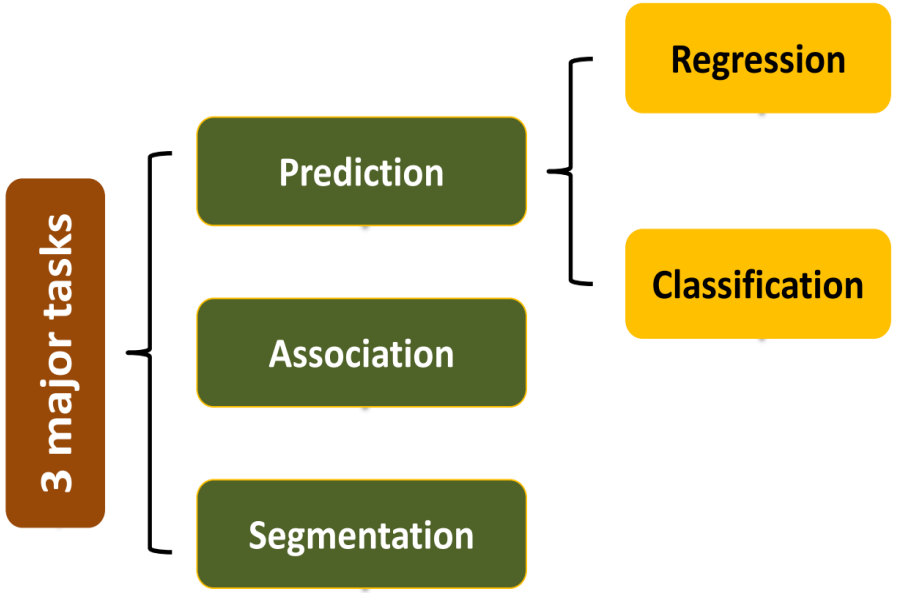
MM5425 商业分析

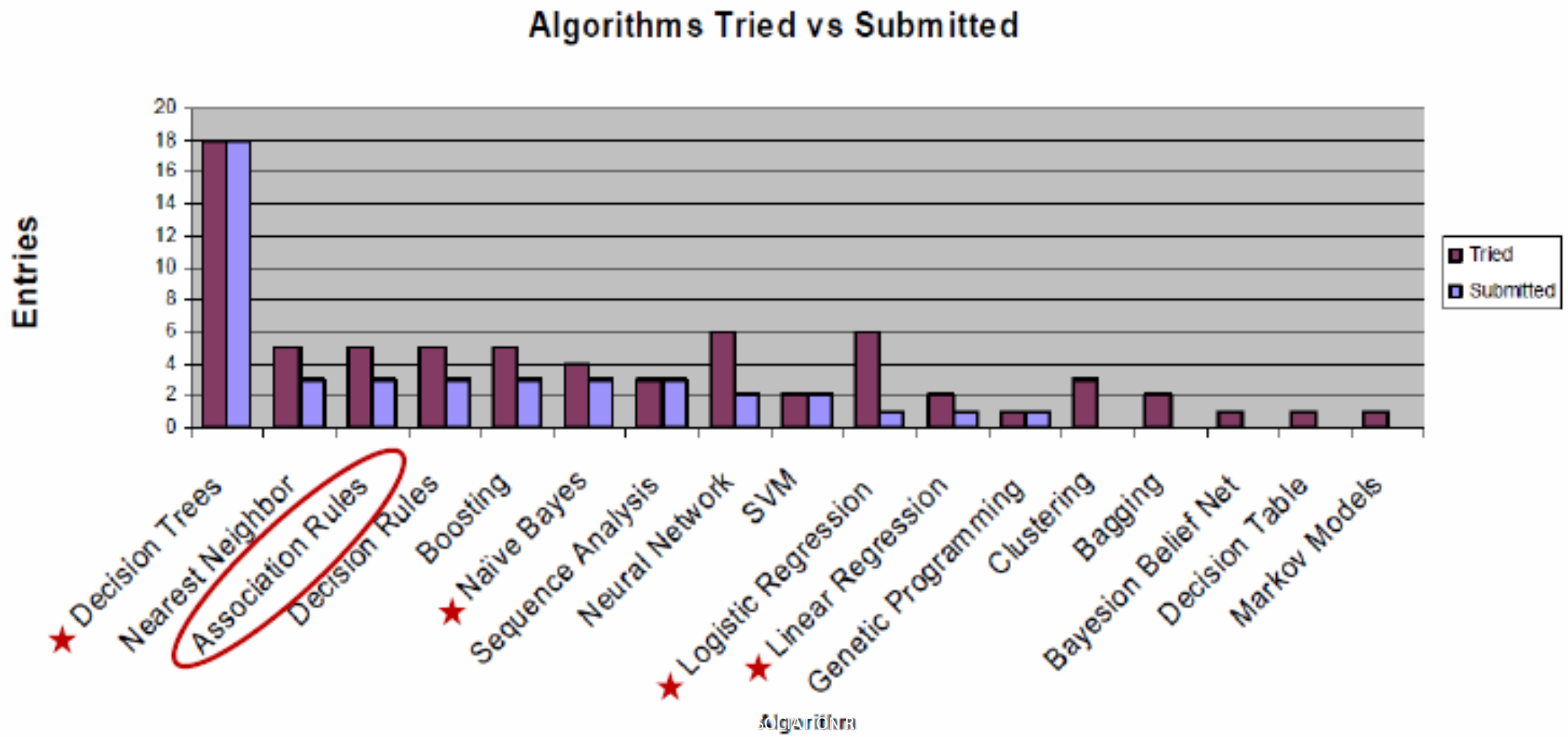
WEEK 9 LECTURE – ASSOCIATION RULE LEARNING

WK9 目录 Contents

- 什么是关联规则学习
- 重要术语与定义
  - 项集、条件（前件）、结果（后件）
  - 支持度 Support
  - 置信度 Confident
  - 提升度 Lift



常用算法

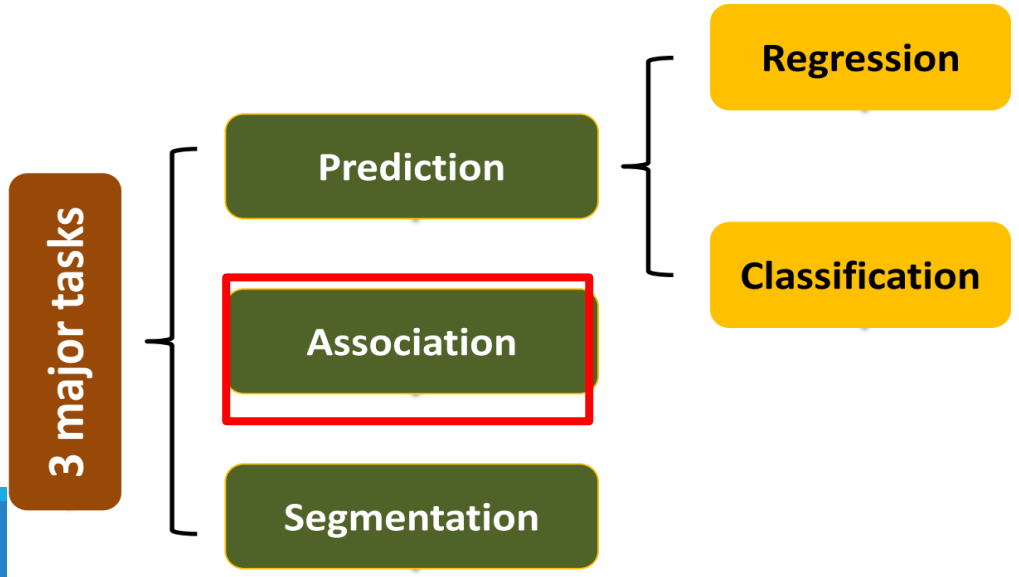

Algorithms Tried vs Submitted
Entries
20
18
16
14
12
10
8
6
4
2
0
Tried
Submitted
★Decision Trees
Nearest Neighbor
Association Rules
Decision Rules
Boosting
★Naïve Bayes
Sequence Analysis
Neural Network
SVM
★Logistic Regression
★Linear Regression
Genetic Programming
Clustering
Bagging
Bayesion Belief Net
Decision Table
Markov Models

关联规则 Association Rule

“IF A CUSTOMER BUYS BREAD, SHE/HE’S 70% LIKELY OF BUYING MILK.”

关联规则挖掘

- 关联规则挖掘是一种流行的无监督学习技术，常用于商业领域帮助识别购物模式。
- 它也被称为“购物篮分析”（Market Basket Analysis）。
- 该方法有助于发现变量（商品或事件）之间有趣的关系（关联性）。



使用关联规则的商业案例

- **交叉销售**

  交叉销售是指向已经购买或表现出兴趣的客户推荐或销售额外产品或服务的做法。

  例如，客户购买了手机后，可以向其推荐手机壳、耳机或延保服务。

- **电商网站中的推荐系统、在线广告优化、产品定价和促销**

  推荐系统是一种根据用户偏好、行为或反馈，为用户提供个性化建议或推荐的系统。

  例如：**Amazon.com、Netflix、Spotify、YouTube、LinkedIn等平台都广泛应用推荐系统。**

- **零售门店设计**

  在零售业中，关联规则可以帮助优化门店布局，将相关产品（如超市中的面包和牛奶）摆放在一起，提升联动销售机会。

Market Basket Data

Transaction NO.	Item 1	Item 2	Item 3	…
1	Beer	Diapers	Chips	
2	Diaper	Orange		
3	Diaper	Milk		
4	Beer	Diaper	Orange	
5	Beer	Detergent		
…				

1	🍎 🍺 ⚪ 🍗
2	🍎 🍺 ⚪
3	🍎 🍺
4	🍎 🍐
5	🍼 🍺 ⚪ 🍗
6	🍼 🍺 ⚪
7	🍼 🍺
8	🍼 🍐

关联“规则”——标准格式

规则格式：如果{一组商品}→ 那么{一组商品}

IF

CONDITION (*C*)

RESULT (*R*)

If {beer} -> {diaper}

什么是有趣的关联?

用于规则 C -> R 的一些标准度量:

- 支持度 Support(R, C)： p(R&C)
  表示同时包含 R 和 C 的交易（“购物篮”）所占的比例。

- 置信度 Confidence(C->R)： p(R|C)
  表示在包含 C 的交易中，同时也包含 R 的比例。

- 提升度和杠杆率 Lift 和 Leverage(C->R)
  用于衡量规则的强度和有趣性。

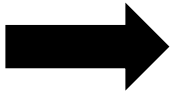
支持度 Support

支持度（Support）：衡量某个商品受欢迎程度的指标，表示包含该商品的交易（购物篮）所占的比例。

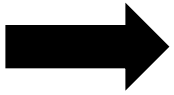
$$\text{Support}(X) = \frac{\#\text{ transactions that contain X}}{\#\text{ total transactions}}$$

支持度 Support

{  ,  } = 4

 Support = 4/10 = 40%

{  } = 5

 Support = 5/10 = 50%

置信度 Confidence

置信度（Confidence，C->R）：衡量关联规则成立的频率，表示在包含 C 的交易中，同时也包含 R 的比例。.

$$Confidence\ (C \rightarrow R) = \frac{Support\ (R, C)}{Support\ (C)}$$

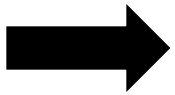
置信度 Confidence

IF  

Confidence = $\dfrac{\#\{\text{  }, \text{  }\}}{\#\{\text{  }\}} = \dfrac{4}{5} = 80\%$

Confidence for this association rule is the likelihood that a transaction contains  given that it contains 

IF 




80% Confidence …… Any problems?

如果有很多人购买尿布会怎样？？

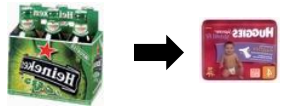
{  } = 8

80% Prevalence of 

……对于任何以尿布为结果项的商品组合（关联规则），其置信度都会很高。

重要度量: Lift (C->R)

提升度（Lift）：通过观察到的支持度与在 C 和 R 独立时的期望支持度之比来衡量。



$$\text{Lift} = \frac{p(R\&C)}{p(R)\,p(C)} = \frac{40\%}{80\%*50\%} = 1$$

对于关联规则来说，要有意义, the Lift must > 1

另一种度量方法：杠杆率（Leverage)

杠杆率（Leverage）：通过观察到的支持度与在 C 和 R 独立时的期望支持度之间的差值来衡量。.

Leverage = p(R&C) - p(R) p(C) = 40% - 40% = 0

对于关联规则来说，要有意义, the Leverage must > 0

练习 Exercise

IF 


What are the Confidence, Lift, and Leverage?

a) 50%, 1, 0
b) 50%, 1.2, 0
c) 70%, 0.8, 1
d) 40%, 0.75, 1
e) None of the above

关联规则用于超过两个商品之间

IF  + 

Support = 2/10

Confidence = 0.2/0.2 = 1

Lift = 0.2/0.2*0.8=1.25

Leverage = 0.2 – 0.2*0.8 = 0.04

如何发现“有趣”的关联规则?

通过为“有趣”的关联规则设定阈值

- 例如，支持度≥0.3，或置信度≥0.5，或两者都满足

关联规则学习算法中常见的策略有三个步骤：

1. 频繁项集生成：找到所有支持度大于最小支持度阈值的项集。
2. 规则生成：从频繁项集中提取所有高置信度的规则。
3. 规则检验：利用提升度/杠杆率去除偶然出现的虚假规则（确保不是巧合）。

如何更快地发现关联规则?

在现实世界中，商品的数量可能非常庞大。

例如，亚马逊有大约3亿种商品，淘宝有超过10亿个商品列表，Spotify有超过3000万首歌曲。

但是，你会尝试所有不同的商品组合吗?

对于大量变量来说，这样做计算成本太高了。

一个聪明的减少搜索空间的方法是：APRIORI算法。

APRIORI算法 (Optional)

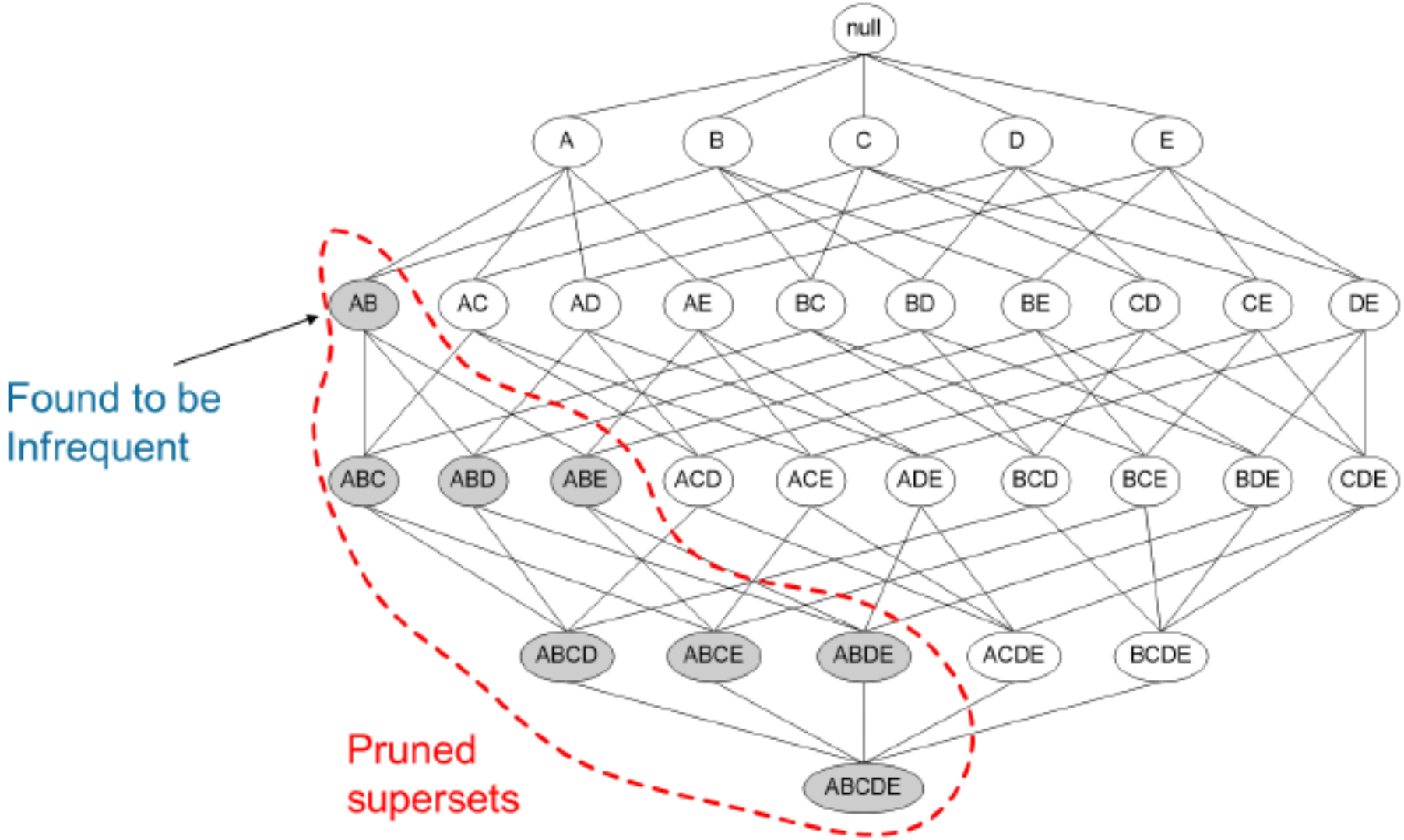
找到频繁项集：即那些支持度达到最小阈值的商品集合。

- 一个频繁项集的所有子集也必须是频繁项集。也就是说，如果{AB}是频繁项集，那么{A}和{B}也应该是频繁项集。
- 如果一个项集不是频繁项集，那么它的所有超集也都不是频繁项集。

设定最小支持度阈值，并从1项集到k项集（k-itemset）逐步迭代地寻找频繁项集。

利用这些频繁项集来生成关联规则。

APRIORI算法 (Optional)



APRIORI算法：一个例子 (Optional)

假设交易数据集由以下几个集合组成：{1,2,3,4}，{1,2}，{2,3,4}，{2,3}，{1,2,4}，{3,4}，以及{2,4}。

将频繁项集的最小支持度定义为3。.

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

Item	Support
{1,2}	3
~~{1,3}~~	~~1~~
~~{1,4}~~	~~2~~
{2,3}	3
{2,4}	4
{3,4}	3

{1,2,3} cannot be frequent
{1,2,4} cannot be frequent
{1,3,4} cannot be frequent
{2,3,4} can be a candidate

Item	Support
{2,3,4}	2

关联规则：其他应用场景

“商品”可以是任何特征：

- 拥有豪华车 => 频繁购买者
- 年龄（“30-39岁”）且收入（“42-48K”） => 购买（“汽车”）

从Facebook挖掘出的关联规则：

- 状态=本科生 & 政治观点=自由派

=> 感兴趣对象=男性 <提升度：(1.66)>

关联规则: 利与弊Pros and Cons

优点

- 可以快速挖掘描述业务/客户等的模式，无需在问题建模上投入大量精力。
- 是生成假设的无与伦比的工具。

缺点

- 目标不够明确。
- 不清楚如何具体应用挖掘出的“知识”。
- 可能会产生大量规则！
- 其中可能只有少数有价值的信息（甚至可能没有）。