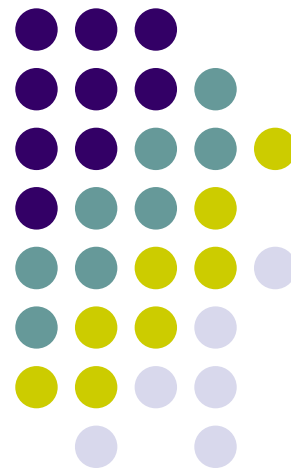


数据分析理论与Python实战

第一章 数据分析是什么

主讲教师：谭伟（13751393385）
计算机科学与技术学院





自我介绍

- 硕士导师
- CCF委员
- 计算机学会会员
- 广东省专家库专家



目录

- 海量数据背后蕴藏的知识
- 数据分析与数据挖掘的关系
- 机器学习与数据分析的关系
- 数据分析的基本步骤
- Python和数据分析

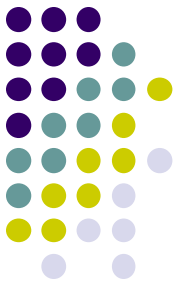


海量数据背后蕴藏的知识

- 数据蕴含了大量知识

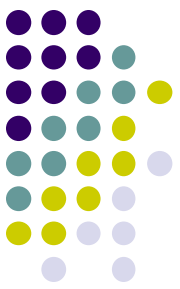


- 规律可用于
 - 解释当前发生的事情
 - 预测未来的情况



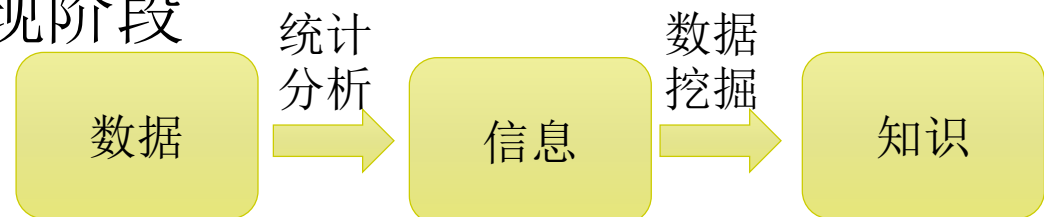
海量数据背后蕴藏的知识

- 数据分析的意义
 - 各行各业每天都在产生和收集大量数据
 - 2012年的微博日发量：4亿条
 - Twitter的信息量每年翻番增长
 - 需要从海量数据中获得有价值的信息辅助决策
 - 了解客户偏好，设计受欢迎的产品
 - 制定合适价格，确保利润同时保证市场
 - 了解市场需求，调整生产计划



数据分析与数据挖掘的关系

- 统计分析的含义
 - 在已定假设、先验约束上，对数据进行整理、筛选和加工，并得到信息的过程
- 数据挖掘的含义
 - 在数据分析得到信息的基础上进一步获得认知，转为有效的预测和决策
- 数据分析=统计分析+数据挖掘
 - 统计分析：数据预处理阶段
 - 数据挖掘：知识发现阶段





机器学习与数据分析的关系

- 机器学习的含义
 - 利用经验来改善计算机系统自身的性能
- 数据分析的含义
 - 识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程，即从海量数据中找到有用的知识
- 两者之间的关系
 - “经验”在计算机系统中主要以数据形式存在，因此机器学习需要对数据进行分析
 - 数据分析过程主要利用机器学习界提供的技术来分析海量数据



数据分析的基本步骤

- 数据收集
- 数据预处理
- 数据分析与知识发现
- 数据后处理



数据分析的基本步骤

- 数据收集

- 大数据时代的数据收集过程，得到的数据特点是大量、冗余、体量大但是信息量少
- 如何从这样的数据中提取出信息的过程是目前数据分析的重点和难点



数据分析的基本步骤

- 数据预处理
 - 将数据转化为信息
 - 主要步骤包括
 - 对数据进行初步的统计方面的分析，得到数据的基本档案
 - 从数据的一致性、完整性、准确性以及及时性四个方面进行数据质量分析
 - 根据发现的数据质量的问题对数据进行清洗，包括缺失值处理、噪声处理等
 - 对其进行特征抽取为后续的数据分析工作做准备



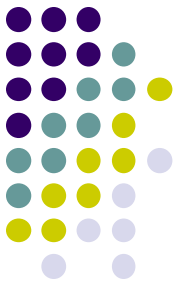
数据分析的基本步骤

- 数据分析与知识发现
 - 将预处理后的数据进行进一步的分析，完成信息到认知的过程
 - 方法分类
 - 有监督学习：分类分析、关联分析、回归分析
 - 无监督学习：聚类分析、异常检测
- 数据后处理
 - 主要包括提供数据给决策支撑系统、进行数据可视化等方面

Python和数据分析

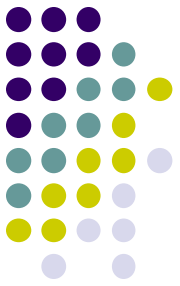


- 专用于实验性数据分析或者领域特定语言包括
 - R语言
 - Matlab
 - Sas
 - Spss



Python和数据分析

- Python语言相比上述语言的优势在于
 - Python是面向生产的
 - 大部分数据分析过程需要首先进行实验性研究和原型构建，再移植到生产系统中
 - Python适用于原型构建，且能够直接将分析程序运用到生产系统中
 - 上述语言无法直接用于生产，需要使用C/C++等语言对算法再次进行实现



Python和数据分析

- Python语言相比上述语言的优势在于
 - 拥有强大的第三方库支持
 - Python的强大功能依赖于第三方库实现
 - 常用数据分析库包括Numpy、Scipy、Pandas、Scikit-learn、Matplotlib等
 - Python的胶水语言特性
 - Python语言的底层可以用C语言实现，使用C语言实现的算法性能很高
 - Numpy的底层使用C实现，其效率比R语言更高