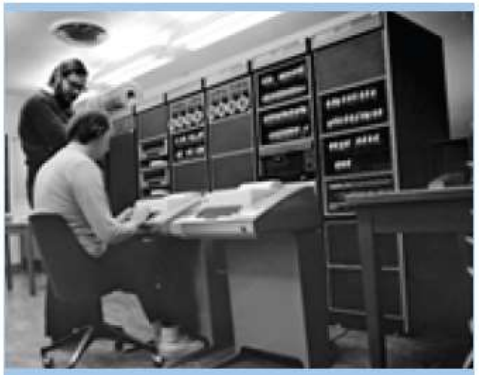
数据挖掘简介

数据

单位	英文	含义和例子
位	Bit	计算机用二进制存储和处理数据，1位是指一个二进制数为：0或1，这是存储信息的逻辑单元。
字节	Byte	计算机存储信息的基本物理单元，存出一个英文字母在计算机上，其大小就是一个字节。
千字节	KB	一页纸上的文字大概是5KB
兆字节	MB	一首普通MP3格式的流行歌曲大概是4MB
吉字节	GB	一部电影大概是1GB
太字节	TB	美国国会图书馆所有等级的印刷版书本的信息量为15TB，截至2011年年底，其网络备份的数据量为280TB。
拍字节	PB	美国邮政局一年处理的信件大约为5PB，Google每小时处理的数据约为1PB。
艾字节	EB	相当于13亿中国人人手一本500页的书加起来的信息量。
泽字节	ZB	截至2010年，人类拥有的信息总量大概1.2ZB
尧字节	YB	超出想象，难以描述

计算机发展史

1980 年以前



- 数据几乎全部驻留在数据中心
- 数据和计算集中化
- 以商用为主

1980-2000 年

- 数据和计算分布化
- 数据中心开始承担管理数据的任务
- 向娱乐领域快速扩展

2000 年至今

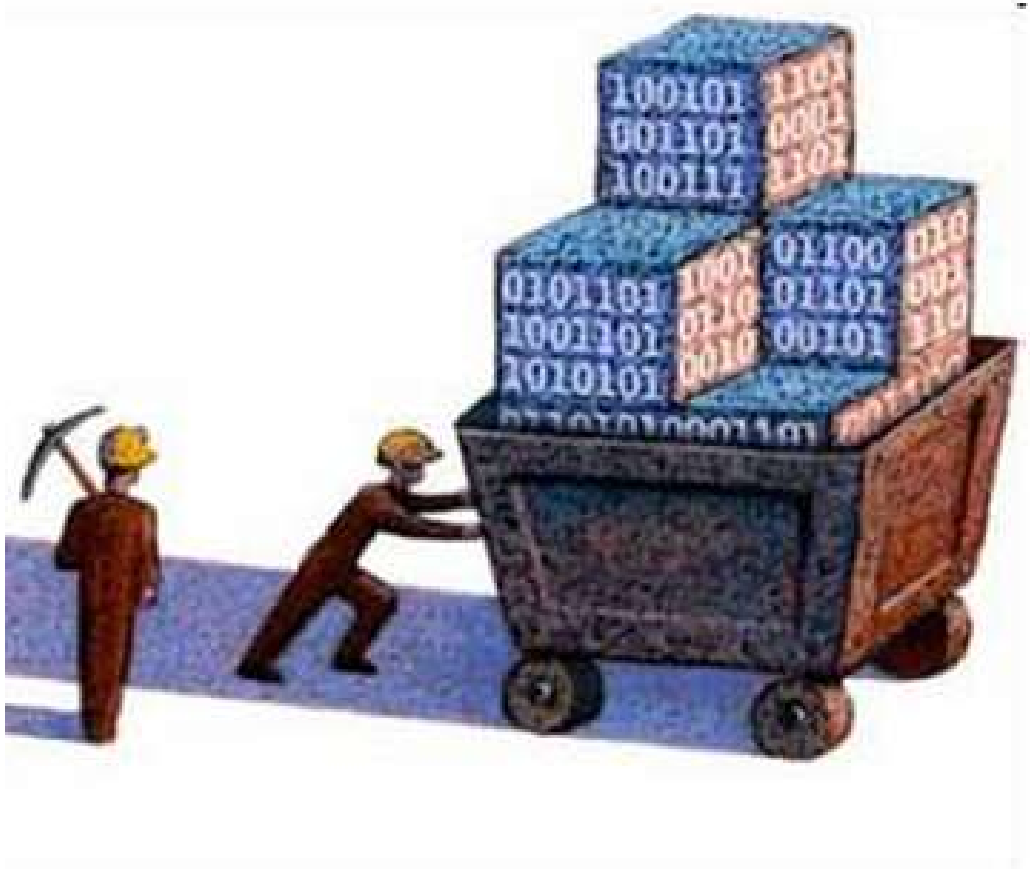
- 数据中心扩展至云基础设施
- 计算进一步分布化；数据开始收缩
- 社交媒体的加入

全球数据量增长


600
400
200
0
12
9
6
3
0
562.55
333.49
150.47
124.05
29.01
16.99
1.18
8.9
5.61
2.69
1.65
0.88
0.37
0.1
0.03
2010
2011
2012
2013
2014
2015
2016
2017
全球平板电脑移动数据流量（PB）
全球智能手机移动数据流量（EB）

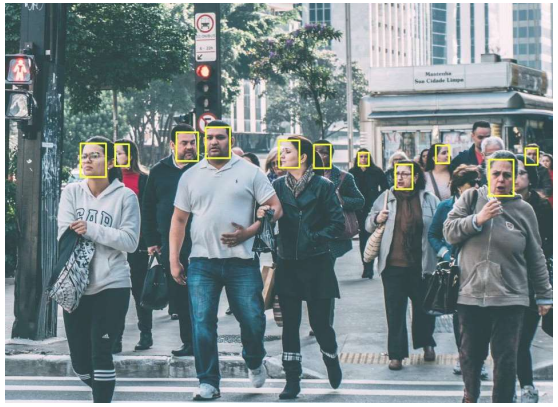
- 2006年估计有450,000台廉价的商品服务器
- 2005年索引了80亿网页
- 目前google有超过200个GFS(google文件系统)集群在运行。而每个集群大约有1000到5000台机器。 GFS存储着高达5PB的数据，成千上万的机器需要的数据都从GFS集群中检索，这些集群中数据读写的吞吐量可高达40GB每秒
- 目前google有6000个MapReduce应用程序在运行，并且以每月编写数百个新应用程序的速度在增长
- BigTable存储着数十亿的URL，数百TB的卫星图像数据和数亿用户的资料
- 每天大约要处理超过20PB的数据量
- 对4,000台机器上约为1PB的数据排序花费约6小时20分左右的时间，并且排序的结果要在48,000块硬盘上来回复制3次

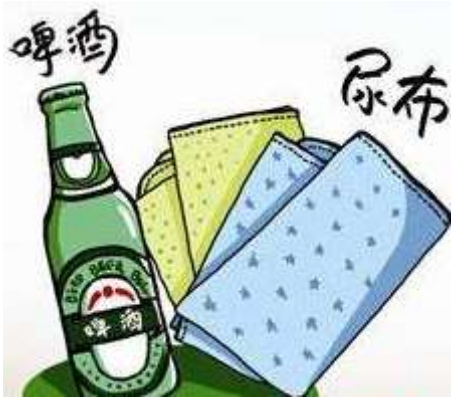
数据挖掘



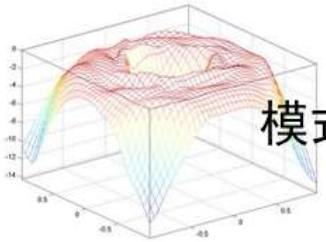
数据挖掘应用实例

恒星风：大规模监控





数据挖掘



模式识别

计算机视觉

机器学习

数据挖掘

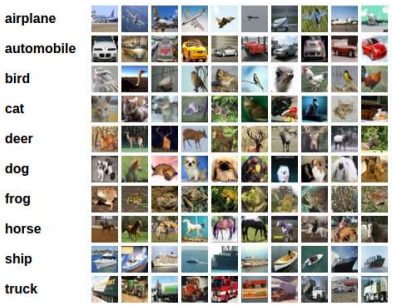
语音识别

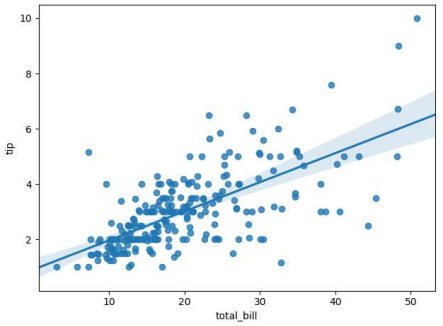
统计学习

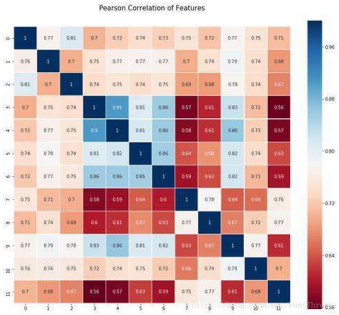
自然语言处理

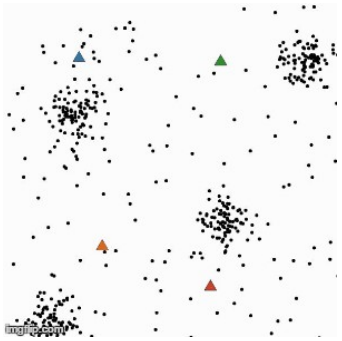
数据挖掘任务

- 预测建模
- 关联规则
- 聚类分析
- 异常检测









讨论

- 下列活动是否是数据挖掘任务：
  - 根据性别划分公司的顾客
  - 根据可盈利性划分公司的顾客
  - 计算公司的总销售额
  - 按学生的标识号对学生数据库排序
  - 预测掷一对骰子的结果
  - 使用历史记录预测某公司未来的股票价格
  - 监视病人心率的异常变化
  - 提取声波的频率

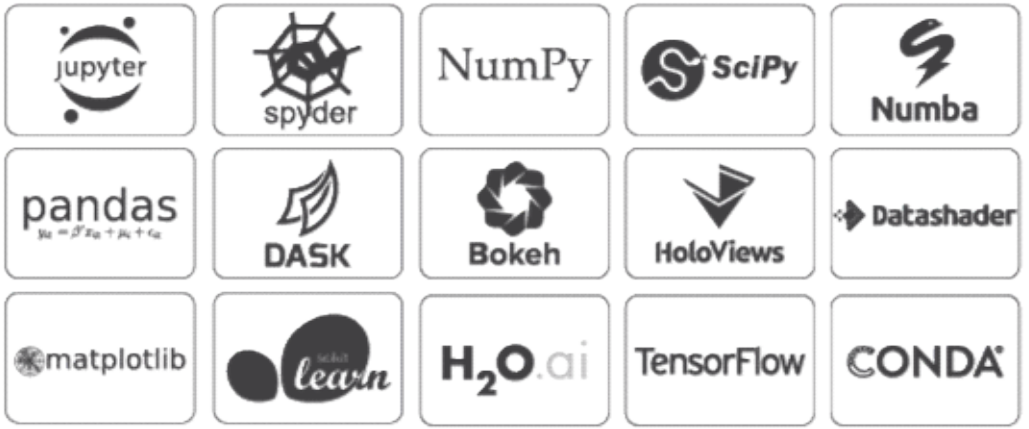
数据挖掘流程

- 问题定义
- 数据
  - 数据清洗
  - 缺失值处理
  - 特征选择
  - 数据集划分
- 建模
  - 方法选择
  - 模型训练
  - 交叉验证
- 生产环境应用

开发工具

- Python

- Anaconda

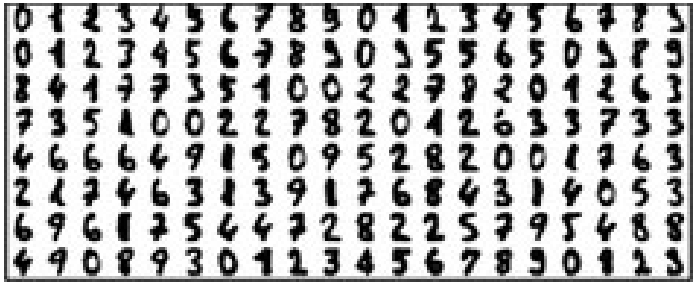


Anaconda环境配置

- Anaconda安装
  - https://www.anaconda.com/distribution/
- 创建新环境
  - conda create –n dm python=3.6
- 进入环境
  - conda activate dm
- 退出环境
  - conda deactivate
- 显示所有环境
  - conda env list

原始数据

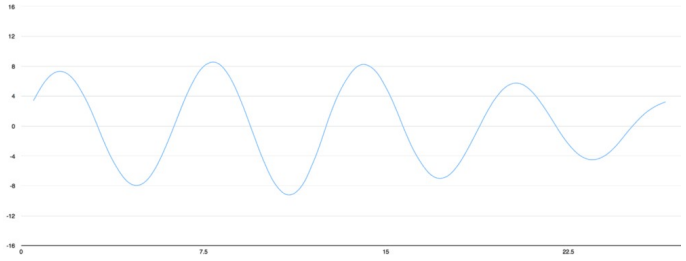
- 图片数据



- 文本数据

7月9日，习近平总书记出席中央和国家机关党的建设工作会议并发表重要讲话，精辟论述了加强和改进中央和国家机关党的建设的重大意义，并对领导干部、青年干部和党务……

- 音频数据

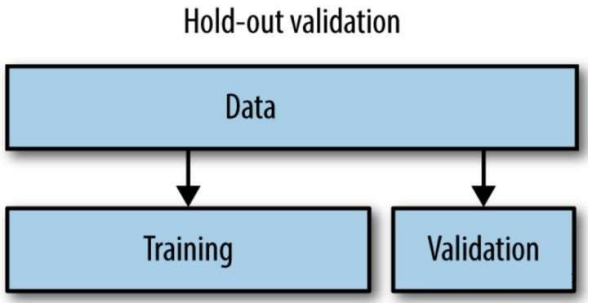


数据预处理

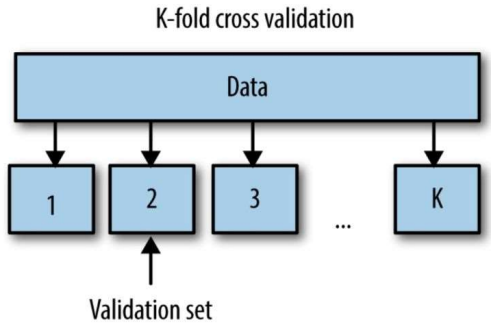
- 标准化
  - 将特征数据的分布进行调整，也就是使得数据的均值维0，方差为1
- 0-1缩放
  - 缩放到一个指定的最大值和最小值(通常是0-1)
- 二值化
  - 将数据特征转变成boolean变量
- 缺失值处理
  - 使用均值、中位值或者缺失值所在列中频繁出现的值来替换

数据集划分

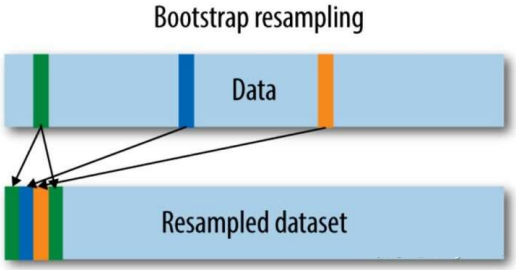
- 留出法



- 交叉验证



- 自助法



数据挖掘方法

- 监督式学习

- 无监督式学习

- 半监督学习

数据挖掘常用算法

- 线性回归算法 Linear Regression
- 支持向量机算法 (Support Vector Machine,SVM)
- 最近邻居/k-近邻算法 (K-Nearest Neighbors,KNN)
- 逻辑回归算法 Logistic Regression
- 决策树算法 Decision Tree
- k-均值算法 K-Means
- 随机森林算法 Random Forest
- 朴素贝叶斯算法 Naive Bayes
- 降维算法 Dimensional Reduction
- 梯度增强算法 Gradient Boosting

分类案例分析

回归案例分析

聚类案例分析