# 第四问

310.

第四问让我们研究不同类比成分之间的关联关系,这里的类别我理解为表一中的四种属性,即

- (1) 按类型,成分之间的关联关系
- (2) 按纹饰,成分之间的关联关系
- (3) 按颜色,成分之间的关联关系
- (4) 按风化程度,成分之间的关联关系

既然是研究不同成分之间的关联关系,因为用关联分析法最为适合,关联分析法的原理和介绍如下:

## 关联规则简介

关联规则挖掘可以让我们从数据集中发现项与项之间的关系,它在我们的生活中有很多应用场景,"购物篮分析"就是一个常见的场景,这个场景可以从消费者交易记录中发掘商品与商品之间的关联关系,进而通过商品捆绑销售或者相关推荐的方式带来更多的销售量。

- 1. 搞懂关联规则中的几个重要概念: 支持度、置信度、提升度
- 2. Apriori 算法的工作原理
- 3. 在实际工作中,我们该如何进行关联规则挖掘

# 关联规则中重要的概念

举一个超市购物的例子,下面是几名客户购买的商品列表

订单编号	购买商品	THE A	-5114	
1 592A25TT TELL CHEE	牛奶、面包、	尿布	287852	antill
20 kg start	可乐、面包、	尿布、啤酒	367	30
3 18 18 035114	牛奶、尿布、	啤酒、鸡蛋	THE PROPERTY OF	2825110
4 BORD RAINE	面包、牛奶、	尿布、啤酒	a catt	20.
5 ASSIST	面包、牛奶、	尿布、可乐	antilla	A A

#### 支持度

支持度是个百分比,它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。

我们看啤酒出现了3次,那么5笔订单中啤酒的支持度是3/5=0.6。同理,尿布出现了5次,那么尿布的支持度是5/5=1。尿布和啤酒同时出现的支持度是3/6=0.6。

#### 置信度

它指的就是当你购买了商品 A, 会有多大的概率购买商品 B。

我们可以看上面的商品,购买尿布的同时又购买啤酒的订单数是 3,购买啤酒的订单数是 3,那么(尿布->啤酒)置信度= 3/3=1。

再看购买了啤酒同时购买尿布的订单数是 3,购买尿布的订单数是 5,那么(啤酒->尿布)置信度=3/5=0.6。

#### 提升度(Lift):

表示含有商品 A 的条件下,同时含有商品 B 的概率,与商品 B 总体发生的概率 之比。

Lift( $A \rightarrow B$ )= P( $B \mid A$ ) / P(B)

# Apriori 的工作原理

Apriori 算法其实就是查找频繁项集 (frequent itemset) 的过程,所以我们需要先了解是频繁项集。

频繁项集就是支持度大于等于最小支持度阈值的项集,所以小于最小值支持度的项目就是非频繁项集,而大于等于最小支持度的项集就是频繁项集。

下面我们来举个栗子:

3.70

假设我随机指定最小支持度是 0.2。首先,我们先计算单个商品的支持度:

	购买商品	9282511	- A Gill	支持度	200	100	285852
0	牛奶面包	all so	tanta	4/5	EB2825114	A TEL CART	. 00
	尿布	at A.A.	474	5/5		Carre	S.B.
	可乐		C85852	2/5	A5515	()	
	啤酒	4		3/5		2051711	
	鸡蛋 ~~~~~~	Ab G	SSIS	1/5	*	28585	A Po

因为最小支持度是 0.2, 所以你能看到商品 鸡蛋 是不符合最小支持度的,不属于频繁项集,于是经过筛选商品的频繁项集如下:

购买商品	21th.	支持度	-0.251111	Clar.	
牛奶	* A551	4/5	5020	AL LOND	A.O.
面包	A Car	2511 M	ER.	K'a'	
尿布 69		5/5	A ASS	4	
可乐	a Citile.	2/5	All tax	08251	A. (
啤酒 4828	30	3/5	**	280	ALL TOP

在这个基础上,我们将商品两两组合,得到两个商品的支持度:

690		- Gro	619		0782	90	
购买商	Buu		- NS5		支持度	-O'Car	&P
牛奶、	面包	all's		05170	3/5	Co.	
牛奶、	尿布		587	A THE	4/5	4.	
牛奶、	可乐	26170	Oth.	is tall.	1/5	982511	, G
牛奶、	啤酒		30	- A551-	2/5	5820	30
面包、	尿布		in a	181	4/5	ista.	
面包、	可乐			2879	2/5	AASS	
面包、	啤酒		6170	Oaki.	2/5	The same of the sa	25/1
尿布、	可乐	4	32812	A	2/5	. 68°	
尿布、	啤酒	Alle .	*ar		3/5	Clar.	
可乐、	啤酒	A G			1/5 5828	C. The	AP.

#### 筛选大于最小支持度(0.2)的数据后

3.70

	P		411	A. C.	Kall's	a light	11h
	购买商	品	028251		支持度	28585	
	牛奶、	面包	20 all fran	THE ALL	3/5	Oliki.	C. arr
	牛奶、	尿布	Sister		4/5	, All	15515
5	牛奶、	啤酒	TO AS	A SERVICE	2/5	C. P.	2511
	面包、	尿布	1929251	Ab Gu	4/5		28585
	面包、	可乐	W. Po	* ant.	2/5	05/10	Olik.
	面包、	啤酒	- Ab Gu	3515	2/5	3282	AAS
	尿布、	可乐	anti-	4112	2/5	Stan	A STATE OF THE STA
	尿布、	啤酒		28585.2	3/5	4 A5512	

#### 在这个基础上,我们再将商品三个组合,得到三个商品的支持度:

	购买商品		1551		. 58Zo	支持度	TIS ASS	
	牛奶、面包	]、尿布	-0251	G.		3/5	All I	01852
	牛奶、面包	1、可乐	28,10	THE PARTY OF THE P	AS	1/5	di.	20.
	牛奶、面包	1、啤酒	Offit.	sista.		1/5	Gio.	
S	面包、尿布	5、可乐	, a.P	55		1/5	- antin	A BA
	面包、尿布	5、啤酒	A. C. C.	082511	Gils.	2/5	32"	**
	尿布、可分	《、啤酒		28,0	antila	1/5	420	

#### 筛选大于最小支持度(0.2)的数据后

_	购买商品	AASSIL AA	ali.		支持度	MIN AS	d
	牛奶、面包、尿布	082511	Glav.	Sist	3/5	V	-8282°
	面包、尿布、啤酒	285	antibe.	A A A	2/5	di	. 2

#### 在这个基础上,我们将商品四个组合,得到四个商品的支持度:

-0.25111

购买商品	MAA	05/14	City.	支持度	
牛奶、面包、尿布、可乐	582	9r	18 a D	1/5	
牛奶、面包、尿布、啤酒	O.184.	tan.	A Line	1/5	
面包、尿布、可乐、啤酒	20	15515		1/5 5828	ALTO

再次筛选大于最小支持度(0.2)数据的话,就全删除了,那么,此时算法结束,上一次的结果就是我们要找的频繁项,也就是{牛奶、面包、尿布}、{面包、尿布、啤酒}。

#### 总结一下上述 Apriori 算法过程:

3.70

- 1. K=1, 计算 K 项集的支持度
- 2. 筛选掉小于最小支持度的项集
- 3. 如果项集为空,则对应 K-1 项集的结果为最终结果
- 4. 否则 K=K+1, 重复 1-3 步
- 5. 我们可以看到 Apriori 在计算的过程中有以下几个缺点:
- 6. 可能产生大量的候选集。因为采用排列组合的方式,把可能的项集都组合 出来了
- 7. 每次计算都需要重新扫描数据集,来计算每个项集的支持度

## 化学成分分析

因此,利用上述关联分析法,基于某种类型的分布下,对组成成分,例如二氧化 硅,氧化钠等等一步一步的进行求其支持度,置信度等,根据这些概率的大小, 我们可以得到在某种类别下,那些成分之间的关联性较强,哪些较弱,比较其差 异性。

这里只是提供一个思路,具体解题方式根据算法一步一步求解,利用 python pandas 数据处理库等工具会较为方便。

ejstantill Olli. Braksti

Will Assistantill Olli. Buaren