北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

相随相伴、谓之关联

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

模式评估

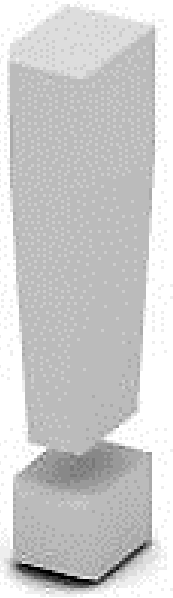
所有模型都是错的，但有些是有用的

所有模型必须经过评估：关联规则也不例外

频繁模式已然产生

根据数据分析的套路，接下来——

如何评估所产生的模式?



一个误导的强关联规则

假设我们对涉及购买计算机游戏*game*和录像*video*的事务感兴趣

在所分析的10000个事务中，数据显示有6000个事务包含计算机游戏，7500个事务包含录像，而4000个事务同时包含计算机游戏和录像

设置最小支持度为30%，最小置信度为60%，将发现下面的规则：

$$buys(x, \text{“game”}) \Rightarrow buys(X, \text{“video”})$$

$$[support = 40\%, confidence = 66\%]$$

可见该规则是强规则，满足最小支持度和最小置信度的要求

然而，**二者是负相关关系**：购买录像video的概率是75%， 比66%还高

从关联分析到相关分析

提升度是一个简单的相关性的度量：

$$lift(\mathrm{A}, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

- 如果提升度小于1，则A的出现和B的出现是负相关的，意味着一个出现可能导致另一个不出现
- 如果提升度大于1，则A的出现和B的出现是正相关的，意味着一个出现蕴含另一个的出现
- 换言之，它评估一个的出现“提升”另一个出现的程度

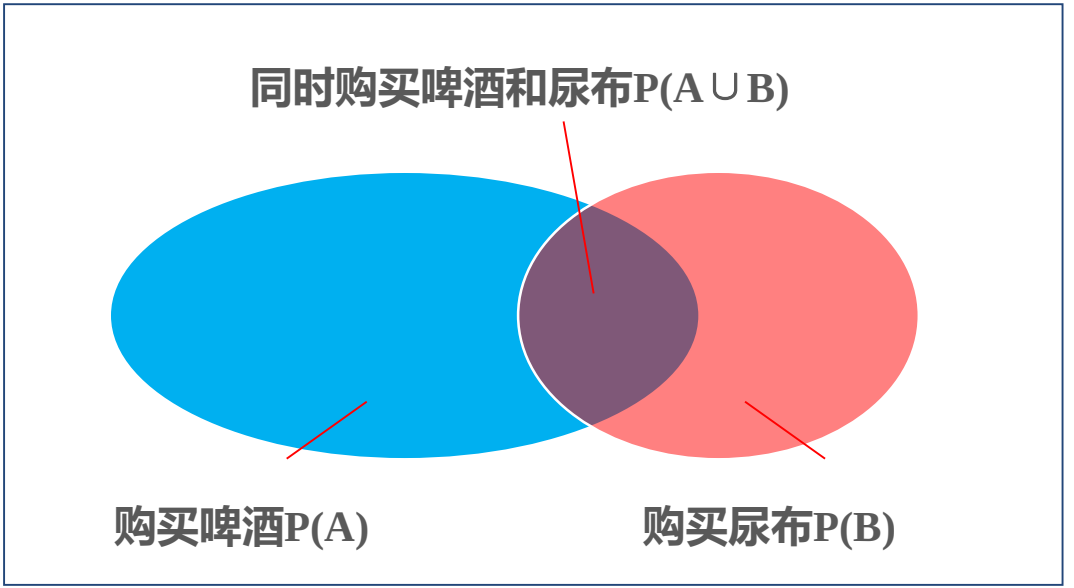
从关联分析到相关分析

提升度是一个简单的相关性的度量：

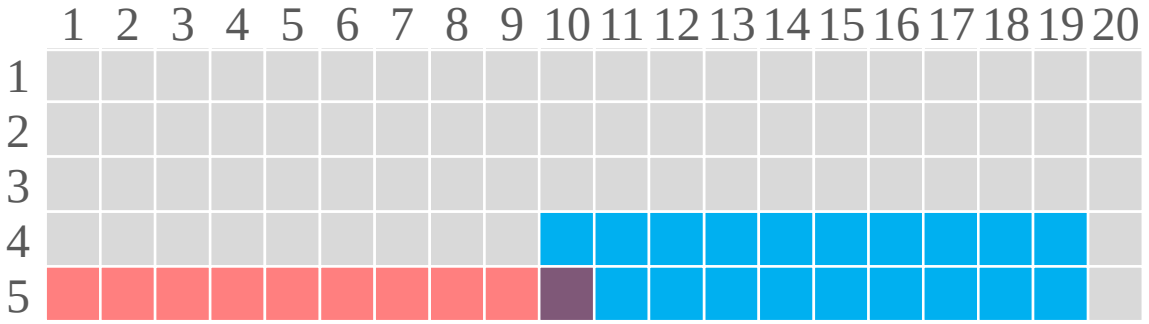
$$lift(\mathrm{A}, B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}$$

- 如果提升度小于1，则A的出现和B的出现是负相关的，意味着一个出现可能导致另一个不出现
- 如果提升度大于1，则A的出现和B的出现是正相关的，意味着一个出现蕴含另一个的出现
- 换言之，它评估一个的出现“提升”另一个出现的程度

理解提升度：文氏图是很好的工具

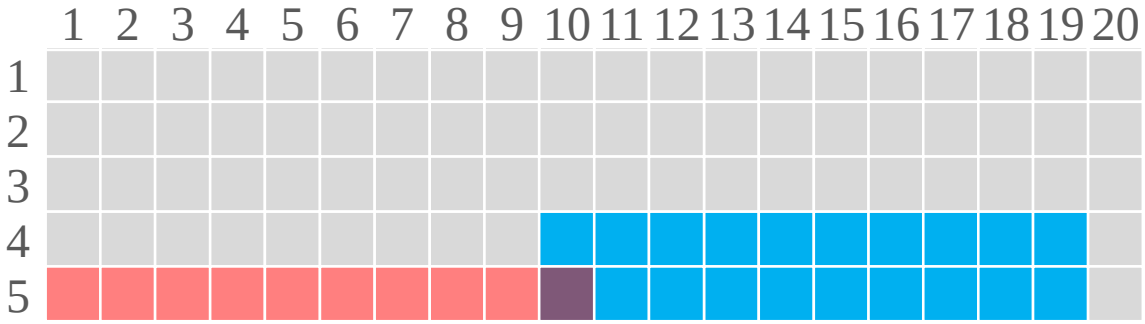


理解提升度：我们还有更好的工具



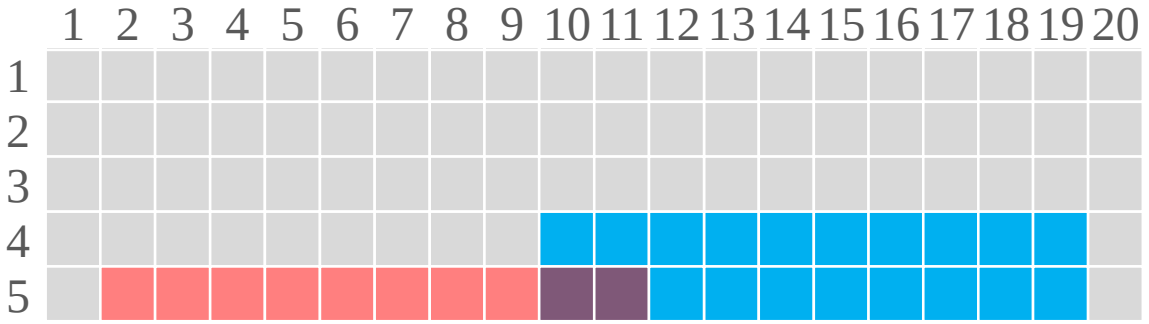
指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	
	尿布→啤酒	重叠部分面积与红色部分之比	
提升度：		（红中之蓝）与（总蓝）之比	

理解提升度：我们还有更好的工具



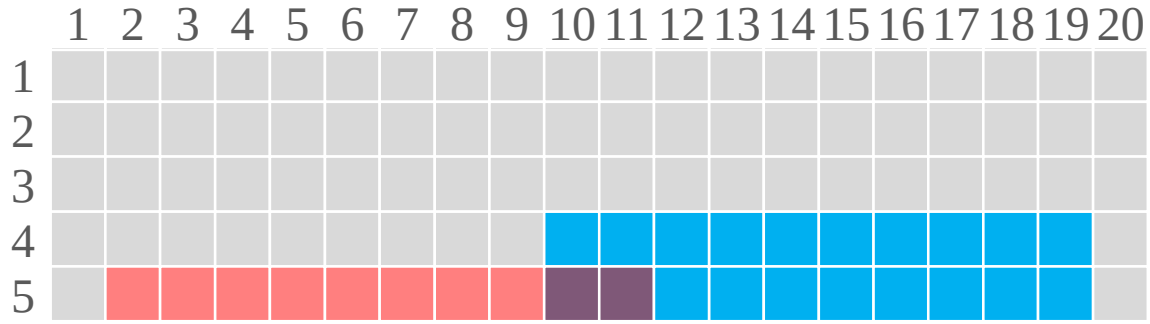
指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	1/100
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	1/20
	尿布→啤酒	重叠部分面积与红色部分之比	1/10
提升度：		（红中之蓝）与（总蓝）之比	0.5

理解提升度：我们还有更好的工具



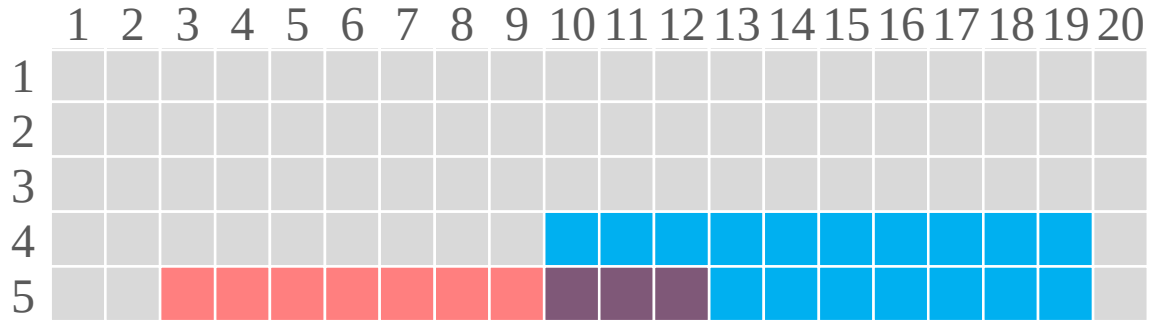
指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	
	尿布→啤酒	重叠部分面积与红色部分之比	
提升度：		（红中之蓝）与（总蓝）之比	

理解提升度：我们还有更好的工具



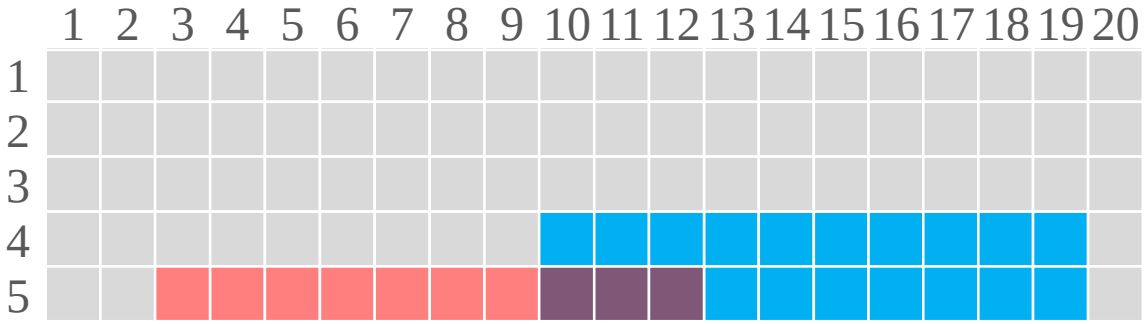
指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	2/100
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	2/20
	尿布→啤酒	重叠部分面积与红色部分之比	2/10
提升度：		（红中之蓝）与（总蓝）之比	1

理解提升度：我们还有更好的工具



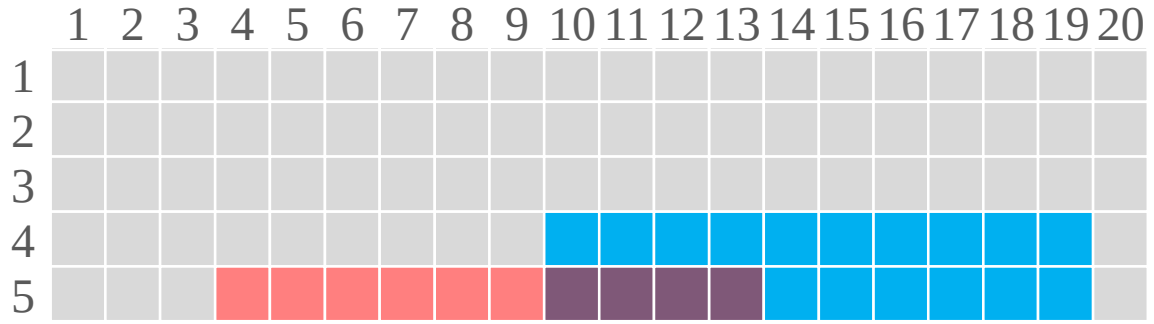
指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	
	尿布→啤酒	重叠部分面积与红色部分之比	
提升度：		（红中之蓝）与（总蓝）之比	

理解提升度：我们还有更好的工具



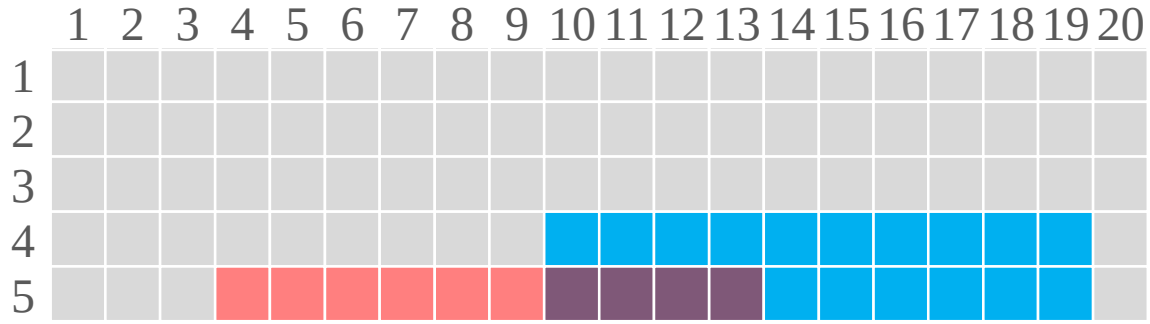
指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	3/100
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	3/20
	尿布→啤酒	重叠部分面积与红色部分之比	3/10
提升度：		（红中之蓝）与（总蓝）之比	1.5

理解提升度：我们还有更好的工具



指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	
	尿布→啤酒	重叠部分面积与红色部分之比	
提升度：		（红中之蓝）与（总蓝）之比	

理解提升度：我们还有更好的工具



指标		计算方法	取值
支持度：		重叠部分面积与总面积之比	4/100
置信度：	啤酒→尿布	重叠部分面积与蓝色面积之比	4/20
	尿布→啤酒	重叠部分面积与红色部分之比	4/10
提升度：		（红中之蓝）与（总蓝）之比	2

谢谢聆听
Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics