1. 概要部分

1.1. 基本概念

定义：从现有的大量数据中，撷取**不明显**、**之前未知**、**可能有用**的信息。
目标：

- 建立决策模型。
- 根据过去的行动预测未来行为。

1.2. 发展历程

数据搜集 -> 数据查询 -> 数据统计 -> 数据分析

1.3. KDD

步骤：

1. Attribute selection：字段选择
2. Data cleaning：数据清洗
    1. 数据污染 Pollution：录入错、数据不是最新（**模式比对算法 Pattern Recognition**）
    2. 缺乏领域一致性 Lack of domain consistency
3. Attribute enrichment：字段扩充（特征工程）
    1. 可以分为外部扩充和内部扩充
    2. 别的地方数据要很容易整合到原始数据中，这一点是必须的
4. Data coding：数据编码（让数据更适合做挖掘）
    1. 移除某些列（也可以在字段选择时做）：和目标的关系不密切、信息量不够
    2. 没有标准做法，如何找到一个最佳用于挖掘的结果
    3. 最后一个步骤：Perform Flatting Operation（摊平运算），近似于特征二值化
    4. 归户：一个客户只能有一笔数据
5. Data mining：数据挖掘
6. Reporting：报告

> 前四个阶段（1~4）是整理挖掘所需的输入，又称为 Data Warehousing，又称作数据预处理、建数据仓库。
> 实际上数据预处理可能开销 60% ~ 80% 的时间，**最重要**。

1.4. 产业标准

标准名称	厂商

标准名称	厂商
CRISP-DM	IBM SPSS
SEMMA	SAS EM

- CRISP-DM全称：Cross Industry Standard Process for Data Mining
  - Business Understanding（商业理解，定题目）
  - Data Understanding（了解数据）
  - Data Preparaion（数据准备，**最耗时间**的步骤）
  - Modeling（建模）
  - Evaluation（模型评估）
  - Deployment（发布）
- SEMMA全称（直接和数据处理相关）：Sample，Explore，Modify，Model，Assets

2. 数据挖掘技术

- 描述性统计
- 可视化技术
- 案例为本的学习

2.1. 描述性统计

最佳优惠定价

客单价：代表一个顾客一次来店消费的总金额
商家手段：互补品摆放、捆绑消费

Naive Predictions：天真预测，比较简单的描述性统计，使用$1-p$，其中 p为该用户会做这个事的概率，$1-p$为**不会做**的概率，完全根据目标字段的概率直接计算。这种计算方法可以当做我们分析的准确度的**基础模型**，模型最终分数必须比这个值要高，低了模型不可用。

2.2. 挖掘模型

2.2.1. 分类

- 描述性数据挖掘（非监督学习）：Descriptive Data Mining
  - 关联规则：Association Rules——哪些事件一起出现（电商、虚拟卖场适用）
    - Apriori
    - FP-Growth
  - 序列模式：Sequential Patterns——哪些事件循序出现（实体卖场适用）
    - AprioriAll

    - 聚类分析：Cluster Analysis——客户分群
        - 「阶层式」Single Linkage，Average Linkage，Complete Linkage
        - 「分割式」K-Means
        - 「分割式」Kohonen Self Organizing Maps（SOM）
        - 「分割式」Two-Step
- 预测性数据挖掘（监督学习）：Predictive Data Mining
    - 分类：Classification
        - 贝叶斯网络：Bayes Net
        - 逻辑回归：Logistic Regresion
        - 决策树：Decision Tree
        - 神经网络：Neural Network
        - 支持向量机：Support Vector Machine
        - K-最近邻：K-nearest Neighborhood
    - 预测：Prediction
        - 线性回归：Linear Regression
        - 时间序列：Time Series
        - 决策树：Decision Tree
        - 神经网络：Neural Network
        - 支持向量机：Support Vector Machine
        - K-最近邻：K-nearest Neighborhood

2.2.2. 重量级站点

- KDnuggets
    - 比较常用的，里面有DataSet（数据集）页面下载相关页面。
    - UCI KDD Archive
    - 数据量不大的站点
- Kaggle
    - 如果数据量比较大，上Kaggle
    - 这个站点是要为企业解决问题的
- Data Castle
- 科赛-Kesci