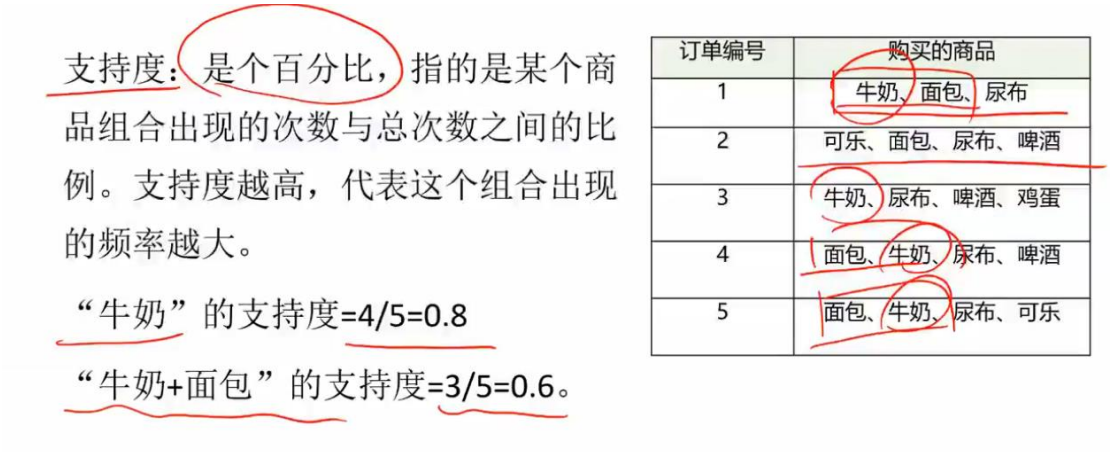
**Thinking1 关联规则中的支持度、置信度和提升度代表的什么，如何计算？**

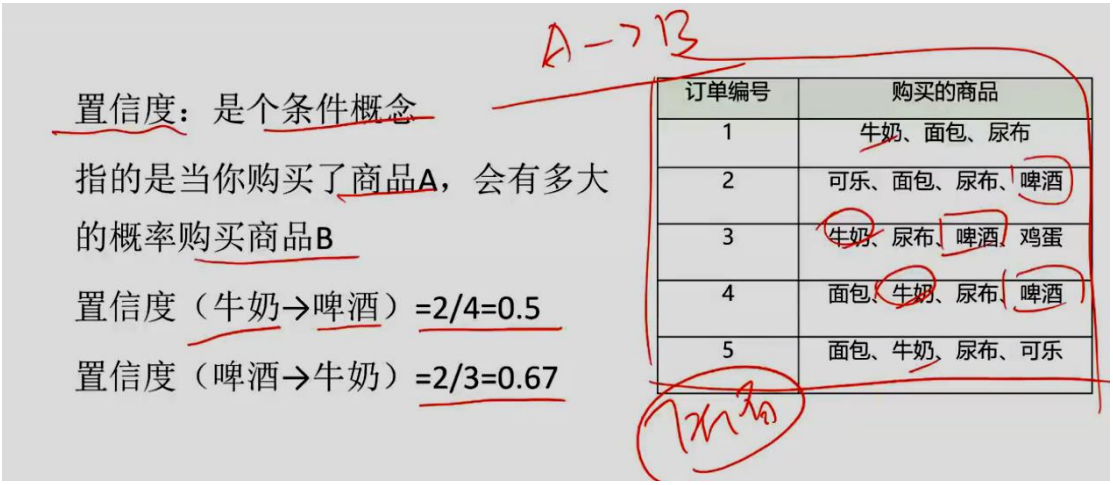
A1：

支持度：是一个百分比，指的是某个商品组合出现的次数与总次数之间的比例。支持度越高，代表这个组合出现的频率越大




| 订单编号 | 购买的商品 |
|---|---|
| 1 | 牛奶、面包、尿布 |
| 2 | 可乐、面包、尿布、啤酒 |
| 3 | 牛奶、尿布、啤酒、鸡蛋 |
| 4 | 面包、牛奶、尿布、啤酒 |
| 5 | 面包、牛奶、尿布、可乐 |

置信度：可信程度，条件概率。购买了商品 A，会有多大概率购买商品 B

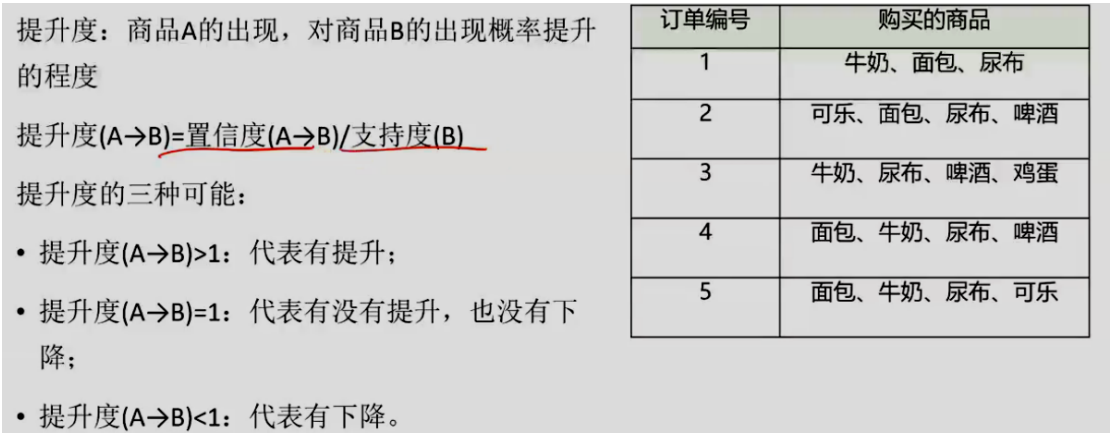



| 订单编号 | 购买的商品 |
|---|---|
| 1 | 牛奶、面包、尿布 |
| 2 | 可乐、面包、尿布、啤酒 |
| 3 | 牛奶、尿布、啤酒、鸡蛋 |
| 4 | 面包、牛奶、尿布、啤酒 |
| 5 | 面包、牛奶、尿布、可乐 |

提升度：商品 A 的出现，对商品 B 的出现概率提升的程度

提升度（A->B）= 置信度（A->B） / 支持度 B

提升度>=3 效果较好




| 订单编号 | 购买的商品 |
|---|---|
| 1 | 牛奶、面包、尿布 |
| 2 | 可乐、面包、尿布、啤酒 |
| 3 | 牛奶、尿布、啤酒、鸡蛋 |
| 4 | 面包、牛奶、尿布、啤酒 |
| 5 | 面包、牛奶、尿布、可乐 |

**Thinking2 关联规则与协同过滤的区别**

A2：

关联规则：使用关联规则算法可以从大量的过往交易数据中获取规则。它可以是会同时被购买的商品之间的关联规则，也可以是按时间依次被购买商品的序列模型。关联规则直接从数据中挖掘潜在的关联，与个人的偏好无关，忽略对于用户的个性化的推荐

协同过滤：分为两类 User-CF，item—CF

基于用户的协同过滤推荐算法先使用统计的方法寻找与目标用户有相同喜好的邻居,然后根据目标用户的邻居的喜好产生向目标用户的推荐

基于物品的协同过滤根据物品之间的相似度对物品进行推荐，有点类似关联规则但是在有用户评分的情况下（如电影评分），协同过滤算法应该比传统的关联规则更能产生精准的推荐

**Thinking3 为什么我们需要多种推荐算法**

A3：

常见的推荐系统算法：

1 基于内容的特征

2 基于协同过滤的推荐

3 基于关联规则的推荐

4 基于效用的推荐（灵活度差，对于用户偏好很敏感）

5 基于知识的推荐

组合推荐：每种推荐算法都有自己使用的场景，在实际使用的过程中灵活使用组合式的方法，提高正确率 or 推荐率

**Thinking4 关联规则中的最小支持度、最小置信度该如何确定**

A4：

通过实验得到预设值

设置最小支持度，开始假设不同的初始值，先看前 20 个，再进行下一步分析

一般来说，最小支持度可能在 0.01-0.5 之间；

最小置信度可能是 0.5-1 之间

**Thinking5 如何通过可视化的方式探索特征之间的相关性**

A5：

很多特征选择里常用的可视化 EDA 方法，最常用的是皮尔森相关系数。也有一些比如 ELI5，SHAP 等有效的可视化特征选择框架。

Action1、2　kaggle 地址：

https://www.kaggle.com/zhangwx95/marketbasket-analysis