本章重难点梳理

温故才能更知新，下面让我们来快速回顾一下本章的内容，为更好的学习后面的内容，打好基础吧。

特征工程的重要性

- 特征工程可以使得机器学习模型更好的达到效果

- 推荐系统常用的特征

- 用户行为信息
- 属性、标签信息
- 用户关系信息
- 内容信息
- 上下文信息

- 原始特征的不足

- 不属于统一量纲
- 信息冗余
- 存在非定量的定性特征
- 存在缺失值

- 特征工程的常见处理方法

- 标准化
  - 较适合本身就呈现正态分布的数据（如价格）
  - 对异常值不敏感
- 归一化
  - 适合本身分布不确定的数据（如哑编码后端分类数据）
  - 对异常值较为敏感
- 二值化
  - 将定性特征转化为定量特征
- 哑编码
  - 将离散属性分类特征转化为0、1向量
- 缺失值补全
  - 常用补0、平均值、中位数等方法

- Apache Spark

- 开源的分布式计算框架
  - 计算速度快：相对于Hadoop有最多100倍的提升
  - 强大的缓存设计：通过简单的接口提供内存+硬盘缓存
  - 部署灵活：支持YARN，k8s等集群管理工具
  - 实时性高：提供专门针对流计算的工具
  - 通用性高：提供多种语言API以及各种业务抽象
- RDD
  - Resilient Distributed Dataset
  - Resilient：良好的容错性和错误自动恢复能力
  - Distributed：天生的分布式
  - Dataset：对用户提供统一的、分布透明的编程接口

- 行为数据采集

- 用户与产品交互时产生的数据，如点赞、收藏、浏览
- 通常由客户端埋点上传
- 为何使用Kafka处理行为数据？
  - 解耦：消息生产者和消费者可以互相独立工作
  - 拓展性：应对用户量快速扩张可以高效扩容
  - 削峰填谷：在活动期间有效保障流量平稳分发
  - 异步通信：适合处理行为数据
- Kafka核心概念
  - Broker：集群中的服务器
  - Topic：消息的逻辑类别
  - Partition：topic下的物理存储单元
  - Producer\Consumer：消息生产、消费者
  - Consumer Group：消费者群组