**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**

* 支持度（support）：是个百分比，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。

例如：

|  |  |
| --- | --- |
| 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 |

“牛奶”的支持度=4/5=0.8

“牛奶+面包”的支持度=3/5=0.6。

* 置信度（confidence）：是个条件概念。指的是当你购买了商品A，会有多大的概率购买商品B。

在上例中：置信度（牛奶→啤酒）=2/4=0.5 （买了牛奶又买了啤酒的次数/买牛奶的次数）

置信度（啤酒→牛奶）=2/3=0.67 （买了啤酒又买了牛奶的次数/买啤酒的次数）

* 提升度（lift）：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)

提升度的三种可能：

提升度(A→B)>1：代表有提升； 即：置信度(A→B)>支持度(B)

提升度(A→B)=1：代表有没有提升，也没有下降； 即：置信度(A→B)=支持度(B)

提升度(A→B)<1：代表有下降。 即：置信度(A→B)<支持度(B)

**Thinking2 关联规则与协同过滤的区别**

* 关联规则是基于整体的transaction。关联规则没有利用“用户偏好”，而是基于整体的购物订单进行的频繁项集挖掘。应用场景：当下的需求；新用户，解决冷启动的问题
* 协同过滤是基于个体的偏好，根据用户的长期偏好做个性化推荐。
* 商品组合使用的是购物篮分析，也就是Apriori算法，协同过滤计算的是相似度。
* 两种推荐算法的思考维度不同，很多时候，我们需要把多种推荐方法的结果综合起来做一个混合的推荐

**Thinking3 为什么我们需要多种推荐算法**

* 实际工作中经常采用
* 每种推荐算法都有自己的使用场景，可以综合考虑

**Thinking4 关联规则中的最小支持度、最小置信度该如何确定**

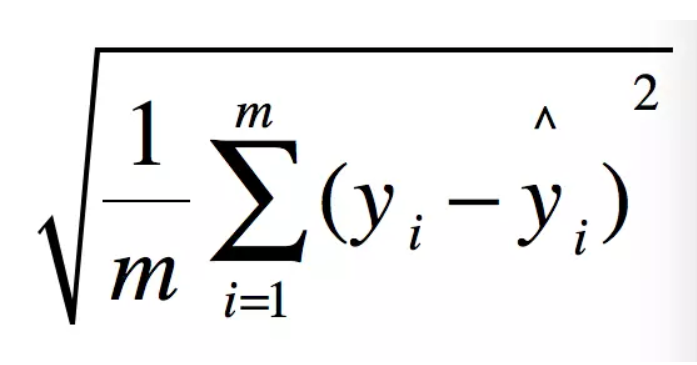
* 最小支持度，最小置信度是实验出来的
* 最小支持度：不同的数据集，最小值支持度差别较大。可能是0.01到0.5之间。可以先设置一个较小的值，从高到低打印出来后再进行设置，例如选择输出前20个项集的支持度作为参考。
* 最小置信度：可能是0.5到1之间。
* 提升度：表示使用关联规则可以提升的倍数，是置信度与期望置信度的比值。提升度至少要大于1。

**Thinking5 都有哪些常见的回归分析方法，评价指标是什么**

* 回归分析（Regression）是确定两种或两种以上变量之间相互依赖的定量关系的统计方法，使用非常广泛：
* 按照涉及的变量的多少，分为一元分析和多元回归分析
* 多项式回归分析：[自变量](https://zh.wikipedia.org/wiki/%E8%87%AA%E5%8F%98%E9%87%8F%E5%92%8C%E5%9B%A0%E5%8F%98%E9%87%8F) x 和[因变量](https://zh.wikipedia.org/wiki/%E8%87%AA%E5%8F%98%E9%87%8F%E5%92%8C%E5%9B%A0%E5%8F%98%E9%87%8F) y 之间的关系被建模为关于 x 的 n 次[多项式](https://zh.wikipedia.org/wiki/%E5%A4%9A%E9%A1%B9%E5%BC%8F)。
* 按照因变量的多少，分为简单回归分析和多重回归分析
* 按照自变量和因变量之间的关系类型，分为线性回归分析和非线性回归分析
* 损失函数可以衡量模型的好坏。常见的损失函数有：
* MSE，均方误差，是在回归问题中比较常用的损失函数。



* RMSE（Root Mean Square Error，均方根误差）是MSE的平方根。



* MAE（Mean Absolute Error，平均绝对误差）

