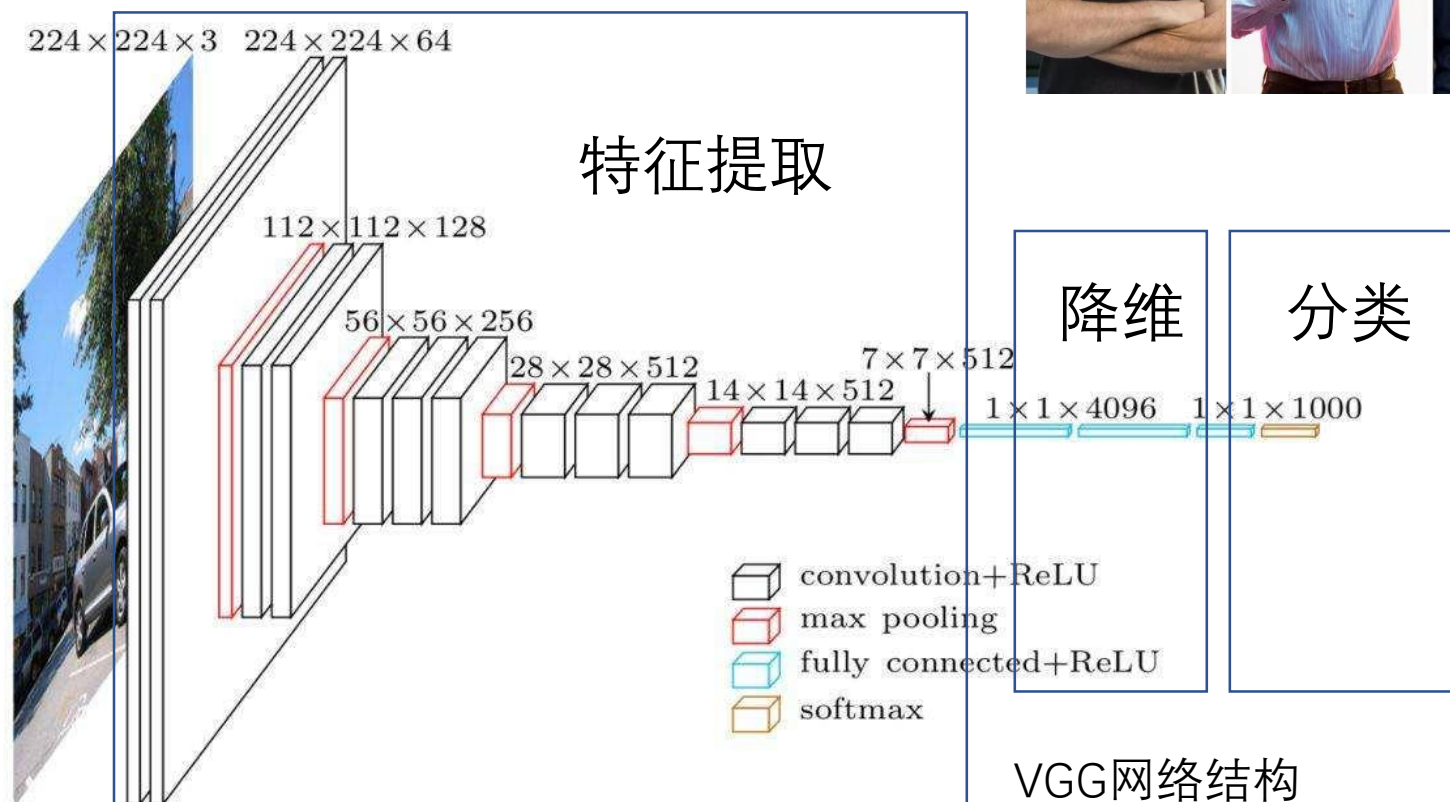
传统方法

→ 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。

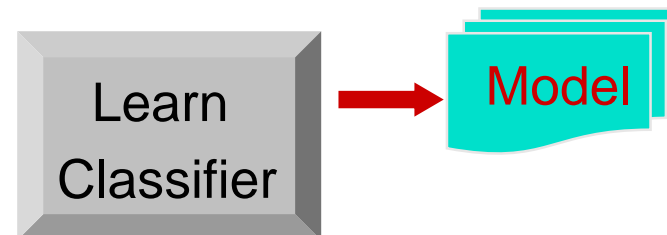
特征或属性

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	?

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

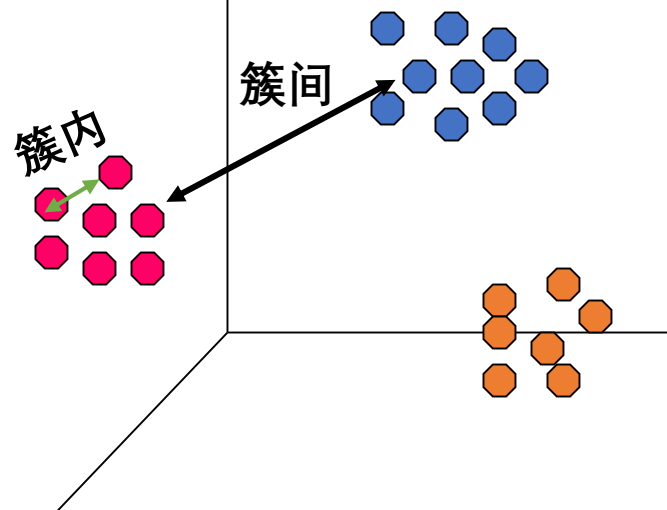
大小	房间数目	是否学区房	离地铁站距离	交付房价 (万/平米)
120	3	是	1.5	2.5
90	2	否	1.0	2.0
90	3	是	2.0	1.8



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： 如果有足够多标记数据，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

- 基于python，对特征表示的数据进行分类、聚类、关联规则挖掘
- 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、密度聚类
 - 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
2	绪论	基本过程、主要任务、课程内容、课外资源
3	python基础	基于numpy的基本语法、向量、矩阵运算
4	Python数据分析	数据读入、统计分析、画图
5	数据类型、相似度衡量	特征类型、连续-离散之间转换、相似度衡量
6	预处理：清洗、规范化	缺失值处理、去噪、数据规范化
7	降维	为什么要降维？主成分分析原理和算法
8	分类：最近邻、模型评估	KNN方法、模型评估：结果度量、交叉验证
9	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
10	分类：朴素贝叶斯	生成式方法、贝叶斯公式、为什么要朴素？朴素贝叶斯方法
11	组合分类	Boosting和Bagging框架及各自代表性算法
12	聚类：k-均值	目标函数、算法、k的值、初始化
13	聚类：层次聚类	基本方法、不同Linkage
14	聚类：密度聚类	DBSCAN算法及讨论
15	关联规则挖掘	常用应用、主要问题、Apriori算法
16	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
1	Python基本语法	编程环境、numpy 基本语法
2	Pyhton数据分析、画图	数据读入、dataframe操作、统计、画图
3	预处理、主成分分析	规约前后的影响、降维及可视化；
4	模型评估、性能度量	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
5	决策树	实现和测试决策树分类
6	朴素贝叶斯、k最近邻	实现和测试朴素贝叶斯分类、分析K最近邻对距离和k的敏感度
7	层次聚类、密度聚类	层次聚类(不同linkage)、密度聚类DBSCAN对参数敏感度
8	关联规则挖掘	实现基本Apriori算法

目的：实现基本算法，了解各个算法的基本特点；

课后作业（从实验中总结出结论；撰写逻辑清晰、格式规范的实验报告）

作业1	组合分类
作业2	K 均值（不同k, 距离、初始化）

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

- Python

1. Scikit 用户手册。 [website](#)

2. 免费电子书: Think Stats (TS) by Allen B. Downey. [PDF](#) | [website](#)

- 机器学习

在线课程: 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。

- 深度学习

在线课程: Andrew Ng: Deep Learning。深度神经网络基础、主要结构、热门应用。

深度学习入门-基于python的理论与实现 (斋藤康毅)

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](#)