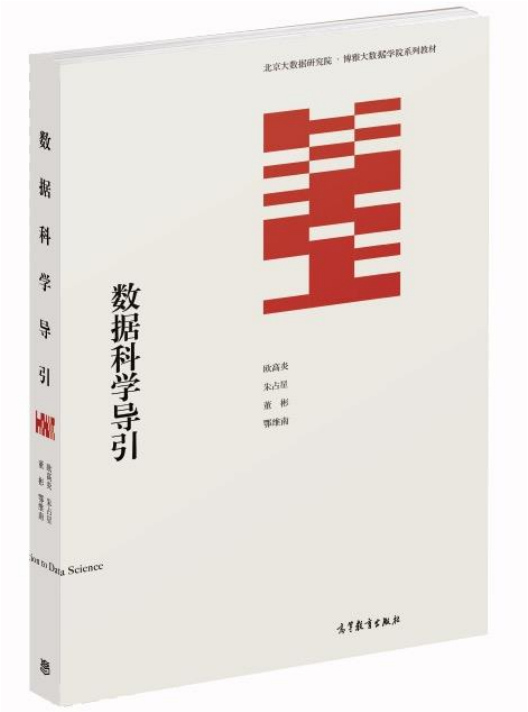
DS 5001 数据科学导引

中国科学技术大学 University of Science and Technology of China

平时成绩40% + 小论文20% + 实践项目40%

- 小论文自选topic，独立完成
- 实践项目组队完成，每组2-4人

参考教材:



课程计划:

1. 绪论&数据预处理
2. 回归模型
3. 分类模型
4. 聚类模型
5. 关联规则挖掘
6. 集成模型
7. 降维
8. 概率图模型
9. 深度学习
10. 文本分析
11. 图与网络分析
12. 知识图谱 (新)
13. 信息检索 (新)
14. 推荐系统 (新)
15. 分布式计算

教师:

何向南
hexn@ustc.edu.cn

徐童
tongxu@ustc.edu.cn

助教:

赵伟豪 博二
zhaoweihao@mail.ustc.edu.cn

吴剑灿 博一
wjc1994@mail.ustc.edu.cn

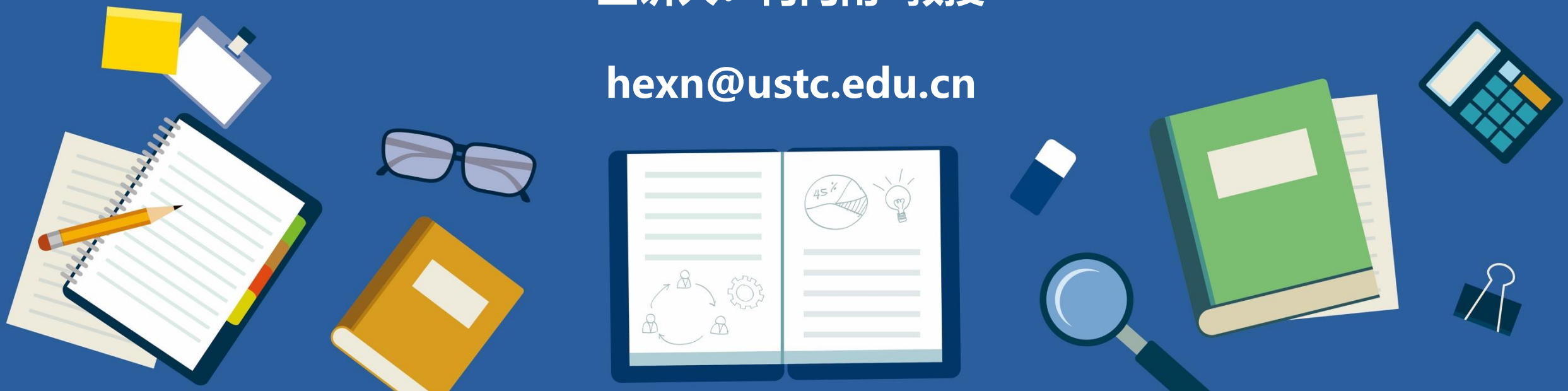
博雅大数据学院
BOYA BIG DATA COLLEGE

数据酷客官网

绪论：数据科学介绍

主讲人：何向南 教授

hexn@ustc.edu.cn



目录

- 数据广泛性

- 数据多样性
  - 结构化数据
    - 关系数据
  - 非结构化数据
    - 网页
    - 文本
    - 图像
    - 视频
    - 语音

- 数据分析的本质都是在解“反问题”

数据科学的内涵

- 用数据的方法研究科学
  - 生物信息学、天体信息学、地球科学等
- 用科学的方法研究数据
  - 统计学、机器学习、数据挖掘、数据库

用数据的方法研究科学的例子：开普勒三大定律

开普勒：分析数据产生价值

表 1 太阳系八大行星绕太阳运动的数据

行星	周期（年）	平均距离	周期^2/距离^3
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

开普勒第三定律:
行星绕太阳运行的周期的平方和行星离太阳的平均距离的立方成正比

用科学的方法研究数据

主要包括：数据采集、存储、和分析

- 常见的数据类型
  - 表格：最为经典的数据，e.g., 行代表样本、列代表特征
  - 点集：很多数据都可以看成是某种空间的点的集合
  - 时间序列：文本、通话和DNA序列等都可以看成是时间序列
  - 图像：可以看成两个变量的函数
  - 网页和报纸：每篇文章都可以看成是时间序列，整个网页或报纸又具有空间结构
  - 网络数据：网络数据本质上是图，由节点和联系节点的边构成

数据分析的基本假设：观察到的数据都是由背后的一个模型产生

数据类型	模型
表格	有监督学习模型
点集	概率分布
时间序列	随机过程（如隐式马氏过程等）
网络	图模型、贝叶斯模型

用科学的方法研究数据

- 数据分析的主要困难
  - 数据量大
  - 维数高（核心困难）
    - 维数灾难：模型复杂度和计算量随着维数的增加而指数增长
    - 如何克服？
      - 将模型限制在一个极小的特殊类里面，如线性模型。
      - 利用数据可能有的特殊结构(例如稀疏性，低维或低秩，光滑性等)，通过正则化和降维来实现。
  - 类型复杂：表格、图像、文本、视频
  - 噪音大：数据在生成、采集、传输和处理等流程均可能引入噪音

用科学的方法研究数据

- 算法的重要性
  - 与模型相辅相成的是算法，以及算法在计算机上实现
  - 从算法角度看，处理大数据有两条思路
    - 降低算法的复杂度：
      - 如随机梯度下降等
    - 分布式计算：
      - 把大问题分解成小问题，然后分而治之. 如著名的MapReduce框架

算法研究的现状

- 算法的研究被分散在两个基本不相往来的领域
  - 计算数学：研究的算法主要针对函数等连续结构。应用于微分方程等。
  - 计算机科学：主要处理的是离散结构，如文本、网络。
- 现实的数据介于两者之间：数据本身是离散的，而数据背后有一个连续的模型
- 要发展针对数据的算法，必须把计算数学和计算机科学的算法有机结合

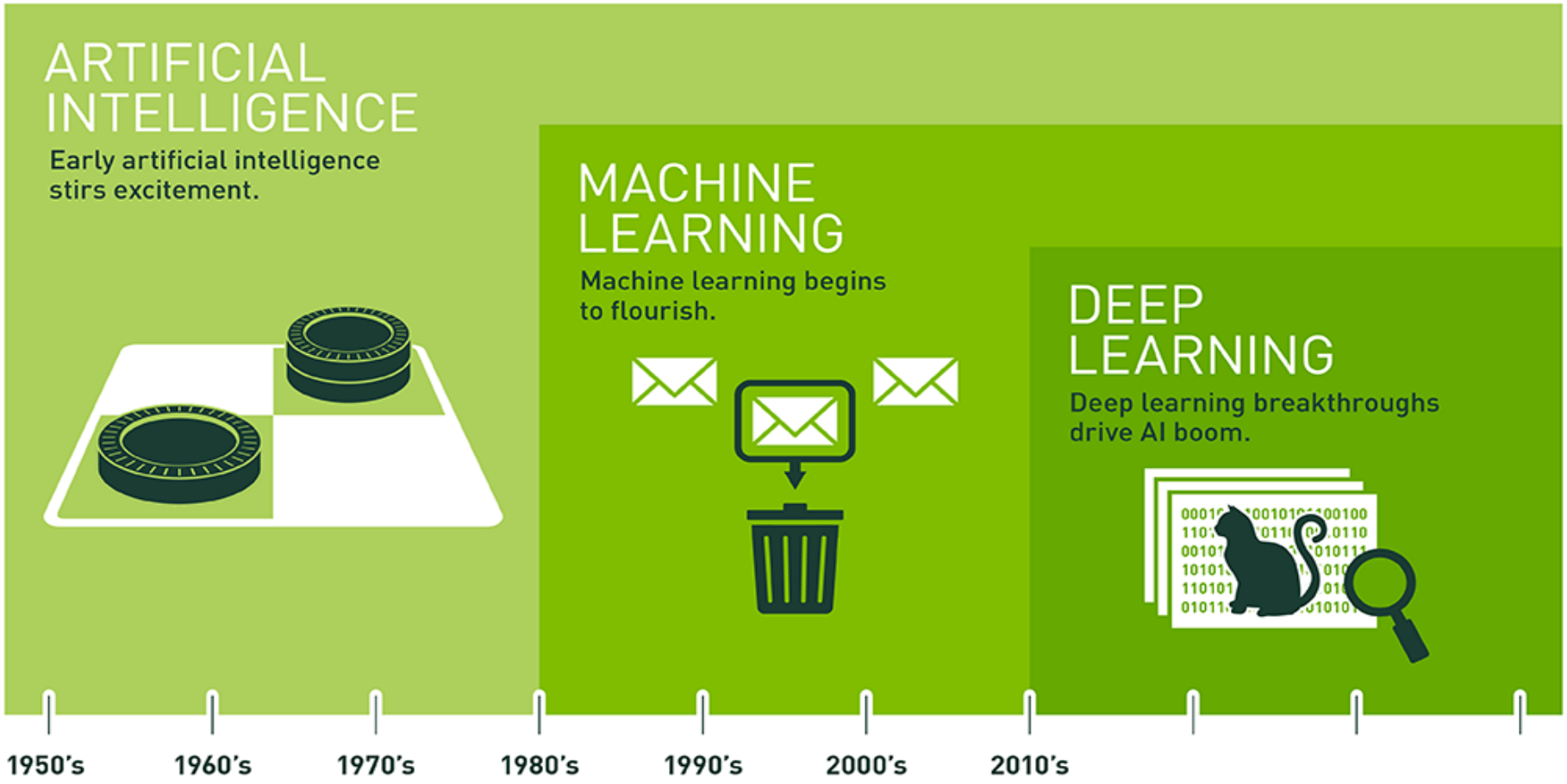
目录

1. 机器学习

- 机器学习有下面几种定义：
  - 机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。
  - 机器学习是对能通过经验自动改进的计算机算法的研究。
  - 机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。

> We define *machine learning* as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).
>
> — 《Machine Learning: A probabilistic perspective》

人工智能、机器学习、深度学习



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

By: Open Data Science on Twitter: "What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?

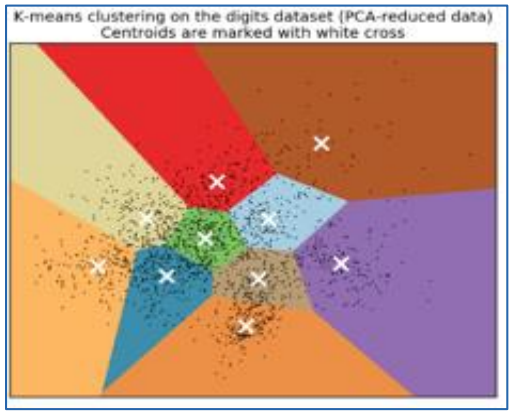
机器学习的问题：概述

博雅大数据学院
BOYA BIG DATA COLLEGE

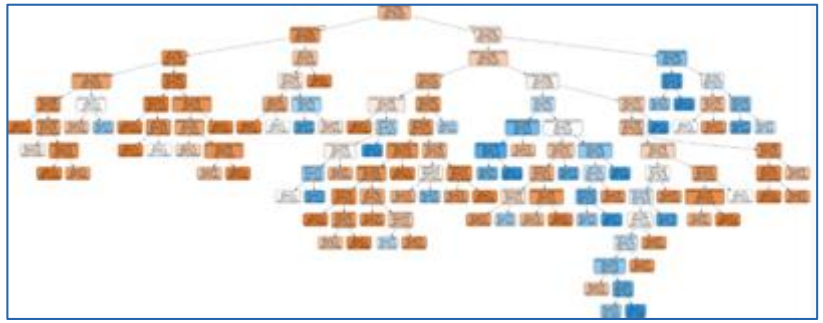
数据

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
14	8	454	220	4354	9	70	1	chevrolet impala

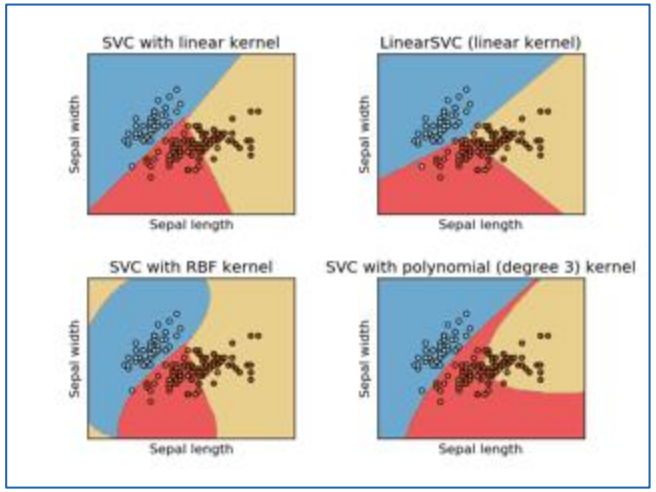
聚类



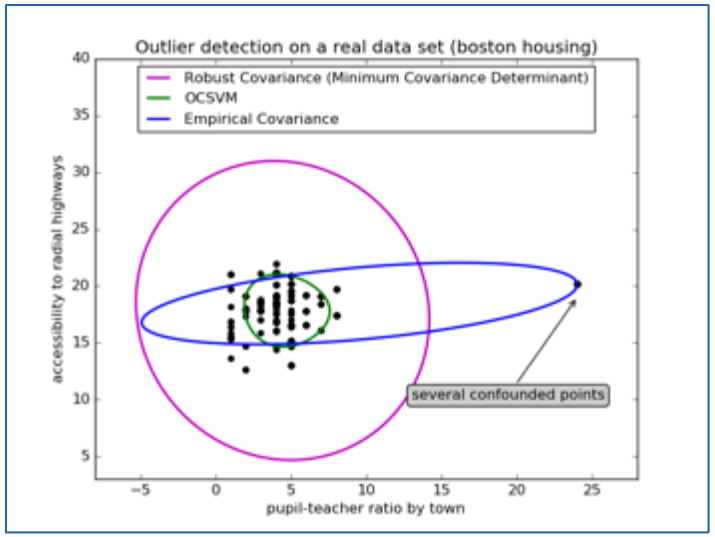
预测



分类



异常值检测



常见机器学习的分析过程

- 基本概念
  - 训练集：用于训练模型的数据集
  - 测试集：用于测试模型的数据集
  - 模型：建立数据的输入$\boldsymbol{x}$和输出y之间的映射关系 $y = f(x)$
  - 损失函数：$L(f(\boldsymbol{x}_i), y_i) = (f(\boldsymbol{x}_i) - y_i)^2$
  - 优化目标：

$$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i))$$

机器学习的基本方法

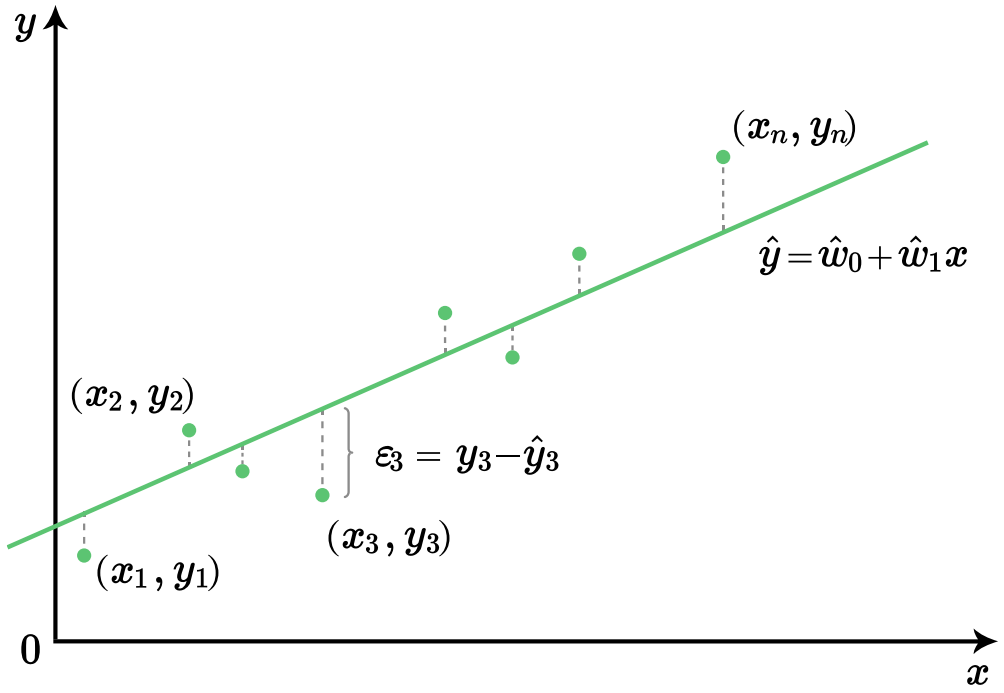
- 有监督学习（supervised learning）
  - 数据集中的样本带有标签，有明确目标
  - 回归和分类
- 无监督学习（unsupervised learning）
  - 数据集中的样本没有标签，没有明确目标
  - 聚类、降维、密度估计、关联规则挖掘
- 强化学习（reinforcement learning）
  - 智慧决策的过程，通过过程模拟和观察来不断学习、提高决策能力
  - 例如：AlphaGo（蒙特卡洛树搜索）

有监督学习

- 数据集中的样本带有标签
- 目标：找到样本到标签的最佳映射
- 典型方法
  - 回归模型：线性回归、岭回归、LASSO和回归样条等
  - 分类模型：逻辑回归、K近邻、决策树、支持向量机等

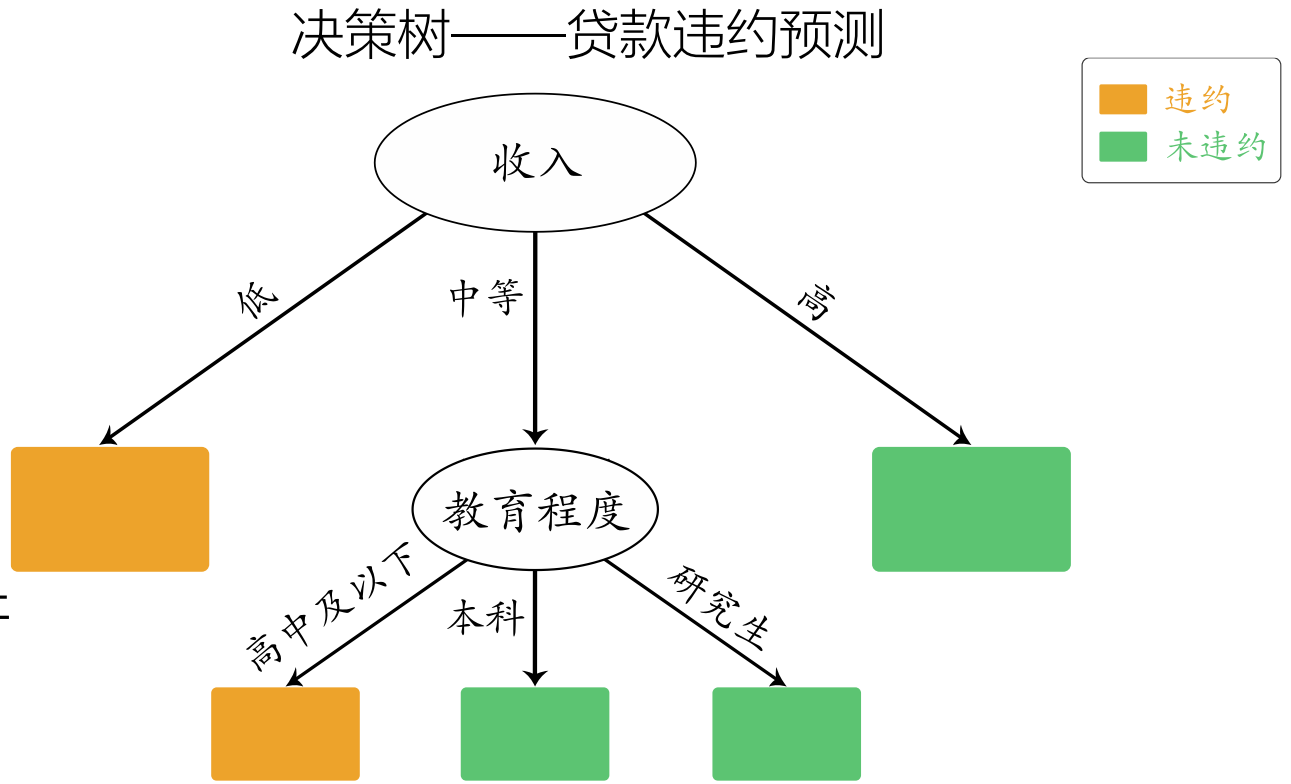
回归

- 典型的有监督任务，样本的标签为连续型，如收入、销量等
- 应用场景：
  - 流行病学：吸烟对死亡率和发病率影响的早期证据来自采用了回归分析的观察性证据
  - 金融：资本资产定价模型利用线性回归以及Beta系数的概念分析和计算投资的系统风险
  - 经济学：预测消费支出，固定投资支出，存货投资，一国出口产品购买，劳动力需求，劳动力供给



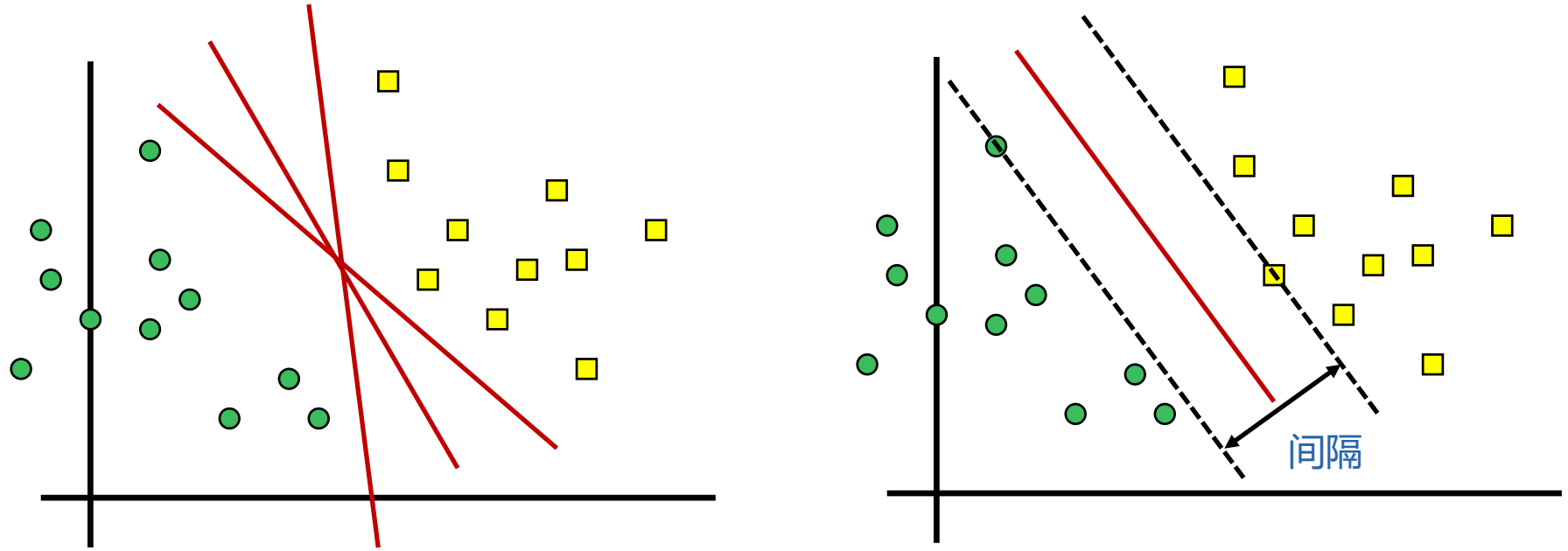
分类

- 典型的有监督学习任务，样本标签为离散型。包括二分类和多分类问题。
- 应用场景：
  - 信用风险评估
  - 预测肿瘤细胞是良性还是恶性
  - 邮件的分类：正常邮件/垃圾邮件
  - 客户流失预测



分类：支持向量机

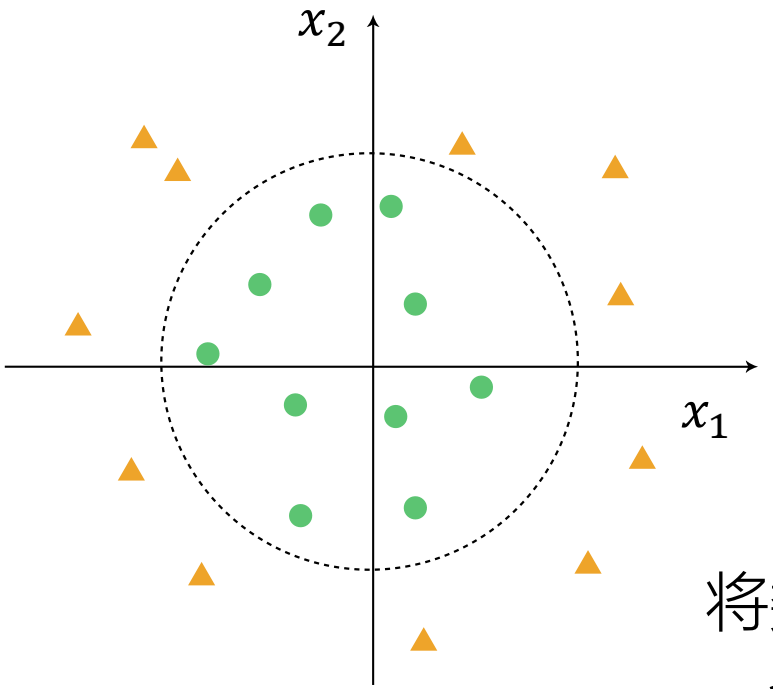
线性可分数据



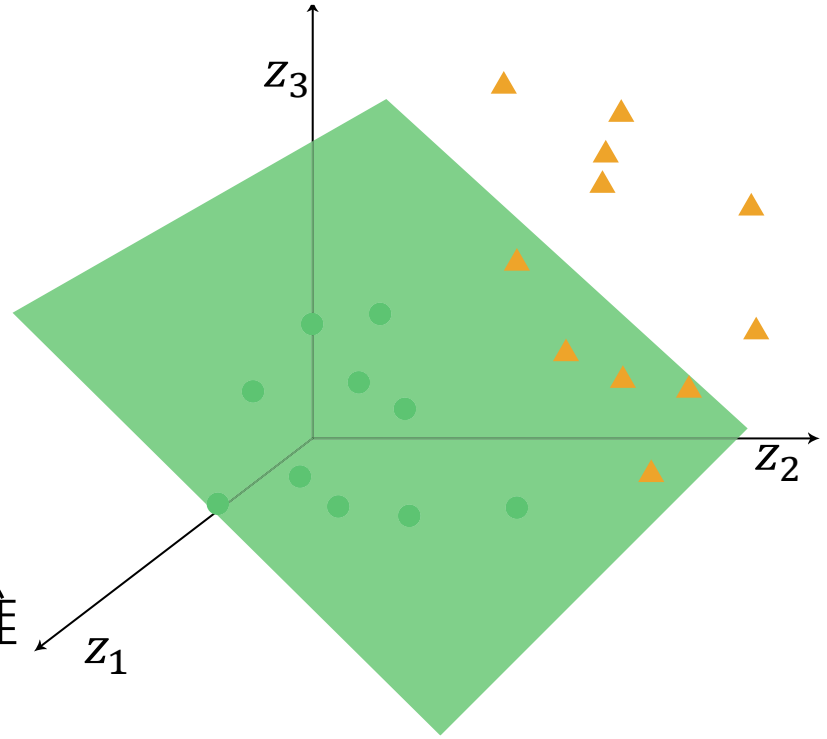
哪一条分割线更好？ ⟹ 具有最大间隔的分割线是最好的！

分类：支持向量机

如果数据并不是线性可分的?



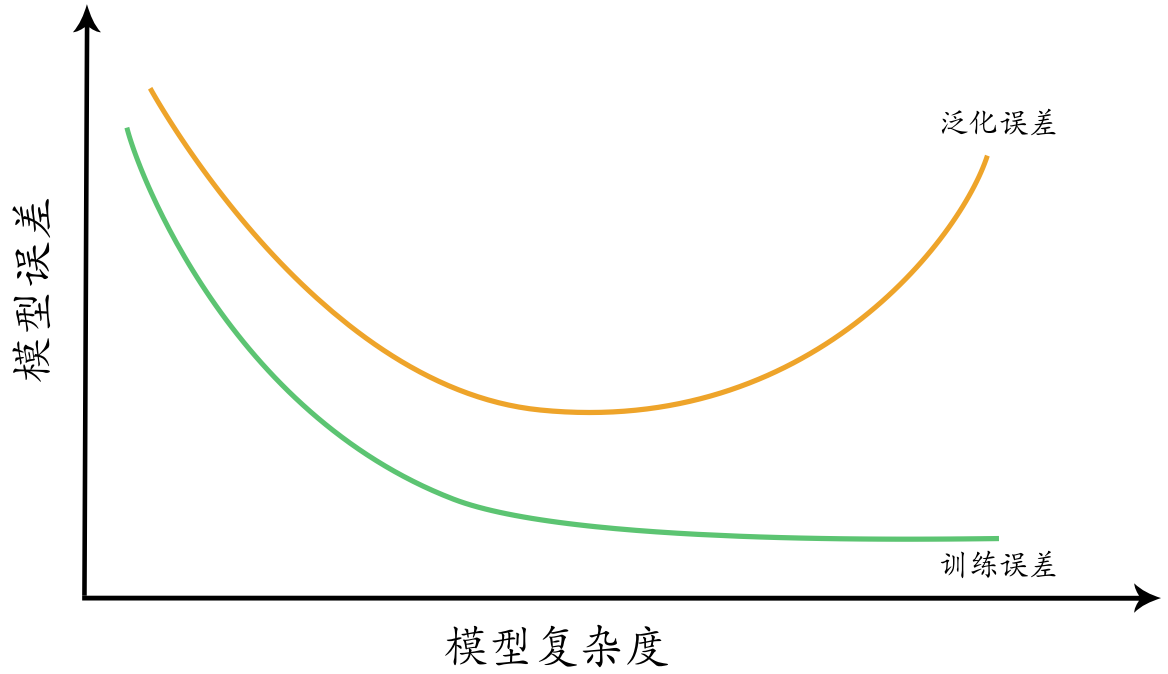
映射trick!



将数据点从2维空间映射到3维空间中，使得数据线性可分

过度拟合问题

- 模型过于复杂(例如参数过多)，导致所选模型对已知数据预测得很好，但对未知数据预测很差。

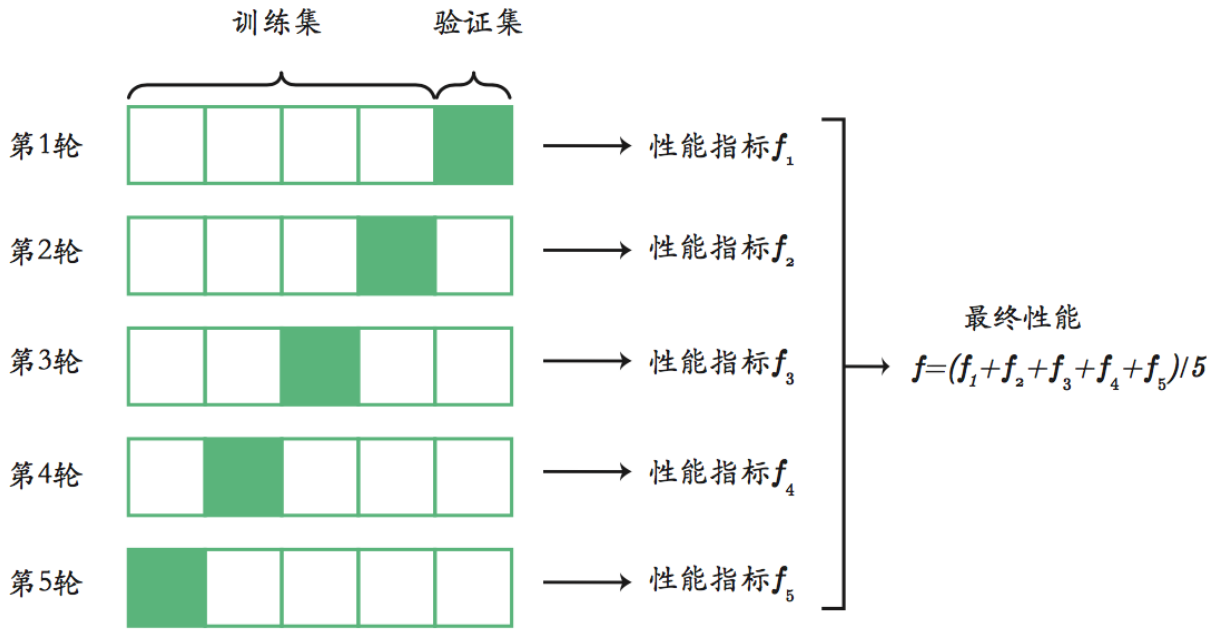


- 正则化：$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda J(f)$

- 正则化是模型选择的典型方法
- 在误差函数上加一个正则项，正则项通常为参数向量的范数
- 在训练误差和模型复杂度之间的权衡

模型选择

- 交叉验证：基本想法是重复地使用数据。将数据集随机切分，将切分的数据集组合为训练集和测试集，在此基础上反复进行训练，测试和模型选择。
- K折交叉验证（k-fold cross validation）
  - 随机地将数据切分为k个互不相同大小相同的子集；
  - 每次利用$k-1$个子集的数据训练模型，余下的数据测试模型；
  - 最后选择在k次测评中平均性能最好的模型。



无监督学习

- 可以处理没有标签的数据
- 根据数据本身的分布特点，挖掘反映数据的内在特性
- 典型方法
  - 聚类、降维、关联规则挖掘等

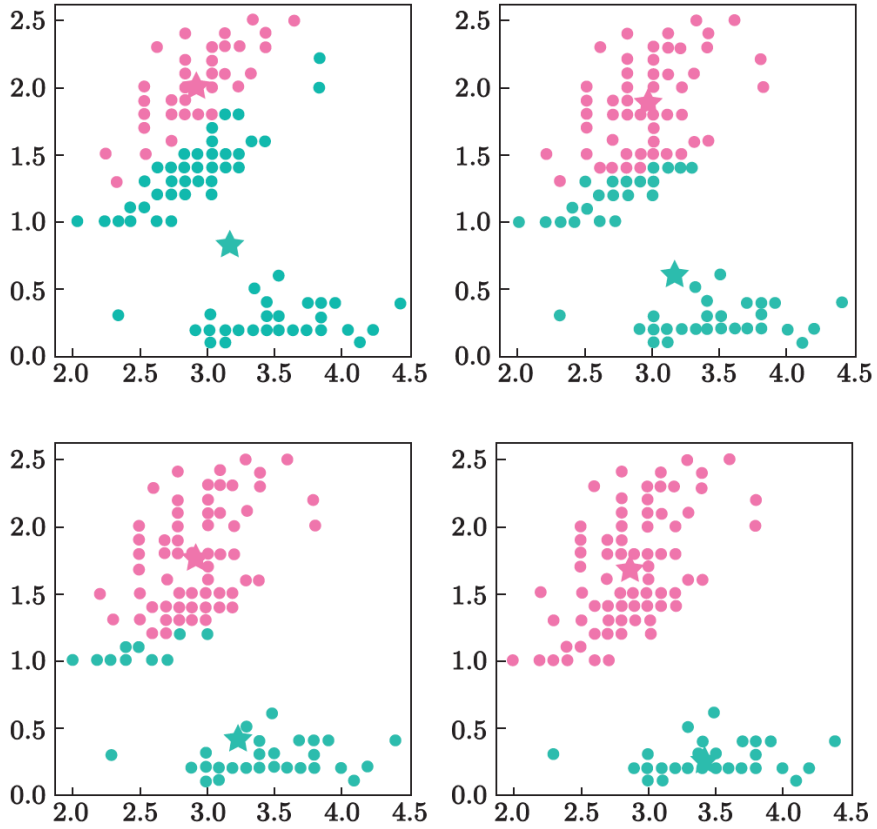
聚类

- 目的：将数据集中相似的样本进行分组，使得：
  - 同一组对象之间尽可能相似；
  - 不同组对象之间尽可能不相似。
- 应用场景：
  - 基因表达水平聚类：根据不同基因表达的时序特征进行聚类，得到基因表达处于信号通路上游还是下游的信息
  - 篮球运动员划分：根据球员相关数据，将其划分到不同类型（或者不同等级）的运动员阵营中
  - 客户分析：把客户细分成不同客户群，每个客户群有相似行为，做到精准营销

无监督学习：例子

K-Means聚类

- 1.选择K个点作为初始质心
- 2. Repeat：
  - 将每个点指派到最近的质心，形成K个簇
  - 重新计算每个簇的质心
- 3. 直到质心不发生变化



2. 关联规则挖掘

- 目的：分析特征之间的关联关系 。
- 应用场景：
  - 购物分析：用于促销、货架管理和存货管理
  - 气象预测：基于关联规则对灾害天气的预测
  - 医疗信息：发现与某种疾病关联的并发症状
  - 推荐系统：找出商品之间的购买关系，从而进行商品推荐

年轻的父亲的购物篮子
（啤酒-尿不湿案例）

TID	项集
1	{面包，牛奶}
2	{面包，尿布，啤酒，鸡蛋}
3	{牛奶，尿布，啤酒，可乐}
4	{面包，牛奶，尿布，啤酒}
5	{面包，牛奶，尿布，可乐}

3. 自然语言处理和文本分析

- 从非结构化的文本数据中提取有用的信息和知识
- 主要问题：分词与词性标注、命名实体识别、句法分析、语义消歧、文本分类和聚类、和情感分析等
- 应用场景：

  1.互联网舆情分析。商品评论：好评/差评；投诉数据分析、法院判决文本

  2. 新闻分类、摘要。新闻类别分类：体育、社会、法制、经济等

  3. 机器翻译。不同语言之间的翻译：中英翻译等

商品评论分析

只是我的手机屏幕上有一小块刮痕，不过平时也不太会注意到，就算了，懒得再申请换货了。

手机不错的，除了一个屎黄色的手机套我不喜欢，其他一切都很完美；用了几天，感觉挺好的

试了几天再来评价，，功能强大，不论是上网还是游戏运行速度都很快，屏幕也很清晰。手感很不错

超好用，N个赞！

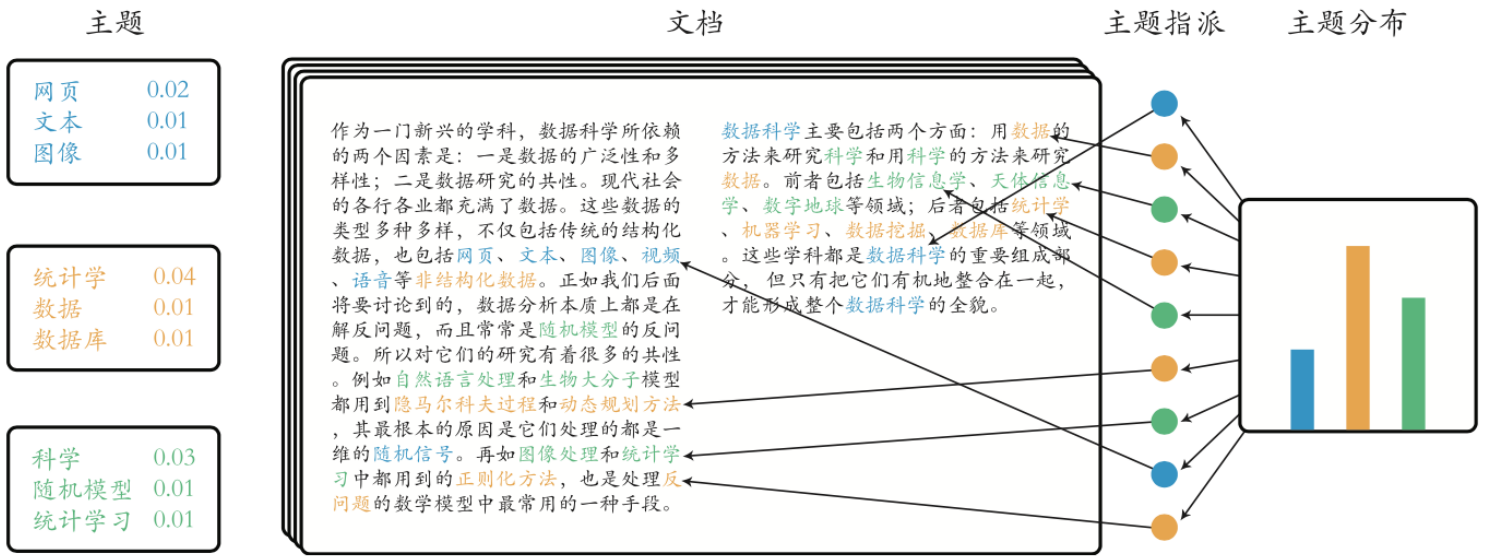
手机挺好的，刚打开用了一天 没有啥毛病，然而我说了能不能给绿色的壳，客服说他们是随机发的，发一个一样颜色的壳这么难吗！！！ 顺丰4天，就这样了.....

说什么呢，高大上。颜值爆表，美的无法形容。使用了几天，很流畅，舒服，内存占用有点高，不过不影响，依然很快，无卡顿。非常满意！稍后上图

评价对象	对象描述词	观点表达式
速度	速度、反应速度	速度快、速度好、速度太慢、速度够快、速度惊人、速度良好、速度真快、
性能	系统、性能、功能	系统流畅、基本流畅、系统稳定、系统简洁、系统快、系统简单、性能优异、性能强大、性能强劲、功能强悍、机子流畅
物流	物流、物流业、发货	物流不错、物流快、物流蛮快、物流很棒、物流超快、物流慢、物流太慢、发货快、发货迅速、发货太慢
客服	客服、服务态度、态度	客服不错、客服小气、客服耐心、客服完美、服务态度差、服务态度很棒、服务态度蛮好、
颜色	颜色	颜色好、颜色漂亮、颜色不错、颜色较暗
屏幕	屏幕、画面、像素	屏幕碎、画面清晰、屏幕清晰、屏幕不灵、屏幕细腻、屏幕亮丽、屏幕小、屏幕很大、屏幕窄、像素好、
手感	手感、质感	手感不错、质感不错、手感好、手感细腻、手感极佳

主题/情感分析

- 主题分析
  - 挖掘海量文本集合中的主题
  - 分析单个文本的主题分布
  - 将文本从词典（十万级）降维到主题（几百）
- 情感分析
  - 政治选举
  - 股票市场
  - 舆情事件
  - 电影票房
  - 用户心情



4. 图与社交网络分析

- 主要问题
  - 中心度、链接分析、社区发现、影响力分析等
- 应用场景：
  - 1. 专家/网页/用户重要度评估

    根据网络结构评估节点重要性，节点：专家、用户、网页等
  - 2. 舆论领袖挖掘：根据信息传播网络，发现舆论领袖、关键人物
  - 3. 欺诈团伙检测

    电信欺诈、交易欺诈、信用卡申请欺诈

5. 分布式计算

- 如何对大规模数据进行处理和分析
- 主要问题：
  - 单机环境下大数据数据处理
  - 集群环境下的大规模数据处理
  - 大规模数据下的建模分析（分布式机器学习）
- 应用场景：
  - 1.大规模数据处理
    并行计算、Hadoop/MapReduce平台
  - 2. 大规模数据下的模型构建
    并行算法、硬件加速（GPU和深度学习）、Spark等分布式架构
  - 3. 算法的并行化、数据的并行化

目录

大数据行业应用

西班牙电信：数据变现

Smart Steps

2012年成立大数据部门：Telefonica Dynamic Insights。推出了名为“Smart Steps”的产品，通过脱敏的用户位置数据，可以对某个时段、某个地点人流量的关键影响因素进行分析。“Smart Steps”为零售商新店设计和选址、设计促销方式、与客户反馈等提供决策支撑，从而帮助零售商更好地理解和满足客户需求、降低成本；也可帮助市政委员会统计、预测各种场景下的人流量。

运输模型

利用实时数据，预测人流量，出行模式以及分析交通网络的变化。

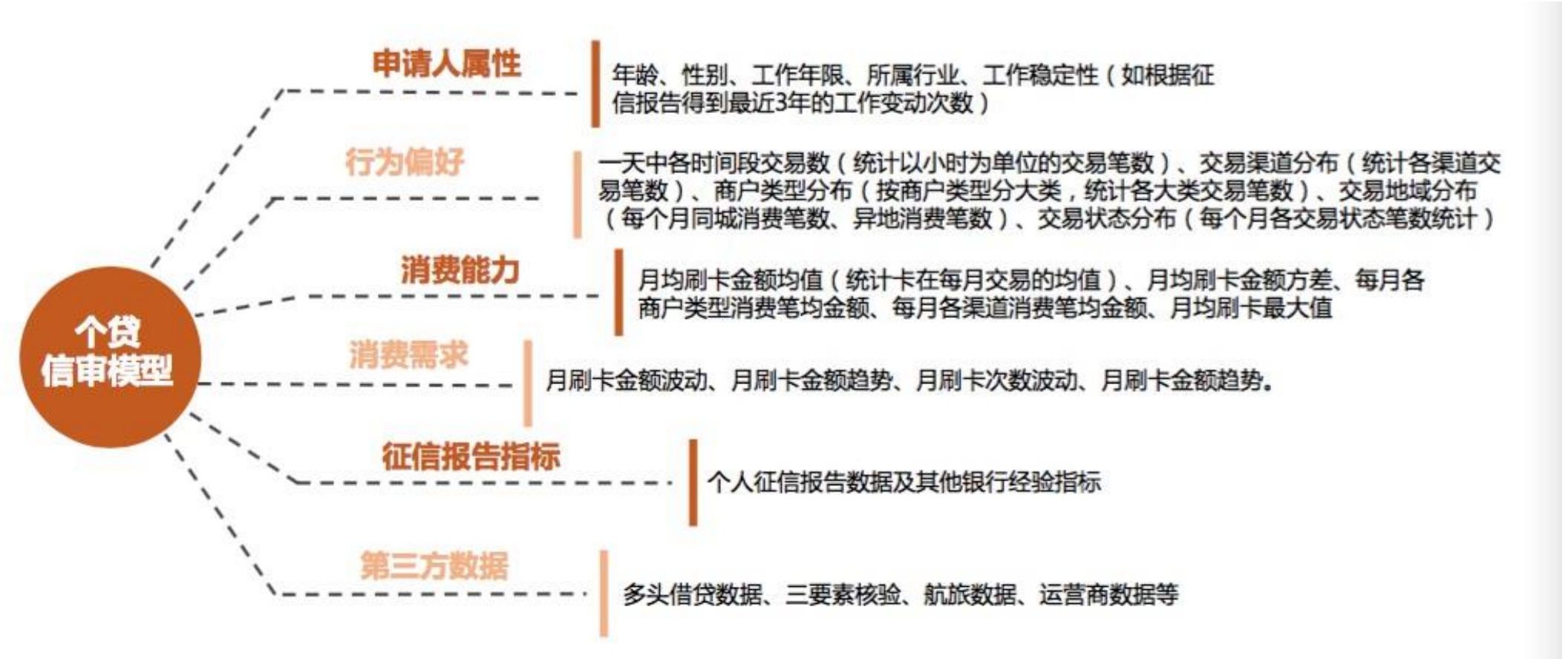
交通建设规划

提供旅途时间，起始、目的地的人流热力图等。

衡量经济发展

通过人流量辅助分析经济运行状况。

消费金融：用户画像和信贷评估

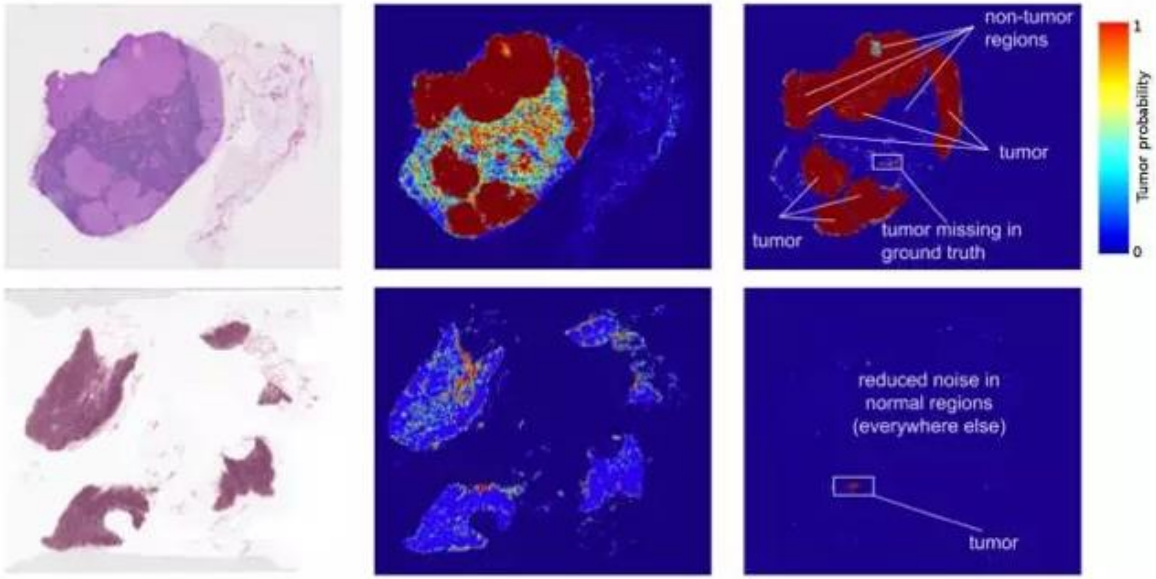


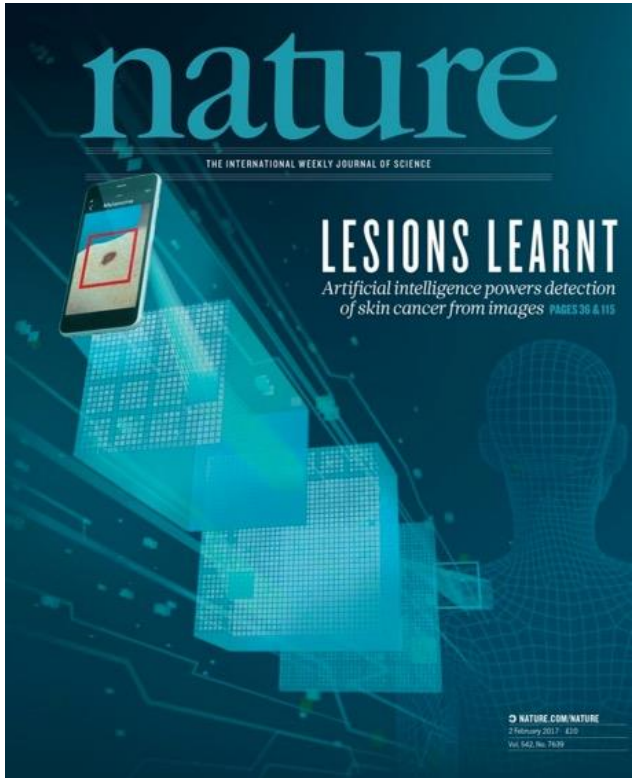
模型评估后决定是否给予信贷

普林科技：开源金融银行卡信用评估

健康医疗：皮肤癌诊断

- 大量患者切片的数据，以及何处病患的标记，训练集充足
- 切片一般都是高清晰度的，一张切片有上千万甚至上亿像素，不便直接训练，专家们将照片切割成了128*128像素的标准大小





Andre Esteva, et al. "Dermatologist-level classification of skin cancer with deep neural networks." Nature (2017)

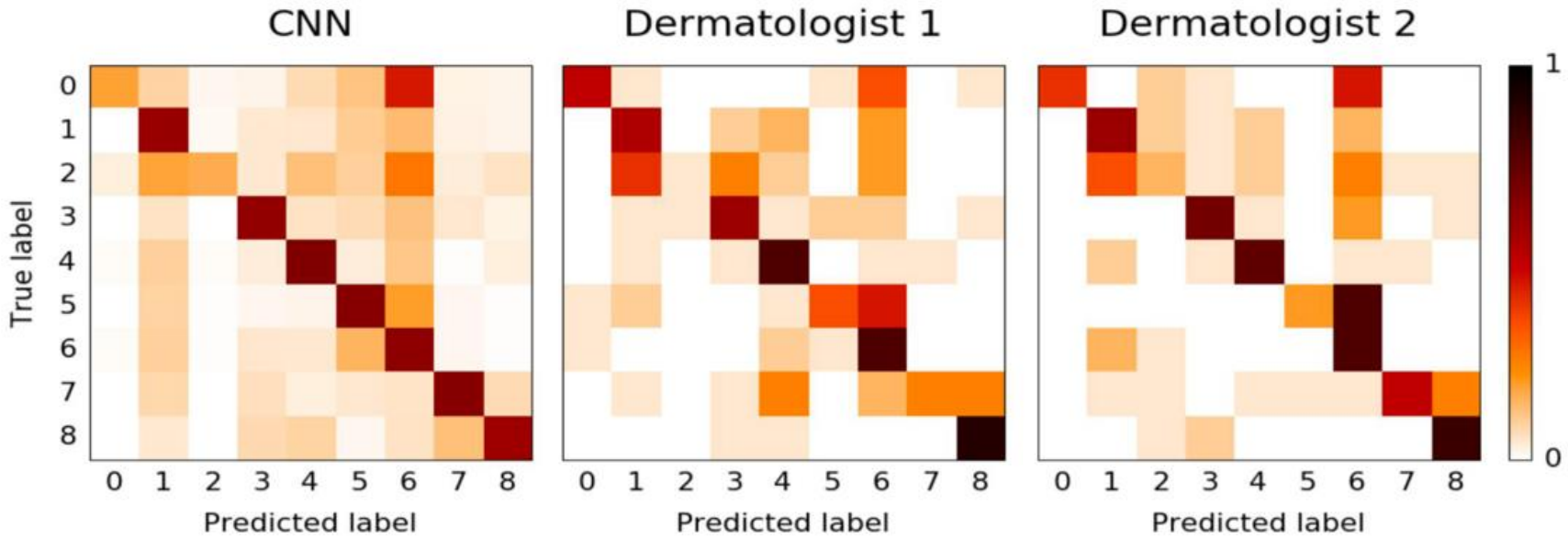
健康医疗：皮肤癌诊断

深度学习: 129,450 临床图片, 757 种疾病

分类准确率: 72.1%

人类病理学家准确率：66%左右。

基于CNN的方法诊断准确率超过了人类病理学家的平均准确率。



总结

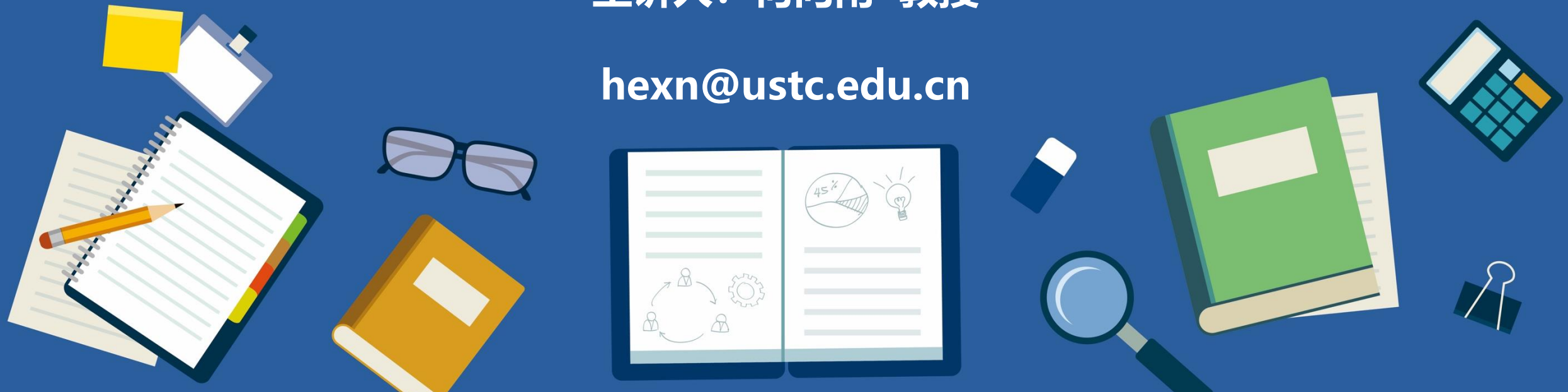
- 数据科学概述
- 数据科学的基本内容
  - 机器学习
  - 关联规则挖掘
  - 自然语言处理
  - 图和社交网络分析
  - 分布式计算
- 大数据行业应用

博雅大数据学院
BOYA BIG DATA COLLEGE

数据预处理

主讲人：何向南 教授

hexn@ustc.edu.cn

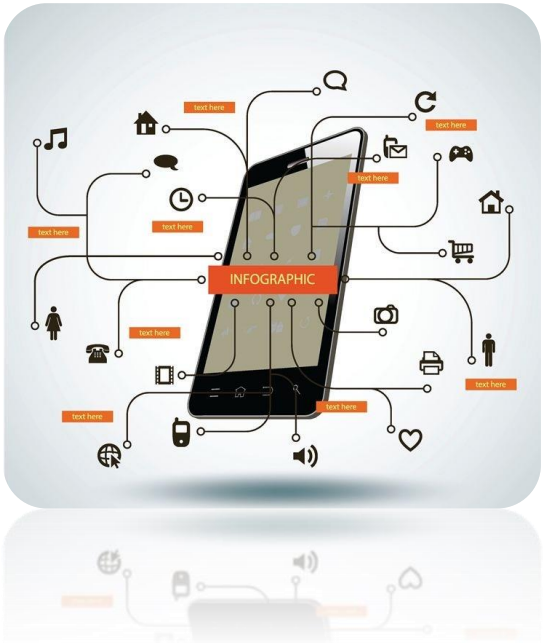


目录

数据类型

- 表格数据
  - 关系记录
  - 数据矩阵
  - 向量
  - 事物数据
- 图和网络
  - 万维网
  - 社交网络
  - 分子结构
- 多媒体数据
  - 文本
  - 图像
  - 视频
  - 音频

行星	周期(年)	平均距离	周期^2/距离^3
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00



样本

- 数据集由样本构成、 一个数据对象表示一个实体
- 相同概念
  - 样本（samples or examples）
  - 实例（instances）
  - 数据点（data points）
  - 对象（objects）
  - 元组（tuples）
- 数据对象由属性（attributes）及其值（value）构成
- 数据库：行(row)为数据对象；列(column)为属性

特征

- 表征样本某个特征的数据域
- 相同概念
- 维（dimension）、变量（variable）、属性（attribute）、字段
- 例如在汽车性能指数：排气量、车重、油耗

连续型 vs 离散型

- 连续型特征
  - 特征可以为实数空间任意取值
  - 例如：温度、身高、价格、时间等
  - 通常由浮点型表示
- 离散型特征
  - 其值域为有限集或可列集
  - 例如：性别、汽车品牌、NBA球队等
  - 布尔型、等级型

数据的基本统计信息

- 获得数据总体印象，更好的理解数据
- 主要内容：度量数据的中心趋势和离散程度、描述数据汇总的图形显示（数据可视化）
- 度量手段：
  - 算术平均值（Mean）
  - 中值（Median）
  - 最大值（Max）
  - 最小值（Min）
  - 分位数（Quantiles）
  - 方差（Variance）

度量集中趋势（Central Tendency）

- 均值（Mean）

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ 或 } \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- 中位数（Median）

  当特征值的项数n为奇数时，处于中间位置的特征值即为中位数；当n为偶数时，中位数则为处于中间位置的2个特征值的平均数。

- 众数（Mode）

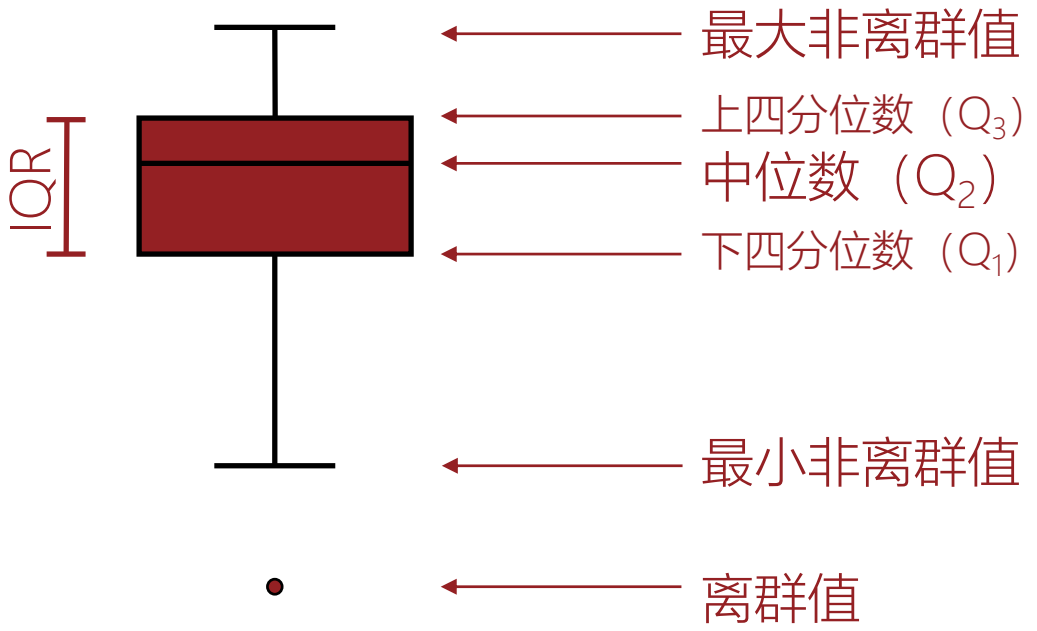
  出现频率最高的值

度量数据散布度（Dispersion）

- 分位数（Quantile）
  - K分位数：令x是一个值，如果在数据集中，百分之K的数据的值都不大于x，则称x为该数据集的K分位数。
  - 分位数：Q1（25th percentile），Q3（75th percentile）
  - 中间分位数范围：IQR=Q3-Q1
  - 方差（Variance）和标准差（Standard deviation）
  - 方差：$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

度量数据散布度（Dispersion）

- 盒图（boxplot）
  - 也称箱线图
  - 盒外的两条线（胡须）分别延伸到最小和最大观测值
  - 直观明了地识别数据集中的离群值，并判断数据集的偏态和尾重



距离

- 在很多场景中需要计算样本的距离或相似度
  - 样本是否重复？两个商品是否相似？客户分群？
- 假设$d(\cdot)$为某种距离函数，则通常需要满足以下条件：
  - 距离通常是非负的: $d(x,y) \geq 0$
  - 一个样本与自己的距离为零: $d(x,x)=0$
  - 距离通常满足对称性: $d(x,y)=d(y,x)$
  - 距离的三角不等式: $d(x,z) \leq d(x,y)+d(y,z)$

数值型数据距离

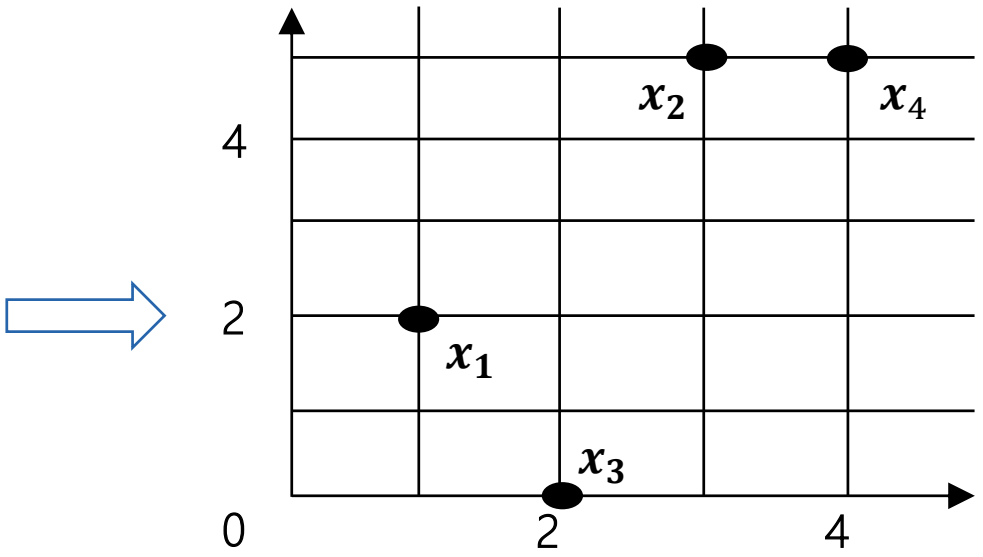
- 闵可夫斯基距离（Minkowski distance）

$$d(i,j)=\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h+\cdots+|x_{id}-x_{jd}|^h}$$

  - $i=(x_{i1}, x_{i2},\cdots, x_{id})$和 $j=(x_{j1}, x_{j2},\cdots, x_{jd})$分别代表两个$d$维的数据对象，$h$ 为序。 上述距离也被称为L_h范式
- 曼哈顿距离（Manhattan distance）
  - $h=1$，L_1 范式 $d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{id}-x_{jd}|$
- 欧式距离（Euclidean distance）
  - $h=2$，L_2 范式 $d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{id}-x_{jd}|^2)}$
- 极大距离（supremum distance）
  - $h=\infty$，L_∞范式 $d(i,j)=\lim_{h\to\infty}\left(\sum_{f=1}^{d}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f=1}^{d}|x_{if}-x_{jf}|$

闵可夫斯基距离示例

点集	特征1	特征2
	1	2
	3	5
	2	0
	4	5



- 曼哈顿距离

L_1				
	0			
	5	0		
	3	6	0	
	6	1	7	0

- 欧式距离

L_2				
	0			
	3.61	0		
	2.24	5.1	0	
	4.24	1	5.39	0

- 极大距离

L				
	0			
	3	0		
	2	5	0	
	3	1	5	0

数值型数据距离

- 余弦相似度（Cosine Similarity）

$$\cos(o_i, o_j) = \frac{\sum_{k=1}^{n}(x_{ik} \cdot x_{jk})}{\sqrt{\sum_{l=1}^{n} x_{il}^2} \cdot \sqrt{\sum_{l=1}^{n} x_{jl}^2}}$$

向量内积

L2 norm

Instance	Team	Coach	Hockey	Baseball	Soccer	penalty	Score	Win	Loss	Season
Instance 1	5	0	3	0	2	0	0	2	0	0
Instance 2	3	0	2	0	1	1	0	1	0	1

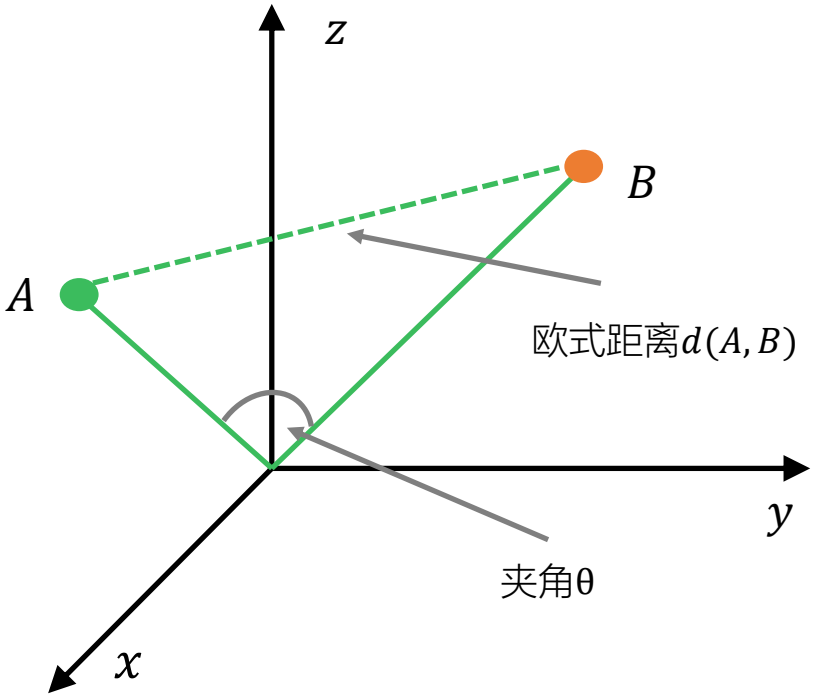
上表中实例 1 和示例 2 的余弦相似度为：

$$\cos(\text{Instance 1}, \text{Instance 2}) = \frac{5*3+0*0+3*2+0*0+2*1+0*1+2*1+0*0+0*1}{(25+9+4+4)^{0.5} * (9+4+1+1+1+1)^{0.5}}$$

$$= 0.94$$

余弦相似度与欧式距离的对比

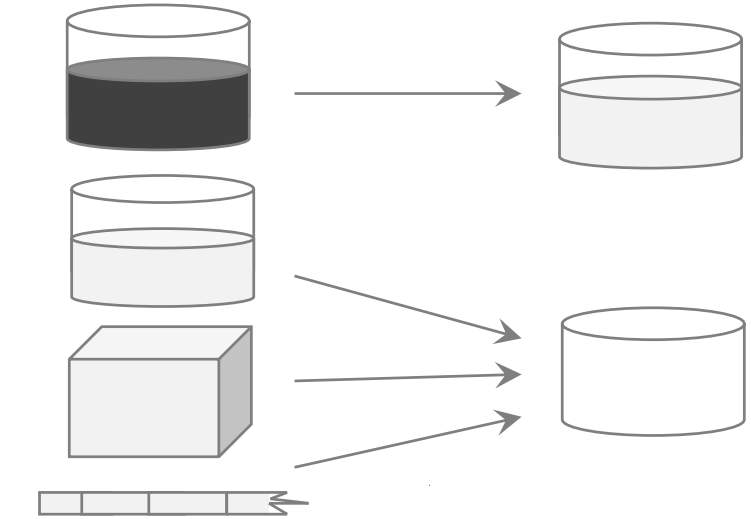
- 衡量角度不同
  - 欧式距离：绝对距离
  - 余弦相似度：方向差异
- 适应模型不同
  - 欧氏距离：数值特征绝对差异，用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异
  - 余弦相似度：对绝对数值不敏感，用于使用用户对内容评分来区分用户兴趣的相似度和差异



数据预处理

- 模型输入数据质量直接影响建模效果
- 在正式构建模型之前往往需要对数据进行恰当的预处理
  - 数据清洗、整合、转换、规约

Garbage in Garbage out



-2,32,100,59,48 → -0.02,0.32,1.00,0.59,0.48

目录

缺失值处理

真实的数据往往因为各种原因存在缺失值：

1）数据采集不全

2）数据丢失

3）数据分析特征设计过程导致的缺失（客户的信用卡欠款额，如果客户没有信用卡）

学生信息数据表

入学年份	性别	年龄	足球	篮球	···	购物	化妆
2012	M	18	0	0	···	0	0
2012	F	18	0	1	···	0	0
2010	M	20	0	1	···	0	0
2012	F	18	0	0	···	0	2
2011	F	18	0	0	···	1	1
2012	F		0	0	···	1	0
2012	F	18	0	0	···	0	0
2011	M	18	2	0	···	0	0
2011	F	19	0	0	···	0	0
2012		18	0	0	···	1	0
2012	F	18	0	0	···	0	0
2011		19	0	1	···	0	0
2012	F	18	0	0	···	0	0
2012	F		0	0	···	0	2
2012	F	18	0	0	···	0	1

删除法

- 删除法通过删除包含缺失值的数据，来得到一个完整的数据子集. 一般从两个角度进行删除：特征 vs 样本
  - 删除特征：当某个特征缺失值较多，且该特征对数据分析的目标影响不大时，　可以将该特征删除
  - 删除样本：删除存在数据缺失的样本。
    该方法适合某些样本有多个特征存在缺失值，且存在缺失值的样本占整个数据集样本数量的比例不高的情形
- 缺点：
  - 它以减少数据来换取信息的完整，丢失了大量隐藏在这些被删除数据中的信息；
  - 在一些实际场景下，数据的采集成本高且缺失值无法避免，删除法可能会造成大量的资源浪费。

均值填补

- 计算该特征中非缺失值的平均值（数值型特征）或众数（类别型特征），然后使用平均值或众数来代替缺失值
- 缺点一：均值填补法会使得数据过分集中在平均值或众数上，导致特征的方差被低估
- 缺点二：由于完全忽略特征之间的相关性，均值填补法会大大弱化特征之间的相关性

随机填补

随机填补是在均值填补的基础上加上随机项，通过增加缺失值的随机性来改善缺失值分布过于集中的缺陷

近似贝叶斯Bootstrap方法

- 假设某特征有n个缺失值，k个正常值
- 从 k 个观测数据集$\{x_1, x_2, \cdots, x_k\}$ 中有放回的抽取 k^*个观测值建立一个新的数据集 x^*
- 对于 n 个缺失值，从数据集 x^*中随机抽取n 个值进行填补

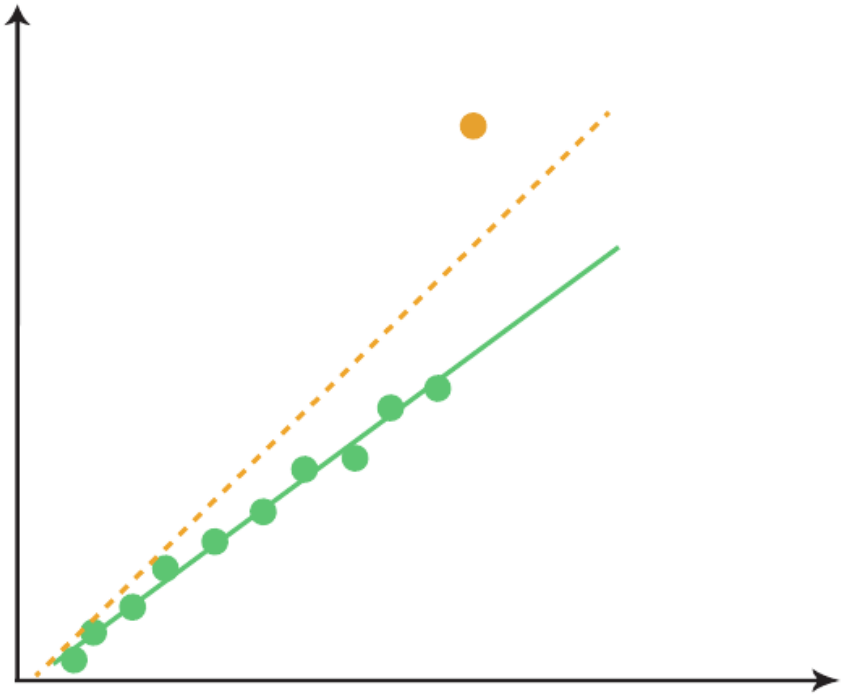
基于模型的填补

- 将缺失特征 y 当作预测目标；
  使用其余特征作为输入，利用缺失特征非缺失样本构建分类或回归模型；
  使用构建的模型预测缺失特征的缺失样本值
- 其他缺失值处理方法
  - 哑变量方法(dummy variable)：对于离散型特征，将缺失值作为一个单独的取值进行处理
  - EM算法填补

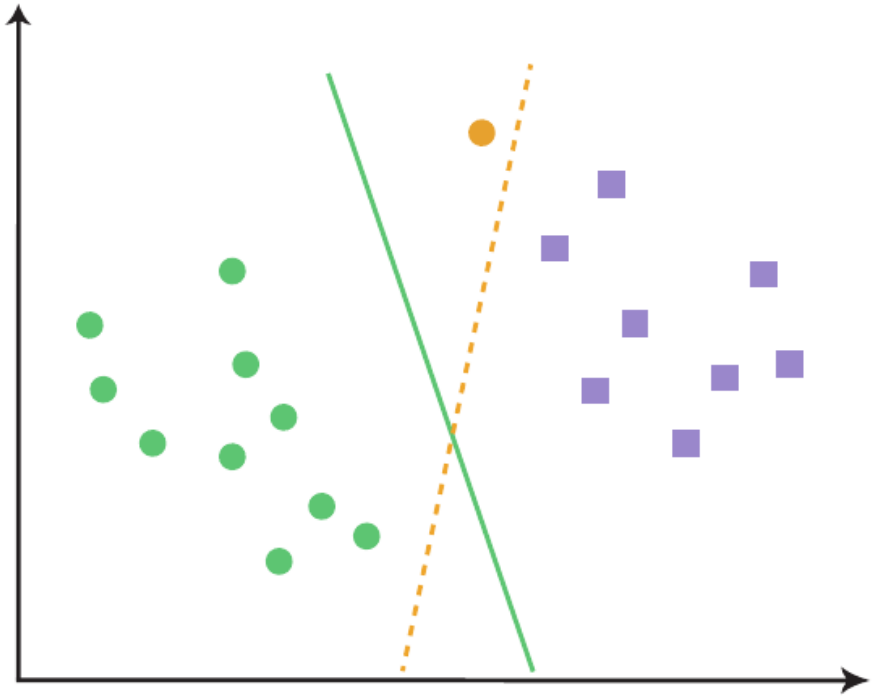
离群值

- Outlier, 指一个数据集中那些明显偏离正常样本的其他样本
  - 一种带有统计学味道的定义是：一个观测与其他观测偏离太多以致于值得怀疑它是由不同的机制所产生的
- 产生原因：自然变异、数据测量、收集误差、人工操作失误等
- 离群值检测可以作为数据预处理的一个步骤，为数据分析提供高质量的数据
- 离群值检测也可以直接用来解决很多应用问题，例如信用欺诈检测、电信欺诈检测、疾病分析和计算机安全诊断等

离群值的影响



回归

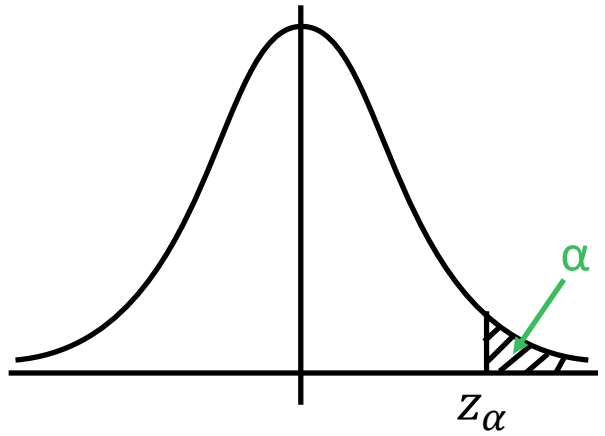


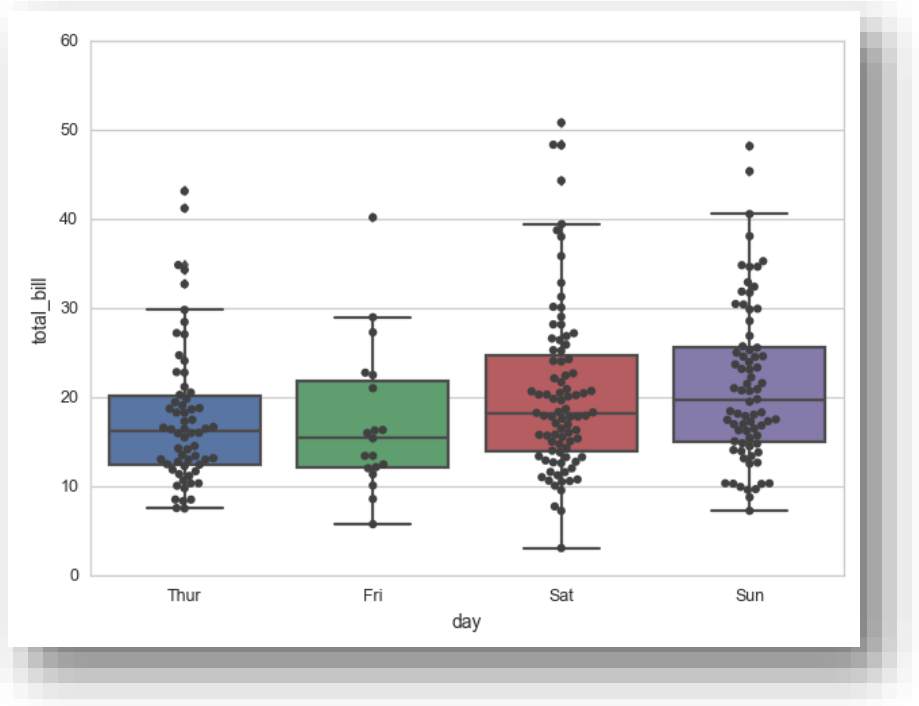
分类

离群值检测方法

1、基于统计的方法

- 在上、下α分位点之外的值认为以异常值
- 盒图观察

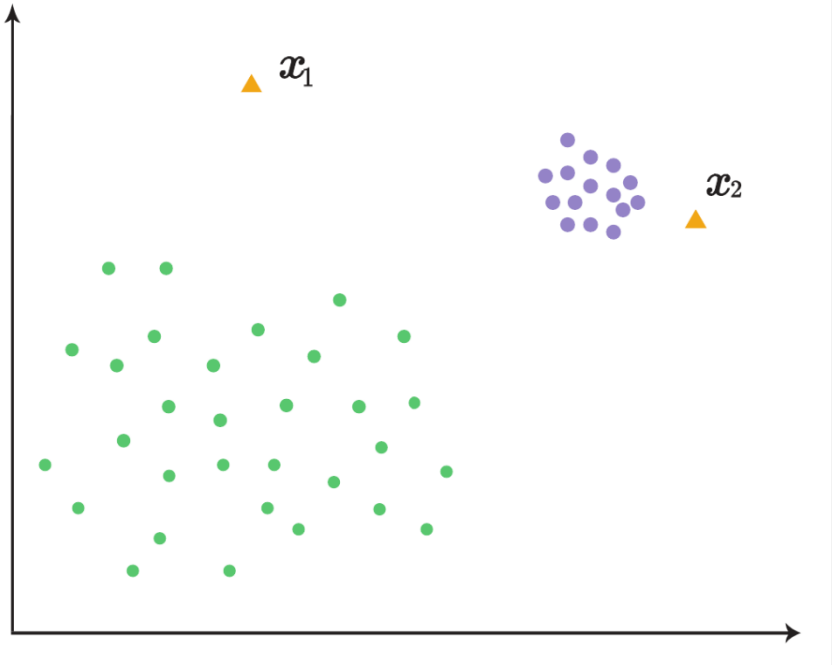




离群值检测方法

2、基于近邻的方法

- 局部异常因子算法（LOF算法，Local Outlier Factor）
- 基本想法：通过比较每个点p和其邻域点的密度来判断该点是否为异常点，如果点p的密度越低，越可能被认定是异常点
- 密度通过点之间的距离来计算，点之间距离越远，密度越低，距离越近，密度越高



离群值检测方法

- 相关定义

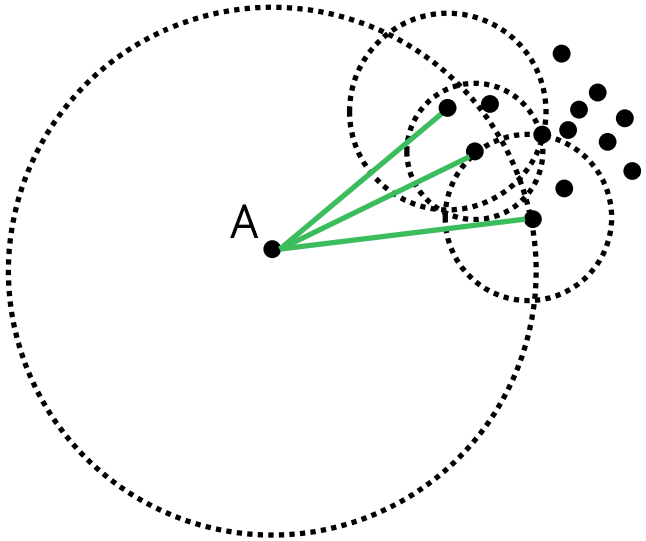
1. $\mathrm{d(A,B)}$：点A与点B之间的距离
2. $\mathrm{d_k(A)}$：点A的第 k 距离，即距离点A第k远的点的距离
3. $\mathrm{N_k(A)}$：点A的第 k 距离领域，即A的第 k 距离以内的所有点，包括第 k 距离的点
4. $\mathrm{rd_k(B,A)}$：点A到点B的第 k 可达距离，计算公式为

$$rd_k(B,A) = max\{d_k(A), d(A,B)\}$$

离群值检测方法

- 相关定义
- 点A局部可达密度（local reachability density）：

$$lrd_k(A) = 1/(\frac{\sum_{o \in N_k(A)} rd_k(A,O)}{|N_k(A)|})$$



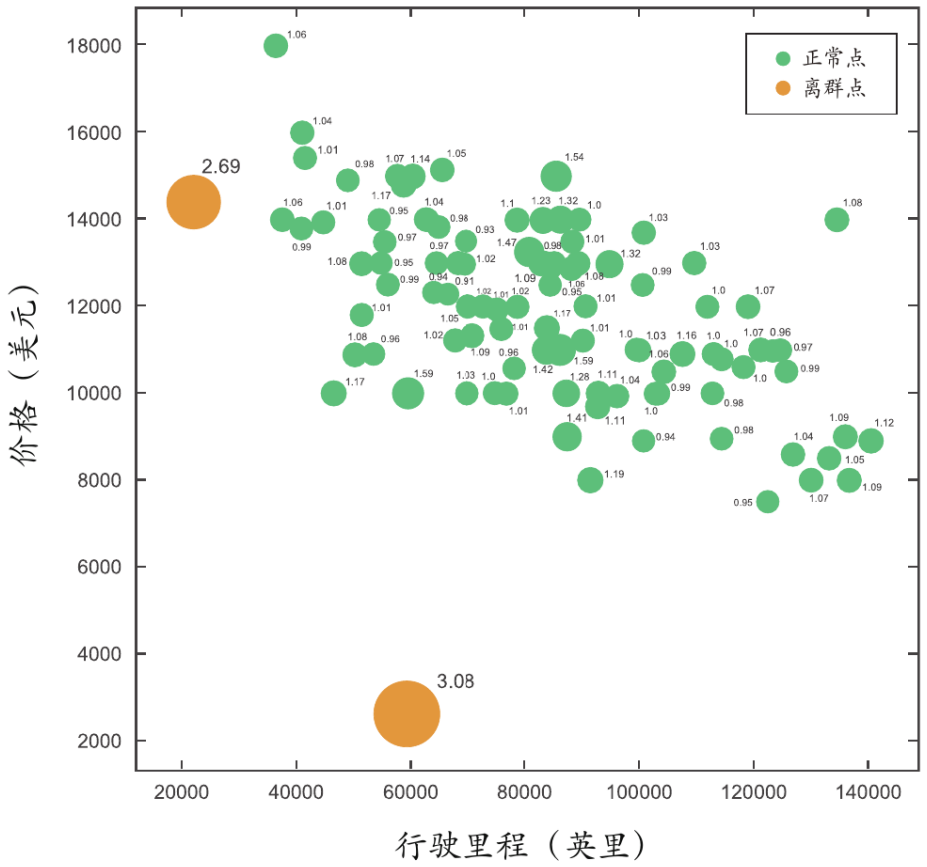
- 表示点A的第 k 领域内点到点A的平均可达距离的倒数
- 局部离群因子（local outlier factor）：

$$lof_k(A) = \frac{\sum_{o \in N_k(A)} \frac{lrd_k(O)}{lrd_k(A)}}{|N_k(A)|} = \frac{\sum_{oinN_k(A)} lrd_k(O)}{|N_k(A)|} / lrd_k(A)$$

- 表示点A的邻域点 N_k (A)的局部可达密度与点A的局部可达密度之比的平均数

离群值检测方法

- 如果 $\mathrm{lof_k}(A)$值越接近于1，说明点A与其邻域点的密度差不多，点A可能和领域属于同一簇
- 如果$\mathrm{lof_k}(A)$越小于1，说明点A的密度高于其领域点的密度，点A为密集点；
- 如果$\mathrm{lof_k}(A)$越大于1，说明点A的密度小于其领域点的密度，点A越可能是异常点。



目录

特征编码

- 模型输入的特征通常需要是数值型的，所以需要将非数值型特征转换为数值特征
- 如性别、职业、收入水平、国家、汽车使用品牌
- 常见方式：数字编码、One-Hot编码

- 数字编码
  - 原特征 收入水平={贫困，低收入，小康，中等收入，富有}
  - 编码后 收入水平={0，1，2，3，4}
  - 缺点：引入了次序关系

特征编码

- One-Hot编码
  - 将包含K个取值的离散型特征转换成K个二元特征（取值为0或1）
  - 优点：
    - 没有引入次序关系：经过编码之后，不同的原始特征之间拥有相同的距离；
    - 对包含离散型特征的回归模型及分类模型的效果有很好的提升
  - 缺点：
    特征数量显著增多，且增加了特征之间的相关性：

$$(f_1 + f_2 + f_3 + f_5 = 1)$$

原始特征取值	f_1	f_2	f_3	f_4	f_5
路虎	1	0	0	0	0
吉利	0	1	0	0	0
奥迪	0	0	1	0	0
大众	0	0	0	1	0
奔驰	0	0	0	0	1

数据标准化（Normalization）

- 为什么要进行数据标准化？
  - 如果一个特征方差远大于其它特征的方差，该特征将会是影响目标函数的主要因素，使算法模型难以学习到其它特征对结果的影响
  - 数据分析及建模过程中，许多机器学习算法需要其输入特征为标准化形式。例如，SVM算法中的RBF核函数，线性模型中的L1、L2正则项，目标函数往往假设其特征均值在0附近且方差齐次；
  - 若样本的特征之间的量纲差异太大，样本之间相似度评估结果将存在偏差
- 常见数据标准化方法：Z-score标准化、Min-Max标准化、小数定标标准化和Logistic标准化

数据标准化

- Z-Score标准化
  - 对特征取值中的每一个数据点作减去均值并除以标准化的操作，使得处理后的数据具有固定均值和标准差，处理函数为：

$$f_i' = \frac{f_i - \mu}{\sigma}$$

  - 其中，f_i'为标准化后各数据点的取值，f_i'为原始各数据点取值，μ为该特征取值的平均值，σ为该特征取值的标准差

  - 适用范围： Z-Score的标准化方法适用于特征的最大值或最小值未知、样本分布非常离散的情况

数据标准化

- Min-Max标准化
  - 又称离差标准化或最大-最小值标准化，Min-Max标准化通过对特征作线性变换，使得转换后特征的取值分布在 [0,1] 区间内。其处理函数为：
  - 其中f_{min}为的最小值，f_{max}为的最大值 $f_i' = \frac{f_i - f_{min}}{f_{max} - f_{min}}$
  - 将特征 f 映射到 [a, b] 区间内：$f_i' = \frac{b-a}{f_{max}-f_{min}}(f_i - f_{min}) + a$
  - 适用范围： 0-1标准化适用于需要将数据简单地变换映射到某一区间中。
  - 缺点：
    - 当有新数据加入时，可能会导致特征的最大值或最小值发生变化，此时便需要重新定义最大值、最小值；
    - 若数据存在离群值，标准化后的效果较差。

数据标准化

- 小数定标标准化
  - 通过移动数据的小数点位置来进行标准化。具体标准化过程中，小数点移动多少位取决于的最大绝对值大小。其处理函数为：

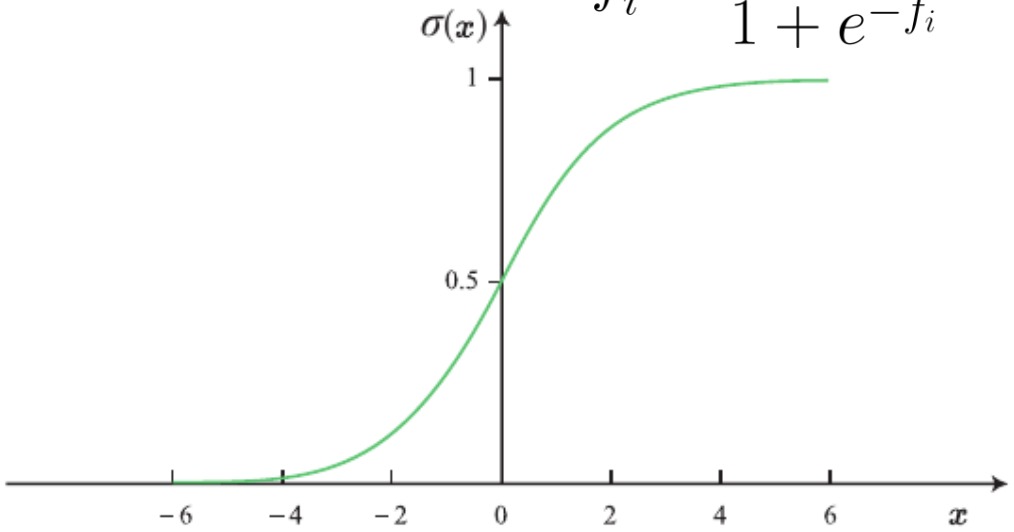
$$f_i^* = \frac{f_i}{10^j}$$

  - 其中 j 是满足条件的 $\max\{f_1', f_2', \ldots, f_n'\} < 1$ 最小整数
    - 如范围为[−3075,2187]，j=4
    - 适用范围：适用于比较分散，尤其是遍布多个数量级的情况，简单实用。
    - 缺点：如果特征取值集中在某几个数量及上，则标准化后的值集中在某几个值附近，不利于后续数据分析时的样本区分；易受到离群值影响。

数据标准化

- Logistic标准化
  - Logistic标准化利用Logistic函数的特性，将映射到 [0,1] 区间内.
  - Logistic函数： $f_i' = \frac{1}{1+e^{-f_i}}$
  - 适用范围：特征取值分布相对比较集中地分布于 0 两侧的情况
  - 缺点：如果特征值分散且远离0，则标准化后的特征值会聚集于 0 或 1 附近，造成原始特征的分布及取值间关系被改变.
  - 因此在应用Logistic标准化方法之前，需要首先分析原始特征取值的分布状况 (可设置temperature)

$$f_i' = \frac{1}{1+e^{-f_i}}$$

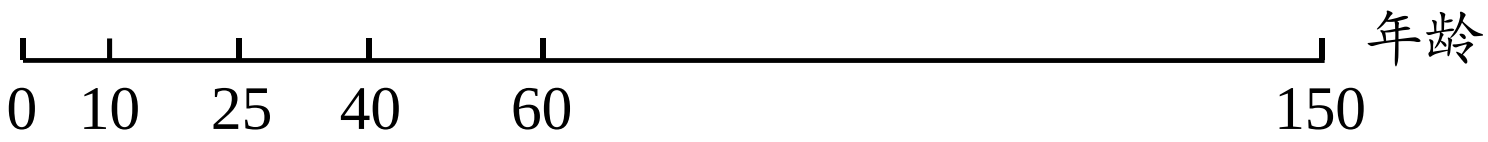


特征离散化

- 为什么要将连续型特征进行离散化处理?
  - 许多算法对特征类型有要求。如关联规则挖掘，ID3决策树，分解机
  - 为更好地**提高算法的精度**。朴素贝叶斯分类算法的正确率比没有处理的情况平均高出10%；
  - 离散化处理本质是**将连续型数据分段**，因此数据中的异常值会直接划入相应的区间段中，进而增强了之后模型对于数据异常值的**鲁棒性**；
  - 离散化后的特征，其取值均转化为有明确含义的区间号，相对于原始的连续型来说，含义更加明确，从而使得数据的**可解释性更强**，模型更易使用与理解
  - 将连续型特征离散化后，**特征的取值大大减少**，这样既减少了数据集对于系统存储空间的需求，又在算法建模中减少了模型的实际运算量，从而提升了模型的计算效率

特征离散化的基本概念

- 将连续型特征的取值范围划分为若干区间段(bin/basket)，然后使用区间段代替落在该区间段的特征取值。
- 区间段之间的分割点称之为切分点(cut point)
- 由切分点分割出来的子区间段的个数，称之为元数 (arity)



- 假设需要将"年龄"这个连续型特征切分成k个区间段，则需要$(k-1)$个切分点。"年龄"特征的取值范围在 [0,150] 之间，通过4个切分点10、25、40和60，将其转化成为5个离散区间段

特征离散化的类型

- 特征离散化目标：在数据信息损失尽量少的前提下，尽可能减少元数
- 按是否参考了数据集的y值信息划分为：
  - 无监督离散化
  - 不参考目标特征 y，直接根据特征本身的分布特性进行离散化处理
    - 等距离散化
    - 等频离散化
    - 聚类离散化
  - 有监督离散化
  - 利用参考数据集中的目标特征 y，将连续型特征进行离散化处理
    - 信息增益离散化
    - ChiMerge离散化等

特征离散化的步骤

- 1、**特征排序**。对连续型特征的取值进行升序或者降序排列，这样做可以减少离散化的运算开销；
- 2、**切分点选择**。根据给定的评价准则，合理选择切分点。（常用的评价准则基于信息增益或者基于统计量）；
- 3、**区间段分割或者合并**。基于选择好的切分点，对现有的区间段进行分割或者合并，得到新的区间段。在离散化的过程中，切分点集合的大小会随之变动；
- 4、在生成的新区间段上**重复第1-3步，直到满足终止条件**（我们可以预先设定元数 k ，作为简单的终止判断标准，也可以设定复杂的判断函数）。

特征离散化

- 等距离散化
  - 根据连续型特征的取值，将其均匀地划分成 k 个区间，每个区间的宽度均相等，然后将特征的取值划入对应的区间从而完成特征离散化
  - 如年龄取值应分布在[0,90]，确定离散化后的区间段个数为5
  - 等距离散化对输入数据质量要求高，无法解决特征存在离群值的问题。
  - 若存在离群值150，则切分点将严重偏移

- $0 \leq$ 年龄 $< 18;$
- $18 \leq$ 年龄 $< 36;$
- $36 \leq$ 年龄 $< 54;$
- $54 \leq$ 年龄 $< 72;$
- $72 \leq$ 年龄 $< 90.$

特征离散化

- 等频率离散化
  - 不要求区间段的宽度始终保持一致，而是尽量使得离散化后每一个区间内的样本量均衡
  - 根据连续型特征的总个数，将其均匀地划分成 k 个区间段，使得每个区间段中的样本数相同，然后每一份数据的取值范围即是对应的特征离散化区间
  - 缺点：有时会将同样或接近的样本划分入不同的区间，容易使得相邻区间段内的数据具有相似的特性。

样本	区间	宽度
$1, 2, 3, 4$	$[1, 4]$	4
$5, 6, 7, 8$	$[5, 8]$	4
$9, 10, 41, 42$	$[9, 42]$	34
$43, 44, 45, 46$	$[43, 46]$	4
$47, 48, 49, 50$	$[47, 50]$	4

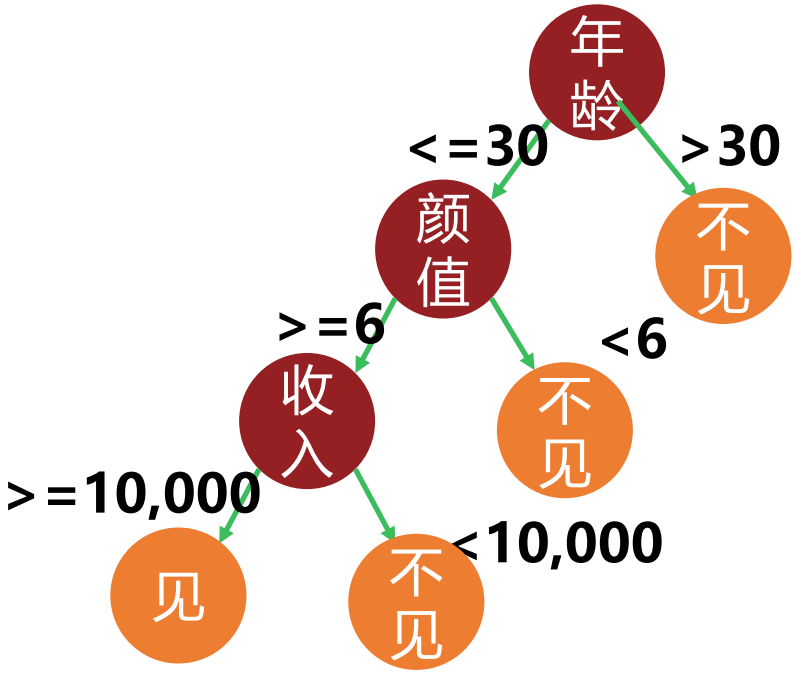
特征离散化

- 聚类离散化
  - 1.对于需要离散化的连续型特征，采用聚类算法(如K-means、EM算法等)，把样本依据该特征的分布划分成相应的簇或类；
  - 2.在聚类结果的基础上，基于特定的策略，决定是否对簇进行进一步分裂或合并。利用自顶向下的策略可以针对每一个簇继续运行聚类算法，将其细分为更小的子簇；利用自底向上的策略，则可以对邻近相似的簇进行合并处理得到新的簇；
  - 3.在最终确定划分的簇之后，确定切分点以及区间个数。

在整个聚类的过程中，我们需要事先确定簇的个数以及描述样本之间的距离计算方式。如何选定簇的个数也会影响聚类算法的效果，从而影响特征的离散化。

特征离散化

- 信息增益离散化
  - 自顶向下的分裂策略
  - 灵感源自于决策树模型建立时基于信息增益的评价标准，用信息增益（熵）来分裂连续型特征
  - 该方法最终所划分的区间个数则由单个特征决策树的叶子结点个数确定，实际应用中需要首先给定单个特征决策树的叶子结点个数。



某女孩相亲抉择决策树

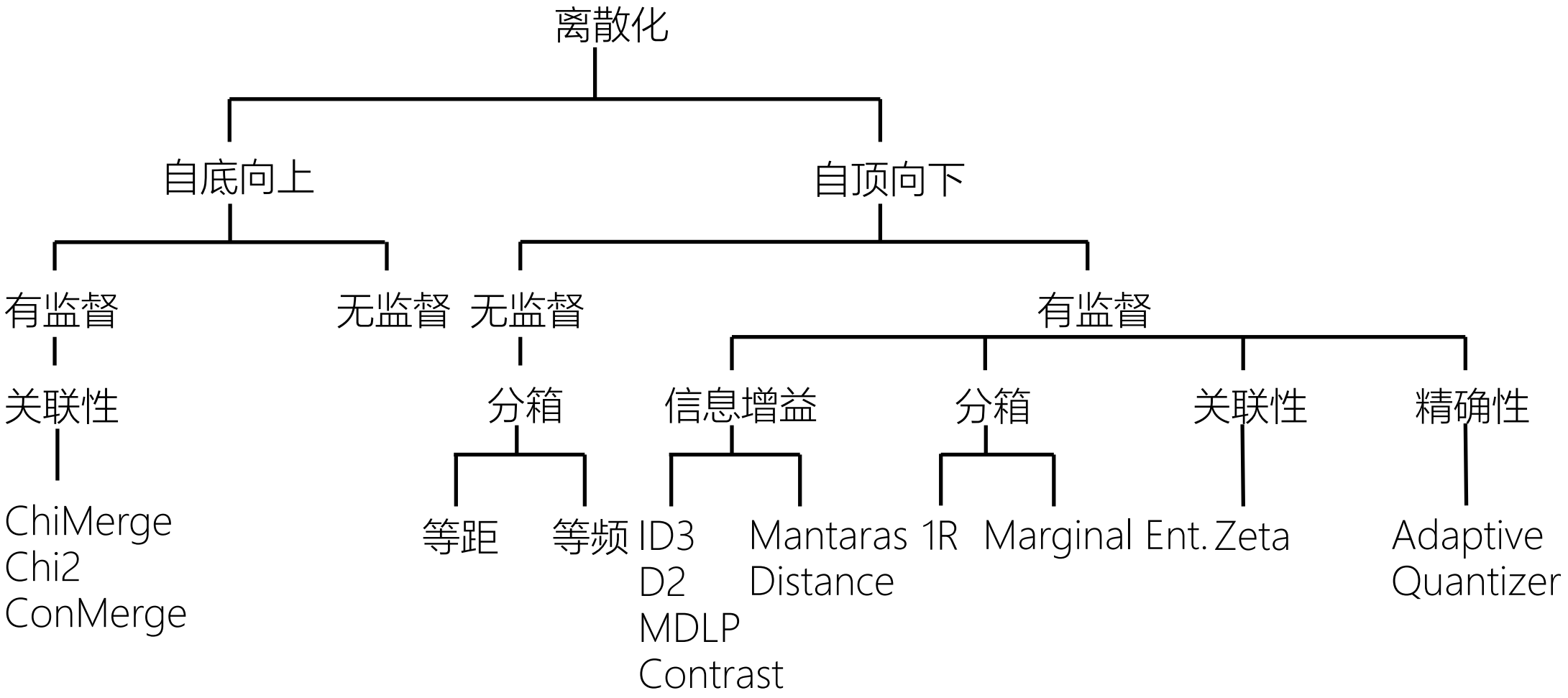
特征离散化

- 卡方离散化
  - 自底向上的合并策略
  - 常用方法：ChiMerge。通过卡方检验判断相邻区间是否需要进行合并。ChiMerge离散化过程
  - 将连续型特征的每个取值看作是一个单独的区间段，并对值进行排序；
  - 针对每对相邻的区间段，计算卡方统计量。卡方值最小或者低于设定阈值的相邻区间段合并在一起。卡方统计量的计算表达式为；

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{C} \frac{(A_{ij} - \mathbb{E}_{ij})^2}{\mathbb{E}_{ij}}$$

  - 对于新的区间段，递归进行步骤 1 和步骤 2 ，直到满足终止条件。

特征离散化小结


离散化
自底向上
自顶向下
有监督
无监督
无监督
有监督
关联性
分箱
信息增益
分箱
关联性
精确性
ChiMerge
Chi2
ConMerge
等距
等频
ID3
D2
MDLP
Contrast
Mantaras
Distance
1R
Marginal
Ent.
Zeta
Adaptive
Quantizer

总结

- 数据的初步诊断与探索
  - 数据类型、数据对象
  - 统计信息
  - 相似性度量
- 缺失值处理和离群值检测
  - 删除法、填补法
  - 基于统计、基于近邻的方法
- 常用的数据转换方法
  - 特征编码（数字编码，One-Hot编码，哑变量编码）
  - 标准化（Z-score标准化，Min-Max标准化，小数定标标准化，Logsistic标准化）
  - 离散化（等距离散化、等频离散化、信息增益离散化、卡方离散化）