Association Rules

Carlos Soares

[incluindo materiais gentilmente cedidos por Alípio Jorge, José Luís Borges e os que acompanham Han, Kamber & Pei]

U. PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

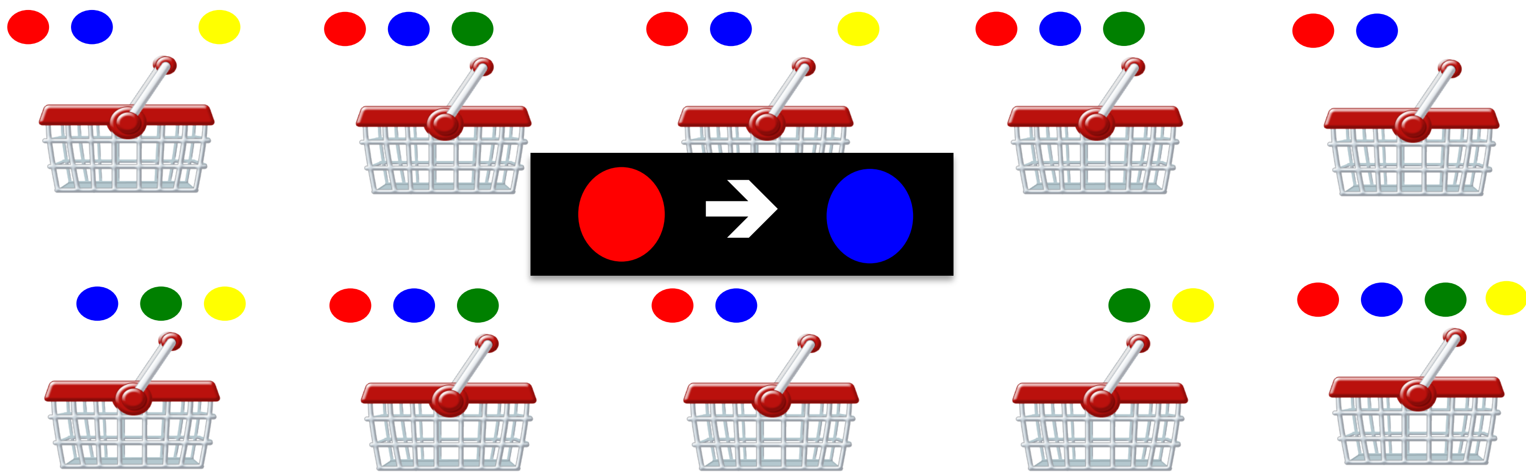
reference materials

U. PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- JMM et al. ch. 6

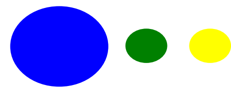
Market Basket Analysis

historical transactions



ongoing purchase

recommend which product?



Plan

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Association rules
- Modeling
- Evaluation
- APRIORI algorithm
- Patterns beyond AR

After the lesson you should be able to

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Identify problems in which association rules can be useful
- Understand basic concepts
  - item, transaction, basket, product
  - recommendation
  - frequent itemset
  - association rule
  - support, confidence and lift
- Perform association rule mining projects using RM
- Interpret and evaluate association rules

Association

http://www.flickr.com/photos/bitliukas/2118201782/

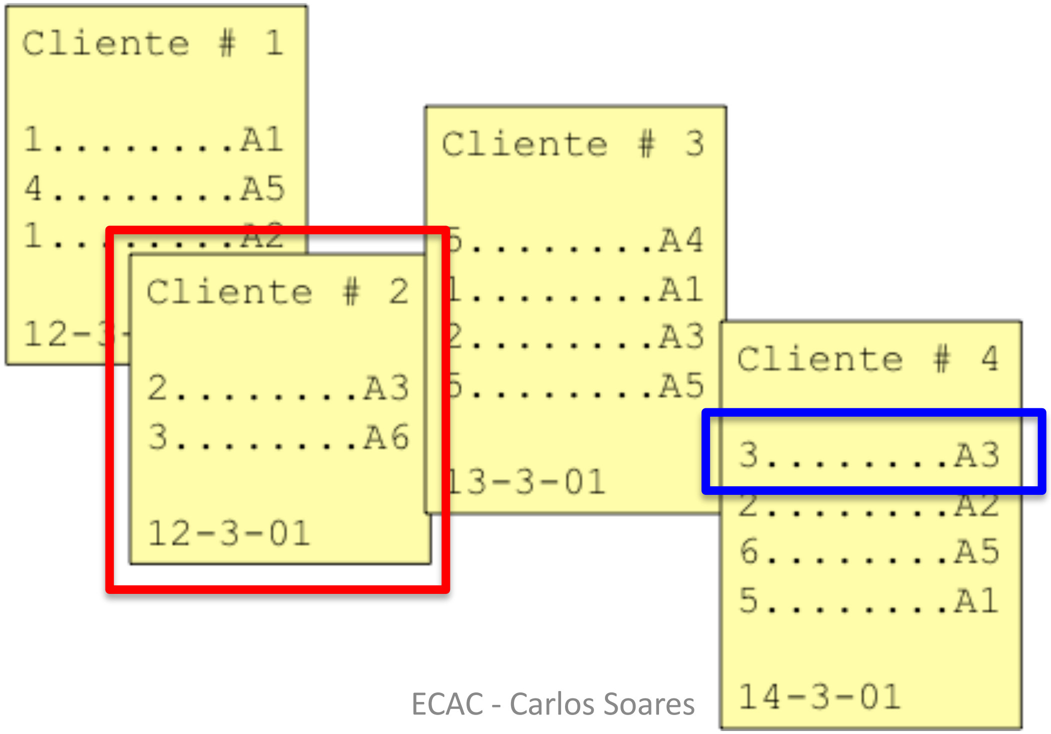
Association

≈

What goes with what

Given…

- set of objects
  - typically transactions or sets of transactions
- each object is also a set (of items)



... Association Model

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- identify sets of items (itemsets) that are typically associated in a transaction
- ... and itemsets that indicate other itemsets

- Typical **examples**
  – market baskets
  – credit card transactions
  – website clickstreams
  – DNA analysis

http://flic.kr/p/8ntacV

http://www.flickr.com/photos/amywatts/1331276401/

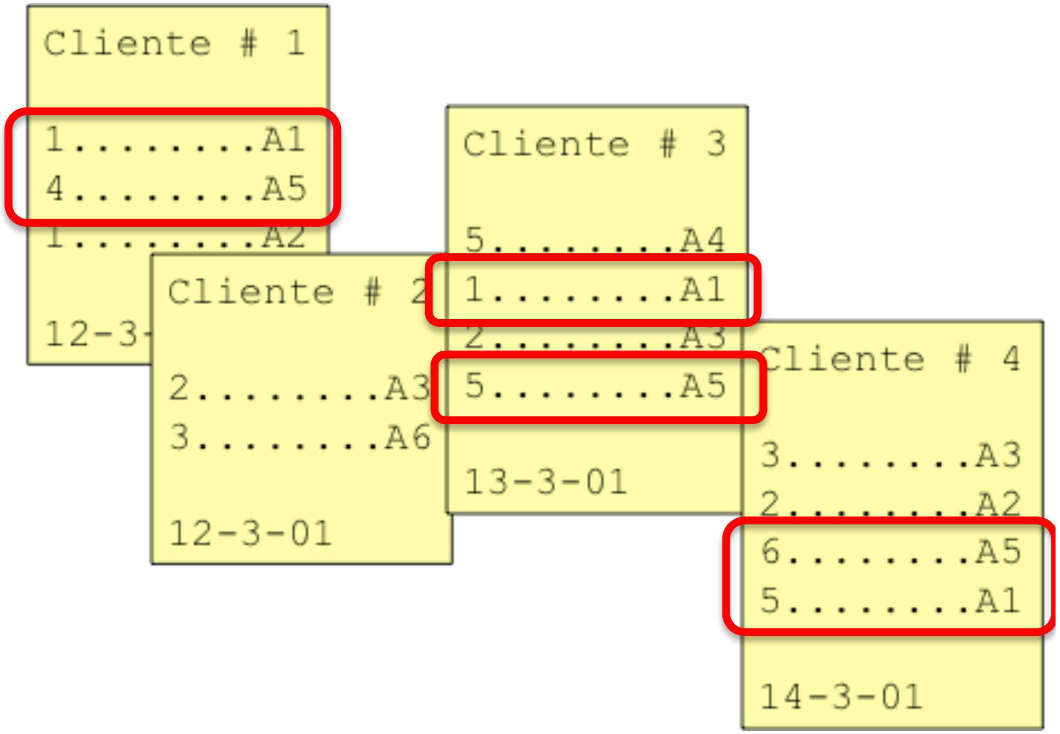
Plan

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Association Rules
- Modeling
  - frequent itemset mining
  - association rule mining
- Evaluation
- APRIORI algorithm
- Patterns beyond AR

Itemset Mining: Definition

- Given
  - A set of transactions $D=\{t_1, t_2, \ldots, t_n\}$
    - each t_i is a set of items/products
  - and a minimum support sup_{min} in [0,1]



- … find frequent itemsets
  - ALL sets of items X such that
  - support(X

Support =
relative frequency of the itemset in the transactions

Association Rules: Definition

Antecedent → Consequent

- Example: {A1, A2} → {A3}
  - read as: if the set of **items** A1 and A2 is observed, item A3 should also be observed
  - **support (same definition as for itemsets)**
    - percentage of baskets where co-occurrence is observed
    - estimates Prob(Antecedent $\cup$ Consequent)
      - example: Prob(A1 & A2 & A3)
  - **confidence**
    - percentage of cases where the occurrence of {A1, A2} correctly anticipates the occurrence of {A3}
    - estimates Prob(Consequent | Antecedent)
      - example: Prob(A3 | A1 & A2)

Association Rules: Interpretation

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- {A1, A2} ➔ {A3}
  - when **items** A1 and A2 are observed, item A3 is also expected

DANGER:

NOT NECESSARILY CAUSAL

- e.g. gas station

  { newspaper} ➔ { gas }

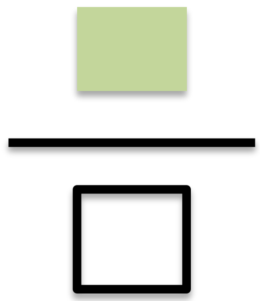
Mining Association Rules: Definition

- Given
  - a set of transactions $D=\{t_1, t_2, \ldots, t_n\}$
    - each t_i is a set of items/products
  - a minimum support sup_{min} in [0,1]
  - and a minimum confidence $conf_{min}$ in [0,1]

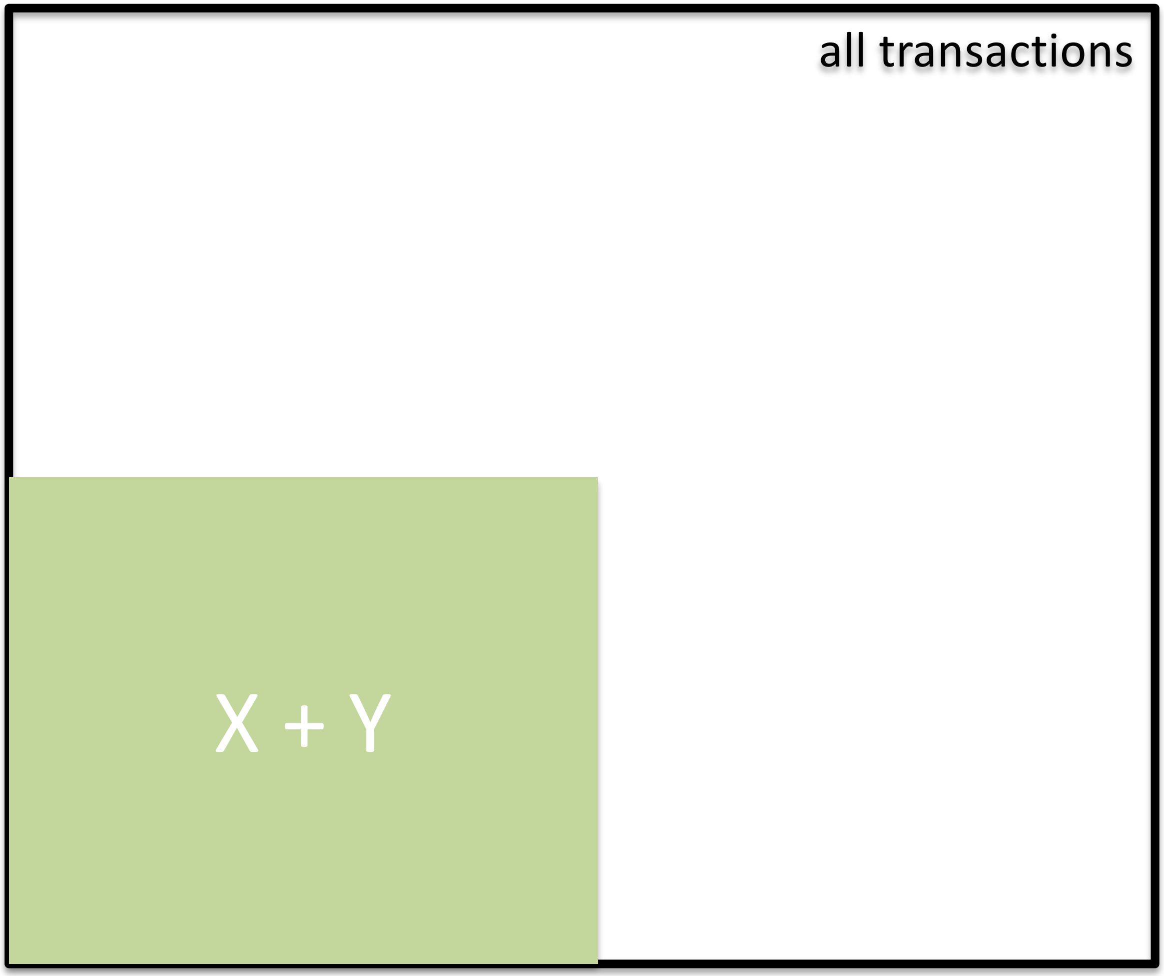
- … find
  - ALL rules $X \rightarrow Y$ where X and Y are itemsets such that
  - $support(X \rightarrow Y) >= sup_{min}$
  - $confidence(X \rightarrow Y) >= conf_{min}$

support vs confidence: support

- X → Y
- sup

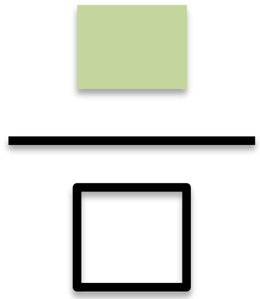


- conf

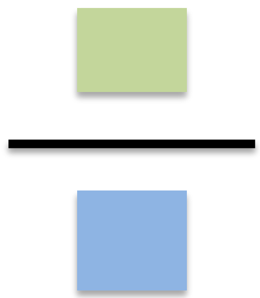


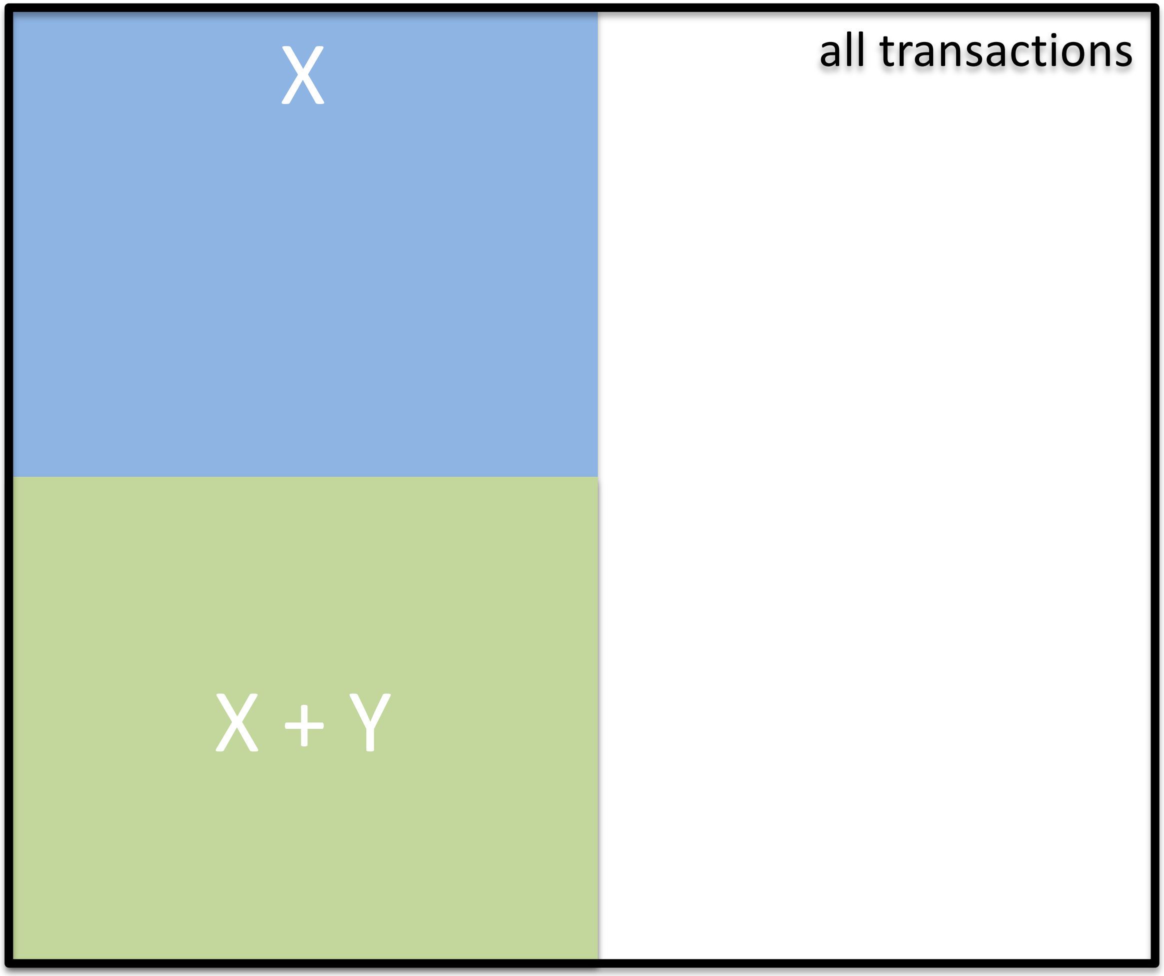
support vs confidence: confidence

- X ➔ Y
- sup



- conf

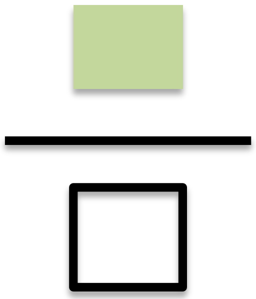




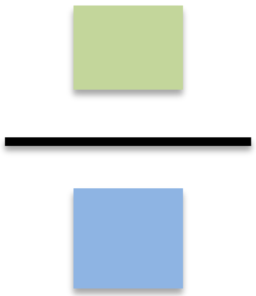
support vs confidence: what about Y?

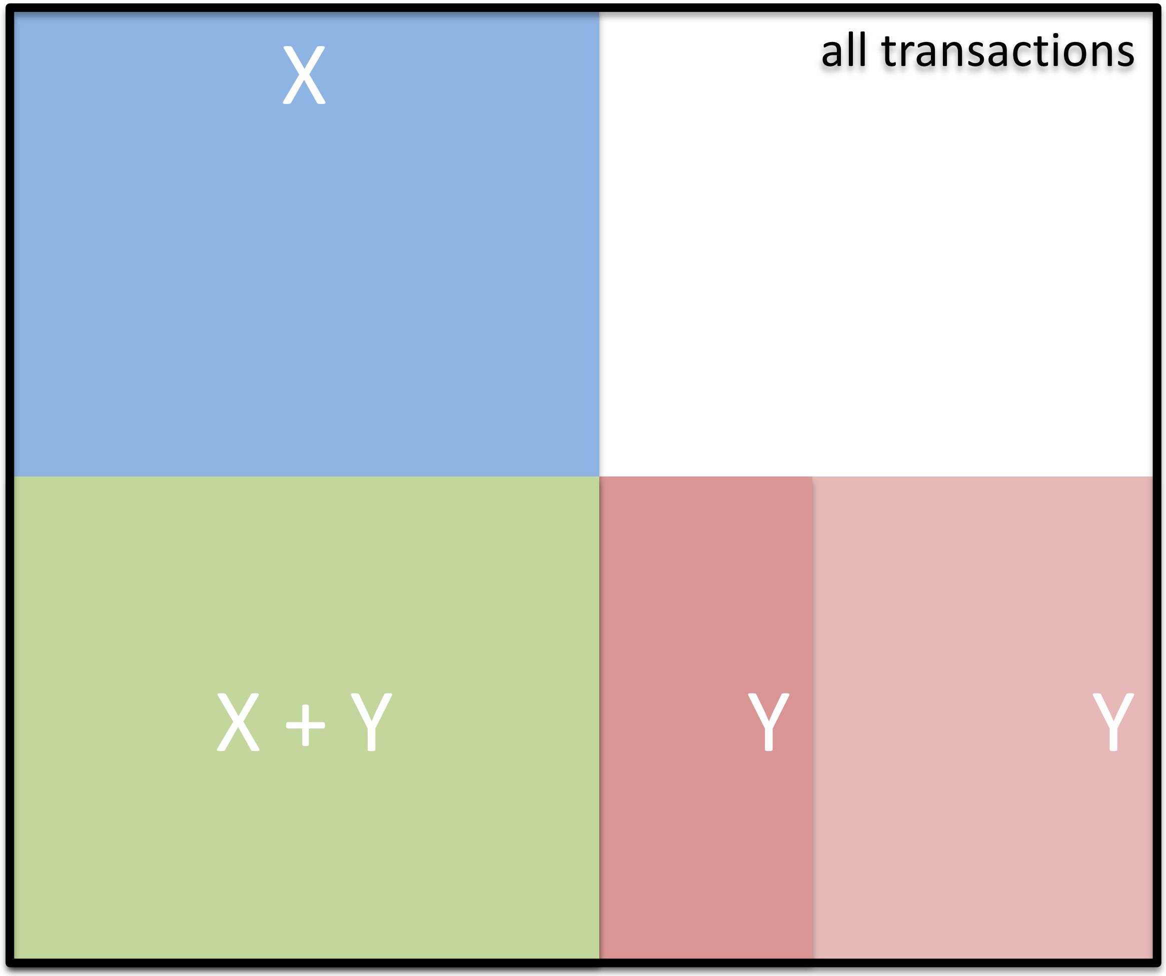
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- X → Y
- sup



- conf





Mining Association Rules: from Frequent Itemsets

- Given the frequent itemset

  {A,B,C}

  – and the support of all its subsets

- ...we can compute the confidence of the rules

  {B,C} ➔ {A}

  {A,C} ➔ {B}

  {A,B} ➔ {C}

  {C} ➔ {A,B}

  {B} ➔ {A,C}

  {A} ➔ {B,C}

$$\text{confidence}(Ant \rightarrow Cons) = \frac{\text{support}(Ant \cup Cons)}{\text{support}(Ant)}$$

Plan

- Association Rules
- Modeling
- Evaluation
  - assess the interest of a rule
  - how to apply rules
- APRIORI algorithm
- Patterns beyond AR

Rule Interestingness: in General

- Interesting rule [Silberschatz & Tuzhilin]
  - unexpected: *deviation from expected or believed*
  - useful (actionable): *estimated benefit*

- Subjective interest
  - interest depends on user knowledge
    - e.g. rule is unexpected
    - … identifies business opportunity
  - involves interesting items
    - focus on particular item (e.g., product)
- Objective interest
  - deviation from statistical independence
  - outstanding values

Trivial, inexplicable and useless rules rule... ☹

Rules Interesting and Statistical Independence (1/2)

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- A $\rightarrow$ B may have high support and confidence but is not interesting

  { jornal } $\rightarrow$ { combustível }

  sup = 5 %

  conf = 95 %

  is neither unexpected nor useful

- Typically
  - A $\rightarrow$ B is interesting if A and B are not statistically independent
  - if A and B statistically independent

  support(A $\cup$ B) $\approx$ support(A) $\times$ support(B)

  confidence(A $\rightarrow$ B) $\approx$ confidence(Ø $\rightarrow$ B)

$$\text{confiança}(A \rightarrow B) = \frac{\text{suporte}(A \cup B)}{\text{suporte}(A)}$$

Statistical Independence: an illustrative example

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

In which case does the presence of A give me any information about the presence of B?

A	B	C
X		X
X		
		X
		X
X	X	
X	X	X
	X	
	X	X

A	B	C
X		X
		X
		X
X	X	
X	X	X
X	X	
	X	X

Rules Interesting and Statistical Independence (2/2)

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- Types of measures
  - ratio

$$\boxed{\frac{probabilidade\ a\ posteriori}{probabilidade\ a\ priori} > 1}$$

$$\text{lift}(A \rightarrow B) = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)}$$

$$\text{conviction}(A \rightarrow B) = \frac{\text{support}(\neg B)}{\text{lift}(A \rightarrow \neg B)}$$

  - difference

$$\boxed{probabilidade\ a\ posteriori - probabilidade\ a\ priori > 0}$$

leverage

  - hypothesis tests

$$\boxed{probabilidade\ a\ posteriori >> probabilidade\ a\ priori\ ?}$$

$\chi 2$

Rule Interestingness: Lift

$$\text{lift}(A \rightarrow B) = \frac{\text{confiança}(A \rightarrow B)}{\text{suporte}(B)}$$

- measures “informativeness” of A relative to B
- lift = 1 means A and B are independent

{ jornal } $\rightarrow$ { combustível }

sup = 5 %

conf = 95 %

sup(B) = 95%

lift({ jornal } $\rightarrow$ { combustível }) = 1 , logo não acrescenta nada

Which Measure to Use When?

(Tan, Kumar, Sritastava @KDD' 02)

measure	range	formula
-coefficient	-1...1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Yule's Q	-1 ... 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Yule's Y	-1 ... 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
Cohen's	-1 ... 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
tsky-Shapiro's	-0.25 ... 0.25	$P(A,B)-P(A)P(B)$
tainty factor	-1 ... 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
dded value	-0.5 ... 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
losgen's Q	-0.33 ... 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P($
lman-kruskal's	0 ... 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\text{ma}}{2-\max_j P(A_j)-max_k P(B_k)}$
al Information	0 ... 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_J)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\text{lo}}$
J-Measure	0 ... 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
Gini index	0 ... 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2] -$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2$
support	0 ... 1	$P(A,B)$
confidence	0 ... 1	$max(P(B\|A), P(A\|B))$
Laplace	0 ... 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
Cosine	0 ... 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
ence(Jaccard)	0 ... 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
_confidence	0 ... 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
odds ratio	0 ... ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
Conviction	0.5 ... ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
lift	0 ... ∞	$\frac{P(A,B)}{P(A)P(B)}$
ctive strength	0 ... ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	0 ... ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

Plan

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

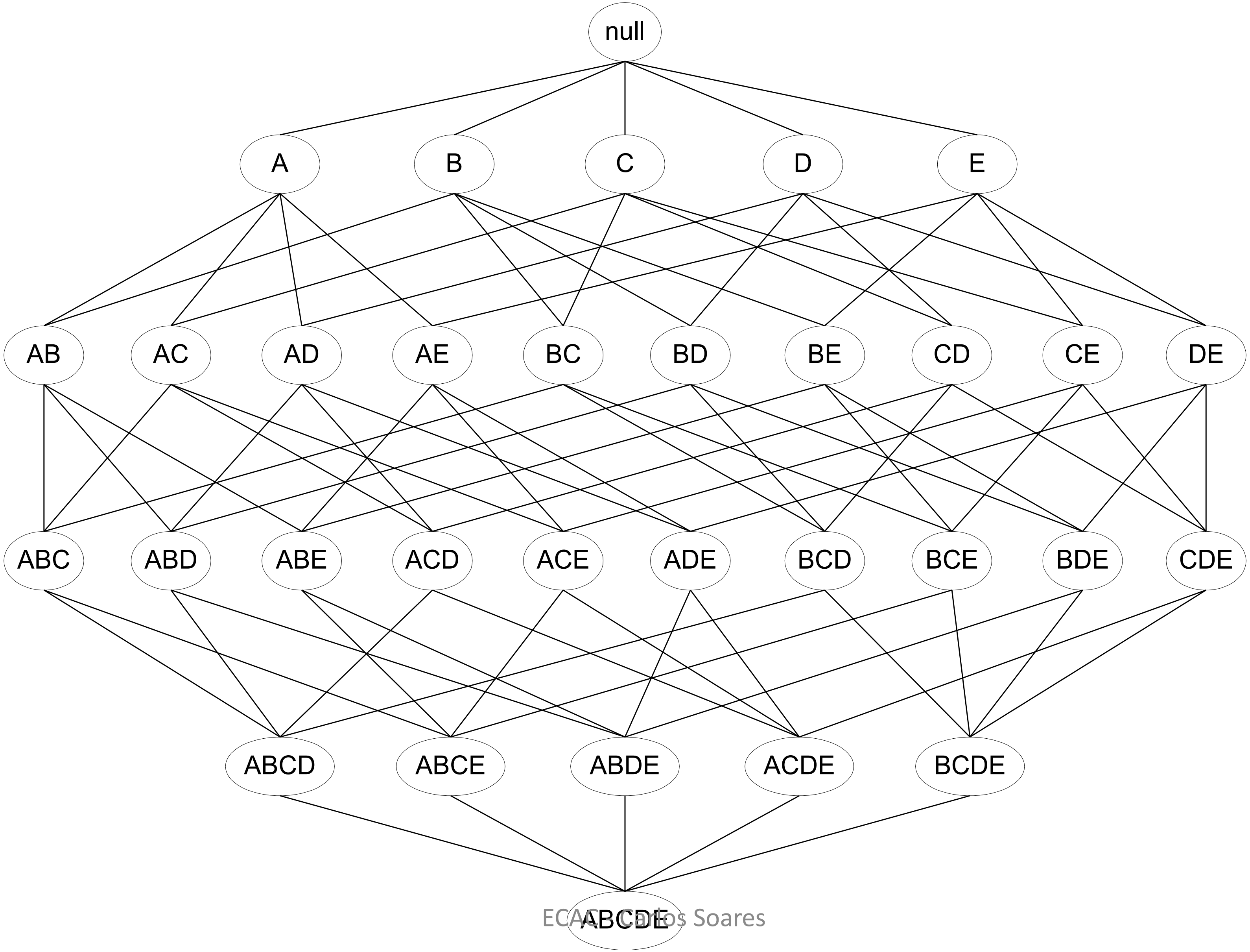
- Association rules
- Modeling
- Evaluation
- APRIORI algorithm
- Patterns beyond AR

APRIORI Algorithm

- Generation of Association Rules
  1. identify frequent itemsets
     - support $\geq sup_{min}$
  2. generate rules from frequent itemsets
     - confidence $\geq conf_{min}$

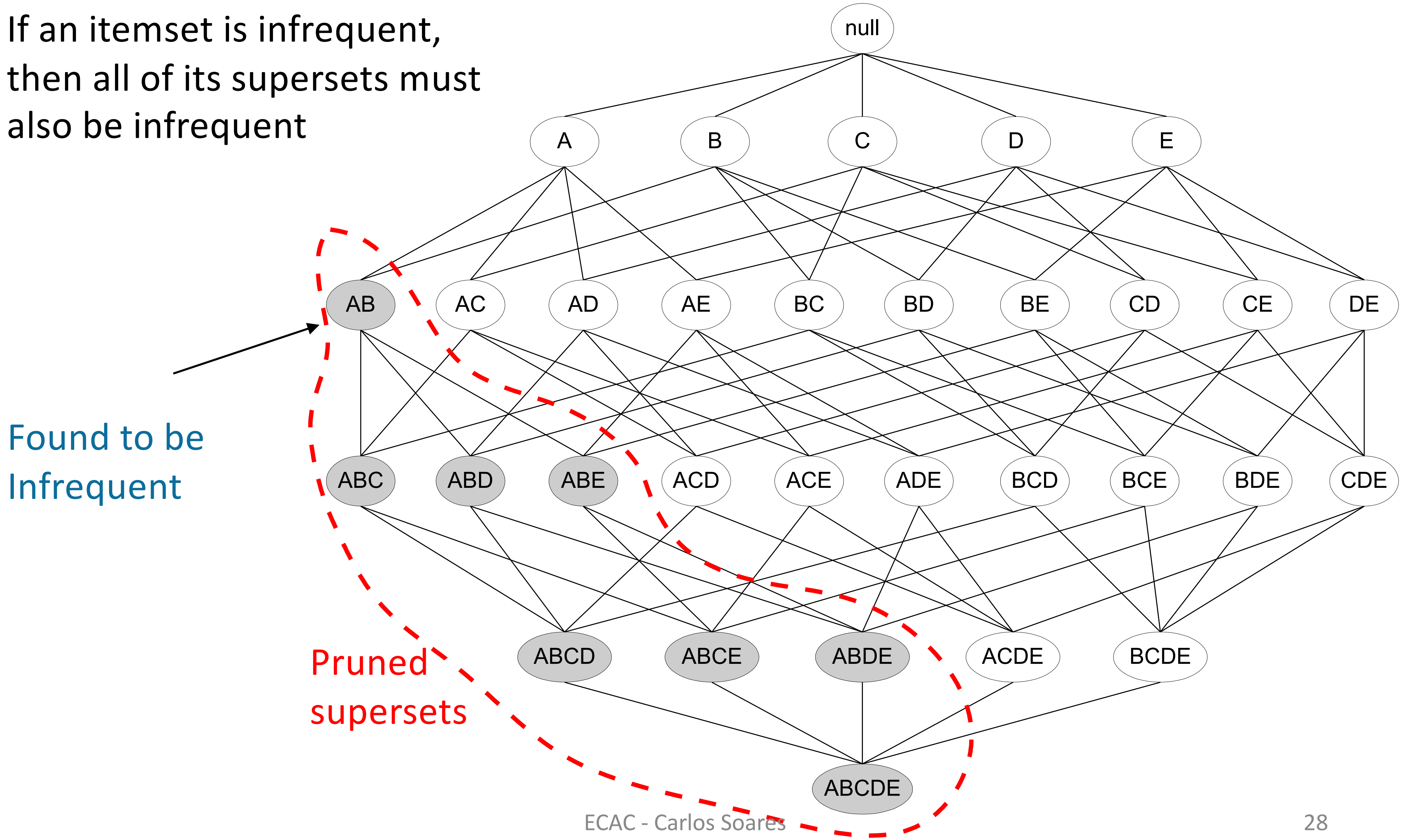
- Worst-case scenario
  - all subsets of all transactions are candidates
  - first step is the most important in AR mining
    - for computational reasons

Itemset Lattice for 5 products



Downward Closure to the Rescue!

If an itemset is infrequent, then all of its supersets must also be infrequent



APRIORI Example (1/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

A	B	C	D
1			
1	1	1	
		1	
1	1	1	1
	1		
1			1
1	1	1	
		1	1
1	1	1	

Pass 1

candidates

{A},{B},{C},{D}

$sup_{min} = 0.4$

(i.e., ~4)

frequent sets

{A},{B},{C}

new candidates

{A,B}

{A,C}

{B,C}

APRIORI Example (2/3)

A	B	C	D
1			
1	1	1	
		1	
1	1	1	1
	1		
1			1
1	1	1	
		1	1
1	1	1	

Pass 2

candidates

{A,B},{A,C},{B,C}

$sup_{min} = 0.4$

(i.e., ~4)

frequent sets

{A,B},{A,C},{B,C}

new candidates

{A,B,C}

APRIORI Example (3/3)

A	B	C	D
1			
1	1	1	
		1	
1	1	1	1
	1		
1			1
1	1	1	
		1	1
1	1	1	

Pass 3

candidates

{A,B,C}

$sup_{min} = 0.4$

(i.e., ~4)

frequent sets

{A,B,C} → new candidates: none

AR from Frequent Itemsets

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- For every frequent itemset X
  - For every non-empty subsets of X, A
    - A $\Rightarrow$ (X-A) is an AR if confidence is higher than the minimum

$$\frac{\text{support(X)}}{\text{support(A)}} \geq \text{conf}_{\text{min}}$$

- Efficient rule generation
  - confidence does not have an anti-monotone property
  - … except within the same itemset
    - i.e., confidence is non-increasing as number of items in rule consequent increases
  - given frequent itemset {A,B,C,D}

confidence(ABC $\rightarrow$ D) $\geq$ confidence(AB $\rightarrow$ CD) $\geq$ confidence (A $\rightarrow$ BCD)

APRIORI Issues

- Major computational challenges
  - Multiple scans of transaction database
  - Huge number of candidates
  - Tedious workload of support counting for candidates
- A few approaches
  - FPGrowth
  - ECLAT
  - CLOSET
  - MaxMiner

Beyond AR

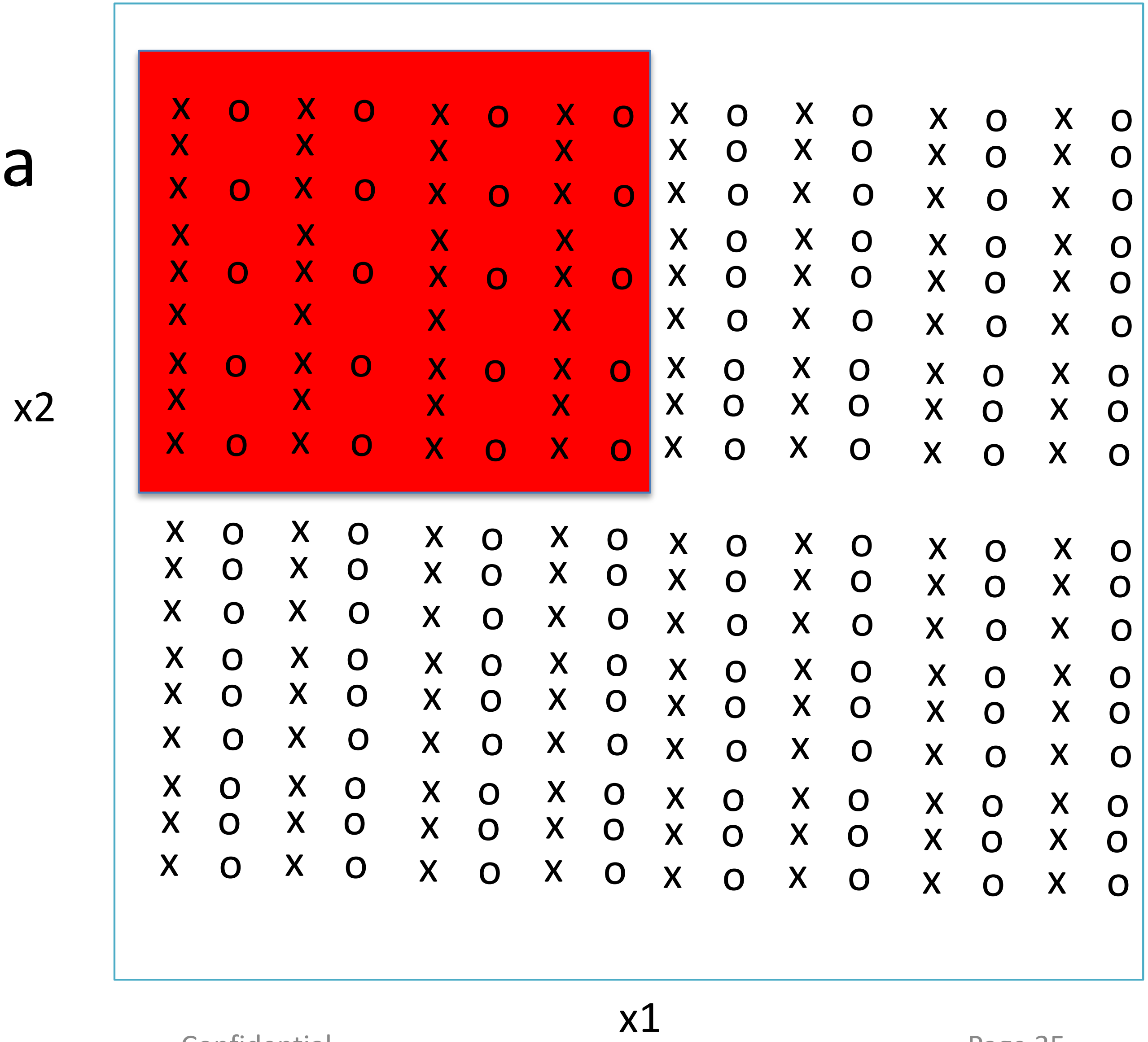
- Frequent pattern mining
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: discriminative, frequent pattern analysis
  - Cluster analysis: frequent pattern-based clustering
  - Data warehousing: iceberg cube and cube-gradient
  - Semantic data compression: fascicles

- Broad applications

beyond AR: subgroup discovery

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- find unusual distributions of a variable of interest
  - unusual = different from the population
  - variable of interest <> target
- … understandable



Further Readings

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Survey
  - Han, J., Cheng, H., Xin, D., Yan, X., 2007. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1), pp.55–86.
- Applications
  - Tom Brijs, Gilbert Swinnen, Koen Vanhoof, Geert Wets: Building an Association Rules Framework to Improve Product Assortment Decisions. Data Min. Knowl. Discov. 8(1): 7-23 (2004)
  - Singh, P., Thomas, A. C., and Sepulveda, A. 2006. Market basket recommendations for the HP SMB store. SIGKDD Explor. Newsl. 8, 1 (Jun. 2006), 57-64. DOI= http://doi.acm.org/10.1145/1147234.1147243