W12 Data anlytics

数据分析概述

- **定义**
  数据分析是处理数据以推断模式、相关性或预测模型的过程。
- **应用场景**
  - **针对个人客户**
    例如：根据客户的购买记录推荐产品。
  - **针对整体客户群**
    例如：决定生产或库存哪些产品及其数量。
- **重要性**
  数据分析是商业决策的关键，尤其是在竞争激烈的市场中。

数据分析的常见步骤

1. **数据收集**
   从多个数据源收集数据并整合到一个位置。
2. **数据转换与加载**
   - 数据需要从源格式提取（Extract）、转换为统一的模式（Transform）并加载到数据仓库中（Load）。
   - 可采用两种方式：
     - ETL（Extract-Transform-Load）：提取后先转换再加载。
     - ELT（Extract-Load-Transform）：提取后直接加载，再进行转换。
3. **生成汇总和报告**
   - 使用仪表盘（Dashboards）显示图表和报告。
   - 使用联机分析处理（OLAP）系统进行交互式查询。
4. **统计分析与模型构建**
   - 使用工具如 R、SAS、SPSS 等进行统计分析。
   - 构建预测模型以辅助决策。

数据仓库（Data Warehousing）

- **定义**
  数据仓库是一个存储从多个数据源收集而来的信息的存储库，具有统一的模式，通常用于决策支持。
- **特点**
  - 存储历史数据，支持趋势分析。
  - 将决策支持查询的负载从事务处理系统中转移出来。
  - 简化查询操作。

数据仓库架构

- 组成部分
  - **数据加载器（Data Loaders）**：负责将数据从源系统传输到数据仓库。
  - **数据库管理系统（DBMS）**：存储和管理数据。
  - **查询与分析工具**：为用户提供交互式查询和分析功能。

数据仓库的设计问题

1. **数据收集方式**
   - **源驱动架构（Source-driven Architecture）**
     数据源主动将新信息发送到数据仓库，可能是连续的或定期的（如每天晚上）。
   - **目标驱动架构（Destination-driven Architecture）**
     数据仓库定期从数据源请求新信息。
2. **数据模式的选择**
   - 需要进行模式集成（Schema Integration）。
   - 数据转换与清理（Data Transformation and Cleansing）：
     - 纠正地址中的拼写错误或邮政编码错误。
     - 合并来自不同数据源的地址列表并去重。
3. **更新方式**
   - 仓库模式可以是源模式的物化视图。
   - 数据更新通常是定期从联机事务处理系统（OLTP）中下载的。
4. **数据摘要**
   - 原始数据可能太大，无法在线存储。
   - 聚合值（如总计或小计）通常足够满足查询需求。

OLAP（联机分析处理）

- **定义**
  联机分析处理是一种交互式数据分析技术，支持数据的在线汇总和多维度查看。
- **常见操作**
  1. **旋转（Pivoting）**
     改变交叉表中的维度。
  2. **切片（Slicing）**
     基于固定值创建交叉表（也称为切块）。
  3. **上卷（Rollup）**
     从细粒度数据聚合到粗粒度数据。
  4. **下钻（Drill Down）**
     与上卷操作相反，从粗粒度数据深入到细粒度数据。

数据挖掘（Data Mining）

- **定义**
  数据挖掘是通过（半）自动化技术分析大规模数据库，发现有效、新颖、有用且可理解的模式的过程。

- **特点**
  - **有效性（Valid）**：模式在一般情况下成立。
  - **新颖性（Novel）**：模式是之前未知的。
  - **有用性（Useful）**：模式可以用于实际操作。
  - **可理解性（Understandable）**：模式易于解释。
- **任务分类**
  1. 分类（Classification）
     - 根据属性值预测实例所属的类别。
  2. 回归（Regression）
     - 预测一个连续值，例如房价或温度。
  3. 关联规则（Association Rules）
     - 找到数据项之间的关联。
  4. 聚类（Clustering）
     - 将相似的点分组为簇。

分类模型

决策树（Decision Tree）

- **定义**
  决策树是一种基于条件分支的分类模型。
- 算法
  - 常见算法包括 ID3、C4.5 和 CART。
  - 可以通过集成方法（如随机森林和提升算法）提高模型性能。

贝叶斯分类器（Bayesian Classifier）

- **定义**
  基于贝叶斯定理，计算实例属于某个类别的概率。
- **简化版本**
  朴素贝叶斯分类器（Naïve Bayes）假设属性之间相互独立。

支持向量机（SVM）

- **定义**
  在 n 维空间中寻找一个超平面，将不同类别的点分开。
- **扩展**
  使用核函数（Kernel Function）实现非线性分割。

回归模型

- **定义**
  回归用于预测一个连续值，而不是类别。
- **线性回归**
  寻找一个线性多项式来拟合数据。
- **非线性回归**
  适用于更复杂的关系，但计算过程更复杂。

聚类（Clustering)

- **定义**
  聚类通过距离度量将相似的数据点分组。
- **应用**
  可用于市场分析、客户分群等。

自监督学习（Self-supervised Learning, SSL)

- **定义**
  自监督学习通过数据自身生成标签，而无需人工标注。
- 原理
  - 隐藏或修改部分输入；
  - 训练模型恢复原始输入或分类变化内容。
- 应用
  - 自然语言处理（NLP）：如 Word2Vec、BERT 等。
  - 计算机视觉（CV）：用于图像处理和识别。