TÉCNICO+
FORMAÇÃO AVANÇADA

Pattern Discovery: Introduction

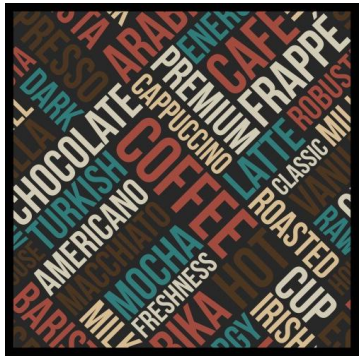
Classic stances on pattern and association rule mining

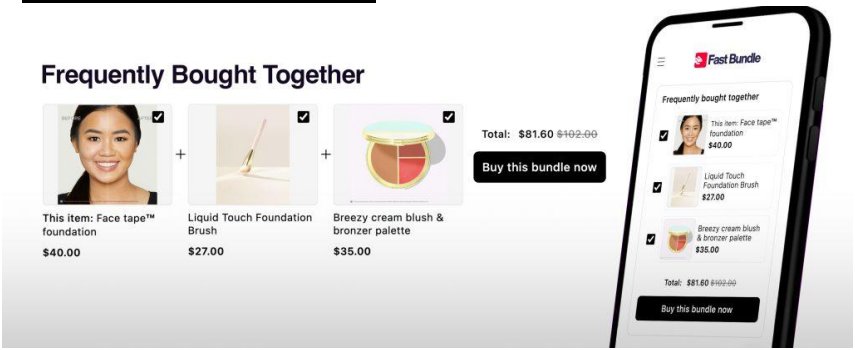
DASH: Data Science e Análise Não Supervisionada
Rui Henriques, rmch@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

Outline

- Pattern discovery
- Statistically significant patterns
- Positive and negative patterns
- Association rules
- Efficient pattern mining
- Quantitative and multi-level patterns
- Knowledge incorporation
- Final remarks

Patterns everywhere!



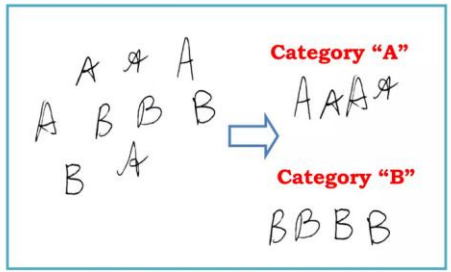


Pattern Recognition

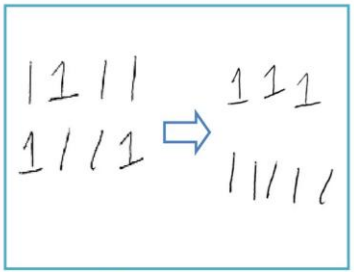
- **Patterns** are local relationships in data, often yielding informative and predictive power
  - *statistical regularities*: correlations, trends, discriminative associations…
  - *structural relationships*: sequences, hierarchies, spatiotemporal layouts (e.g. word order in text, shapes in images)…
  - *latent representations*: hidden features that efficiently summarize content…
- *Locality* in reference to the extent of the content space (e.g., subset of features, time, text, space)
- **Pattern recognition** underlies all learning tasks by humans and machines
  - *prediction*: patterning used for driving? Image classification? Question answering?
  - *description*: patterning used for clustering behaviors? Describing series? Detecting outliers?

Pattern Recognition

- Why patterns matter:
  - artificial agents do not understand meaning directly
  - patterns offer a way to minimize (descriptive or predictive) error on observed data
  - good generalization capacity if those patterns also hold in unseen data
    - *useful patterns* (signals) → generalize
    - *spurious patterns* (noise) → overfitting
- Patterns seem to be everything and nothing!
  - indeed *pattern recognition* became a proxy name to machine learning



Classification


Clustering

Pattern Discovery

- Patterns form the ground for **knowledge acquisition**
  - yet, to this end, we need to be able to access and interpret them
  - *problem*: many patterns in ML are not clearly defined or easily interpretable (e.g. latent patterning in neural networks)
- Two paradigms: recognition of **implicit** *versus* **explicit** patterns
  - for **descriptive ends**: we want to move toward *explicit patterns*
    - well-defined and human-understandable (easy to inspect and validate)
    - e.g. if–then rules in decision trees, handcrafted features (e.g. “knee angle > θ”), …
- **Pattern discovery**: subfield of ML dedicated to the retrieval of (explicit) patterns

Pattern Discovery

- From now on: **patterns** as local relationships in data that are both *explicit* and *relevant*
- What makes a pattern **relevant**?
  - **statistical significance** (≠ spurious)
  - **actionability** (ability to be used to guide decisions in a given domain)
  - **non-triviality** and **novelty** (can expand existing knowledge)
  - **informative power** (e.g. summarization capacity) and/or **predictive power**
  - **non-redundancy** (with other patterns in voluminous solutions)
- **Pattern mining**: discovery of all relevant patterns in a given dataset

Patterns everywhere!

- Different data structures ⇒ different patterns
  - **multivariate** patterns: co-occurring features, e.g. $(\text{age} > 50 \vee \text{BMI} > 35) \wedge \text{drug} = A$
  - **transactional** patterns: itemsets, e.g. $\{\text{milk}, \text{bananas}\}$
  - **vision** patterns: edges, textures, shapes, object parts
  - **language** patterns: syntactic structures, semantic associations, phrase usage
  - **temporal** patterns: trends, motifs, cycles, anomalies
  - and more…
    - *graph* patterns: subgraphs, e.g. social communities
    - *relational* patterns: associations spanning multiple tables, e.g. $\text{buys}(y, x) \wedge \text{retailer}(z, y)$
    - *spatiotemporal* patterns, *multi-event* patterns…

Why mining patterns?

- Acquiring **knowledge** from **non-trivial** and **actionable patterns**
  - *classic*: basket analysis, cross-marketing, web navigation, DNA sequence analysis, catalog design, sales campaign…
- Patterns further form the foundation for many essential data mining tasks
  - **multivariate association** and **causality analysis**
  - **feature extraction** from complex data
    - e.g. patterns in media, spatiotemporal, stream data as features
  - **prediction**: associative classifiers and regressors
  - **clustering**: pattern-based clustering (e.g. shapelet-based clustering)
  - semantic **data compression** (e.g. fascicles)

Applications

- **social networks:** communities of individuals with shared interests, correlated activity and/or coherent intercommunication; content aggregation from comments and tags
- **text data:** group content-related documents to support searches, suggestions and tagging
- **e-commerce:** browsing and shopping patterns
- **financial/trading:** profitability patterns for buy/hold/sell trading points
- **collaborative filtering:** groups of users with similar preference patterns
- **omics data:** functional processes and pathways
- **physiological data:** patient groups with coherent stimuli- or disease-conditional responses
- **clinical data:** risk profiles from health records
- **biological networks:** modules of correlated genes, proteins or metabolites

Outline

- Pattern discovery
- **Statistically significant patterns**
- Positive and negative patterns
- Association rules
- Efficient pattern mining
- Quantitative and multi-level patterns
- Knowledge incorporation
- Final remarks

Origins… basket analysis

- Some of the origins of pattern mining resort back to **market basket analysis**…
- A **transactional database** is a set of transactions
  - each transaction (basket) is a set of items
- Patterns often given by:
  - **frequent itemsets**, i.e. co-occurring items in a number or percentage of transactions
  - **association rules**, i.e. items that discriminate occurrence of other items
- Accordingly, frequent itemset mining (**FIM**) and association rule mining (**ARM**) aim at finding all those patterns

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

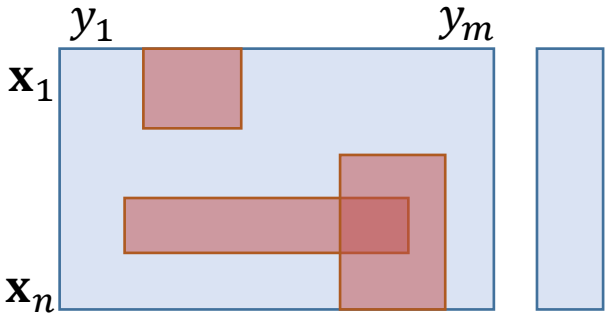
{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

… extended to multivariate data

Patterns given by:

- **value expectations on a subset of variables**
  - the pattern **coverage** is the set of observations satisfying those expectation
  - the pattern defines a **subspace**
- **association rules**, i.e. values on some variables that discriminate the values on other variables
  - **supervised setting**: input features in the antecedent and outcomes in the consequent
  - **unsupervised setting**: free associations between input features



Frequent *vs* relevant patterns

- Considering the itemset {water, potatos}
  - Can it be considered frequent if its support is 5%? 20% 50%?
  - What about the itemset {diapers, flowers}?

- **Challenge**: impossible to define a single support threshold as *item probabilities highly vary*!
  - same challenge in multivariate data (e.g. brown skin and blue eyes)
  - implication: **frequency is not good proxy for relevance**
    - e.g. the pattern “dog bites human” is as frequent as “human bites dog” on a given dataset, yet the latest is much less expected so surely more important!
- Replace frequency by **statistical significance**

Statistically significant patterns

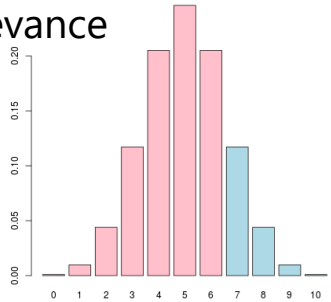
- Considering a transactional database of 1000 baskets
  - the probability of a user buying water is 30% and buying potatoes is 10% (probabilities directly estimated from the database)
  - what is the null joint probability of buying $\varphi = \{\text{water},\ \text{potatoes}\}$?
    Under independence assumption:
    - $p_{\text{null}}(\varphi) = p_{\text{null}}(\text{water} \cap \text{potatos}) = p(\text{water}) \times p(\text{potatoes}) = 0.3 \times 0.1 = 0.03$

- Assuming that we observe 35 baskets in the database with water and potatoes:
  Is $\{\text{water},\ \text{potatoes}\}$ statistically significant?
  - to answer this, we need to test "Given $N = 1000$, what is the probability of at least $n = 35$ users buying water and potatoes knowing $p_{\text{null}} = 0.03$?"

Statistically significant patterns

- Given $N = 1000$, what is the probability of observing at least $n = 35$ users buying water and potatoes with $p = 0.03$?
  - can be answered by computing the tail of a Binomial, i.e. $p(n > 35 \mid n \sim Binomial(N, p))$
  - $p(n \geq 35)$=0.15, as 0.15 is above reference significance levels (0.05 or 0.01) for statistical tests, we can say that is not unexpectedly low and therefore not relevant
  - if $n = 50$, then $p(n \geq 50)$=2.37E-4, hence $\{\text{water},\ \text{potatoes}\}$ is relevant in such dataset

- Summary: **joint probability** and **binomial testing** to assess relevance
  - How? scipy in python, excel…
    - 1 – cumulative Binomial with n, N and p parameters (also known as survival function)



Outline

- Pattern discovery
- Statistically significant patterns
- **Positive *vs* negative patterns**
- Association rules
- Efficient pattern mining
- Quantitative and multi-level patterns
- Knowledge incorporation
- Final remarks

Rare and negative patterns

- Why to solely pursue unexpectedly frequent patterns (aka *positive patterns*)?
- As we moved from clusters to anomalies…
  … we can move from unexpectedly frequent to unexpectedly infrequent patterns
- **Rare patterns**
  - very low support but interesting (e.g. buying diamond)
- **Negative pattern**
  - patterns that are unexpectedly rare (instead of frequent)
  - *example*: buying Ford Expedition (SUV car) and Toyota Prius (hybrid car) together is unlikely since Ford Expedition and Toyota Prius are negatively correlated
  - rare/infrequent negative patterns can even be more interesting than frequent ones!

Negative patterns

- **Support-based** definition
  - If patterns $P1$ and $P2$ are both frequent or significant, yet rarely occur together, i.e., $\sup(P1 \cup P2) \ll \sup(P1) \times \sup(P2)$
    - then $P1$ and $P2$ are negatively correlated
  - Example: a store sells needles A and B, with $p(A) = p(B) = 0.5$
    - assuming only one transaction contained both A and B
    - when there is a total of 200 observations, $\sup(A) \times \sup(B) = 0.25$ and $\sup(A \cup B) = \frac{1}{200} = 0.005 < \sup(A) \times \sup(B)$
- Other definitions available, e.g. **Kulzynski**-based definition
  - If patterns $P1$ and $P2$ are frequent, yet $(P(P1|P2) + P(P2|P1))/2 < \epsilon$, then $P1$ and $P2$ are negatively correlated

Outline

- Pattern discovery
- Statistically significant patterns
- Positive *vs* negative patterns
- **Association rules**
- Efficient pattern mining
- Quantitative and multi-level patterns
- Knowledge incorporation
- Final remarks

Association rules

- Recall the concept of an association rule, R: $A \Rightarrow B$
  - where A is the **antecedent** (set of features) and B is the **consequent** (set of features or outcomes)
  - if B is an outcome of interest, R is also termed **discriminative pattern**

- Considering the following transactional database
  - $\varphi = \{\text{Socks}, \text{Tie}\}$ is an itemset with coverage {T1,T2}, support 0.5, and p-value (Binomial test with n=2, N=4 and null probability $p(\varphi) = ¾ \times ¾ = 0.56$) of 0.4, hence not unexpectedly frequent
  - What about $\text{Socks} \Rightarrow \text{Tie}$? Is it relevant?

T1	Shoes,Socks,Tie
T2	Shoes,Socks,Tie,Belt,Shirt
T3	Shoes,Tie
T4	Shoes,Socks,Belt

Association rules in multivariate data

Association have been further classified into different categories:

- **Boolean** association rule

$$\text{Keyboard} \Rightarrow \text{Mouse}\ [\text{sup}=6\%, \text{conf}=70\%]$$

  - the choice in transactional, sequential and categorical multivariate data

- **Quantitative** association rule

$$\text{Age} \in [26,30] \Rightarrow \text{Cars} \in \{1,2\}\ [\text{sup}=3\%, \text{conf}=36\%]$$

  - the choice in numeric data

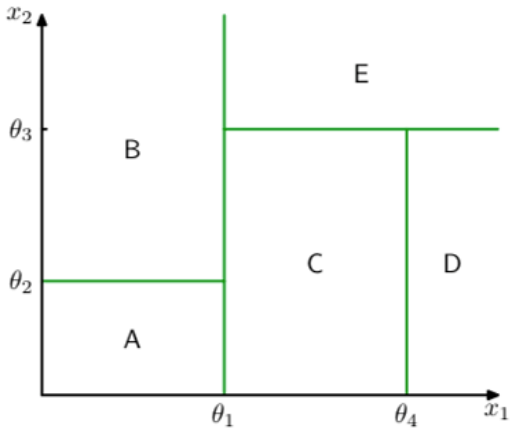
- **Hybrid** association rules

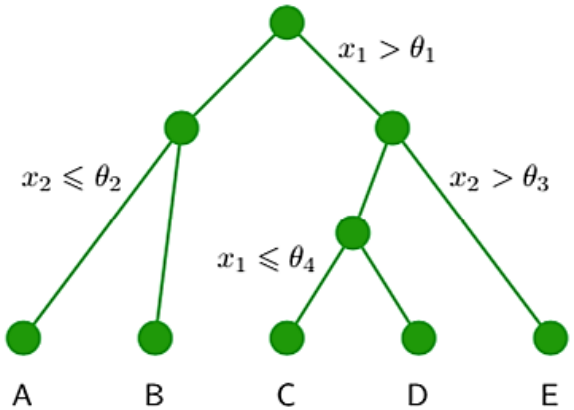
$$\text{Age} \in [26,30] \wedge \text{Keyboard} \Rightarrow \text{Mouse} \in \{1,2\}$$

  - the choice in mixed multivariate data, transactions with numeric outcomes, etc.

Association rules are already familiar…

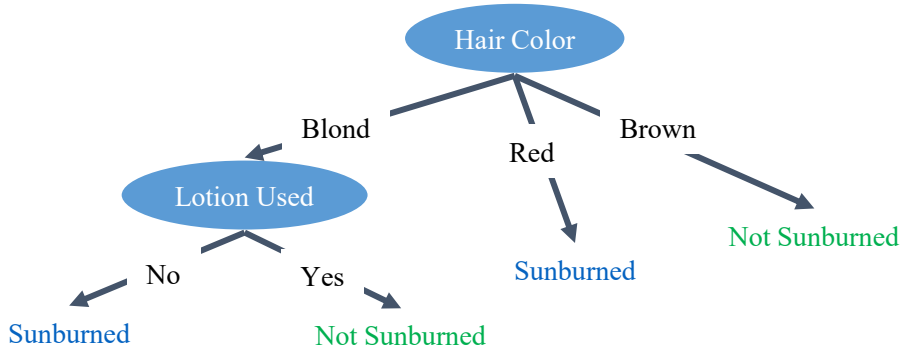
- Recall: **decision trees** for prediction
  - each path from root to leaf is an association rule
  - *classification* (classes on leaves) and *regression* (quantities on leaves)





Association rules are already familiar…

Name	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned (positive)
Dana	blonde	tall	average	yes	none (negative)
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none



If the person's hair is blonde and the person uses lotion then nothing happens

If a person's hair color is blonde and the person uses no lotion then the person turns red

If the person's hair color is red then the person turns red

If the person's hair color is brown then nothing happens

Relevant association rules

- Any given association rule, e.g. $\mathrm{A} \Rightarrow \mathrm{B}$, can be further characterized by its:
  - **support**, fraction of observations that satisfy the rule, i.e. $\sup(\mathrm{A} \Rightarrow \mathrm{B}) = \sup(\mathrm{A} \cap B)$
  - **confidence**, fraction of observations with the antecedent in which the consequent is also satisfied, i.e.

$$\mathrm{confidence}(\mathrm{A} \Rightarrow \mathrm{B}) = \mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{P(A \cap B)}{P(A)} = \frac{\mathrm{support}(\mathrm{A} \Rightarrow \mathrm{B})}{\mathrm{support}(\mathrm{A})}$$

- Recovering $\mathrm{R}: \{\mathrm{Socks}\} \Rightarrow \{\mathrm{Tie}\}$
  - support of R is 50% (2/4)
  - confidence of R is 66.67% (2/3)
  - are the observed support and confidence high enough? Is the rule relevant?

T1	Shoes,**Socks,Tie**
T2	Shoes,**Socks,Tie**,Belt,Shirt
T3	Shoes,Tie
T4	Shoes,Socks,Belt

Relevant association rules

- Example: among 5000 students
  - 3000 play basketball
  - 3750 eat cereal
  - 2000 both play basketball and eat cereal

	basketball	not basketball
cereal	2000	1750
not cereal	1000	250

  - play basketball ⇒ eat cereal [sup=40%, conf=66.7%]
    - *Misleading*! Overall percentage of students eating cereal is 75%, higher than 66.7%
  - play basketball ⇒ not eat cereal [sup=20%, conf=33.3%]
    - Far more *accurate*! Although lower support and confidence!
- How to more accurately measure a rule's relevance?

Interestingness: lift

- **Lift** measures the **discriminative power** (also known as **surprise**) of the rule

$$lift(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A) \times P(B)}$$

  - in contrast with confidence, takes into account the probability of the consequent
    - in fact, $\text{lift}(\text{A} \Rightarrow \text{B}) = \frac{\text{conf}(\text{A} \Rightarrow \text{B})}{\text{P}(\text{B})}$
  - lift < 1: A and B negatively correlated, if the value is less than 1
    - the higher, the greater the relevance of the rule
  - lift > 1: A and B are positively correlated
  - lift = 1: A and B are independent, $P(A \cap B) = P(A)P(B)$

Interestingness

- Classic pattern mining methods define simplistic thresholds:
  - FIM aims at finding all patterns satisfying a *minimum support* threshold
  - ARM aims at finding all rules satisfying a *minimum support* and *confidence*
    - rules satisfying both thresholds are called **strong**
- Modern pattern mining methods further consider:
  - *statistical significance* criteria
    - e.g. p-value<0.1 under binomial test from null data model)
  - *interestingness* criteria, e.g. lift>1.3…

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Rule	Support	Lift
X⇒Y	25%	2
X⇒Z	37.50%	0.9
Y⇒Z	12.50%	0.57

Interestingness

- many other possibilities…
- DISA is a Python package that implements most of them:

https://github.com/JupitersMight/DISA

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1 . . . 1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 . . . 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 . . . 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 . . . 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 . . . 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 . . . 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 . . . 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 . . . 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 . . . 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 . . . 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\Sigma_i P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i))}$
J	J-Measure	0 . . . 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 . . . 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 . . . 1	$P(A,B)$
c	confidence	0 . . . 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 . . . 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 . . . 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 . . . 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 . . . 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 . . . ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 . . . ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 . . . ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective 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	χ^2	0 . . . ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

Outline

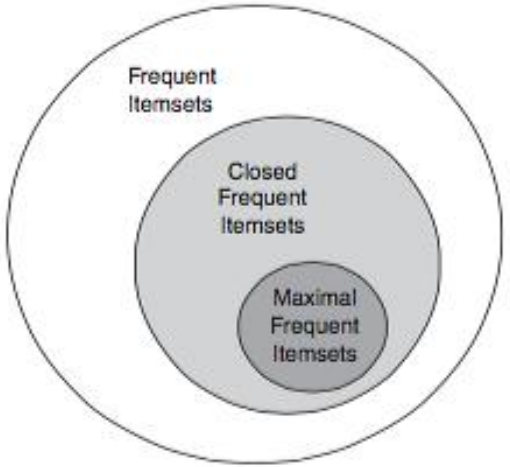
- Pattern discovery
- Statistically significant patterns
- Positive *vs* negative patterns
- Association rules
- **Efficient pattern mining**
- Quantitative and multi-level patterns
- Knowledge incorporation
- Final remarks

Pattern mining challenges

- There is usually a massive number of patterns on real-life databases
  - thousands or hundreds of thousands!
- How to handle large pattern sets?
  - pursuing **condensed representations**
  - returning **dissimilar patterns** only
  - **filtering less relevant** patterns
  - outputting **top patterns** only
  - using **background knowledge** to guide the mining process
    - focusing the process according to user expectations and available knowledge

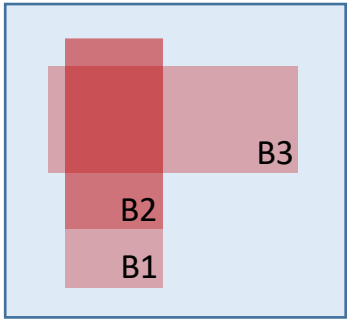
Condensed patterns

- A long pattern contains a combinatorial high number of subpatterns
  - Example: $\varphi = \{yi_{1}, \ldots, yi_{100}\}$ contains $\binom{100}{2} + \binom{100}{3} + \cdots + \binom{100}{99} = 1.27 \times 10^{30}$ sub-patterns!
- Solution:
  - Mine closed patterns or maximal patterns only
  - A pattern $P1$ is **maximal** if $P1$ is frequent and there exists no frequent super-pattern$P2 \supset P1$
  - A pattern $P1$ is **closed** if $P1$ is frequent and there exists no super-pattern $P2 \supset P1$, with the same support as $P1$



Condensed patterns: maximal and closed

- Given a pattern φ
  - a subspace with a smaller pattern, $|\varphi(B1)|<|\varphi(B3)|$, generally has more supporting observations (**vertical shape**)
  - a subspace with a larger pattern , $|\varphi(B3)|>|\varphi(B2)|$, generally has lower coverage (**horizontal shape**)
- In the right example:
  - Closed patterns? B1 and B3
  - Maximal patterns? B3
- **Closed** representations is a **lossless compression** of patterns
  - default representation when pattern mining
- **Maximal** pattern representations can cause the **loss** of relevant patterns
  - vertical shaped patterns are generally neglected in preference towards horizontal ones



Other patterns

- **Top *k* patterns**
  - compact pattern solution w.r.t. significance, interestingness, dissimilarity (redundancy-aware)
- **Generative** pattern-based models
  - concise **graphical models** that explain a set of patterns
- **Contrast patterns** from two datasets
  - pattern-centric highlighting of differences
- **Noisy** (fault-tolerant) **patterns**
  - supporting observations accommodate a parameterizable number or percentage of errors
- Mining truly interesting patterns:
  - incorporating (domain-driven) measures of **actionability**, **surprise**, **novelty**…
- Mining **emerging patterns**:
  - considering recency weighting in stream or temporal data

Efficient pattern mining

- Exhaustive discovery of patterns is a highly computational heavy task
- Principle to boost efficiency: **downward closure property**
  - any subset of a frequent pattern is also frequent
  - if {beer, diaper, nuts} is frequent, so is {beer, diaper}
    - i.e., transactions with {beer, diaper, nuts} also contain {beer, diaper}
- Three major approaches mining approaches (details in *Appendix*)
  - **Apriori** family (Agrawal and Srikant)
  - **FP-growth** family (Han, Pei & Yin)
  - **Vertical** family (Charm and Zaki)

Association rule mining

- Association rules can be exhaustively generated from the found patterns with PM approaches
  - efficiently combine disjoint patterns on the antecedent/consequent
  - efficiently test interestingness criteria
- Upper bounds to the pattern length in the consequent can be imposed
- What if we have outcomes of interest?
  - **patterns** can be found within each **outcome-conditional dataset**
    - yet one pattern can be statistically significant for multiple outcomes
    - *solution*: post-test the **discriminative power** of each rule $P \Rightarrow outcome$ (e.g. lift)
  - straighforward PM extension to impose variables of interest to appear in the consequent

Colossal patterns

- **Colossal patterns** are *lengthy* patterns
  - generally yield higher importance than small patterns
- Many real-world tasks need mining colossal patterns
  - **high-dimensional data domains** (biomedicine, text/media, social networks…)
- Problem: a subpattern of a frequent pattern is frequent, data with lengthy patterns have an **explosive set of patterns** 😐
  - the downward property of classical approaches is insufficient
    - whether we use breadth-first search (e.g. Apriori) or depth-first search (e.g. FPgrowth)
  - closed/maximal patterns partially alleviate, yet not solve
    - still need to find scattered large patterns

Pattern fusion

- *Philosophy*: jump out of the swamp of mid-sized patterns and quickly reach colossal patterns
  - mid-sized patterns are called **core patterns**
  - a colossal pattern is composed by multiple **core patterns** and a few **small-sized patterns**

- **Pattern Fusion** is an approach to this end using tree structures
  - traverses the tree in a bounded-breadth way, only expands a bounded-size candidate pool
    - only a fixed number of patterns are used as starting nodes — avoiding the exponential search space
  - identifies “shortcuts” whenever possible (e.g. agglomeration of patterns in the pool)
    - pattern fusion shortcuts will direct the search down the tree much more rapidly towards the colossal patterns

Outline

- Pattern discovery
- Statistically significant patterns
- Positive *vs* negative patterns
- Association rules
- Efficient pattern mining
- **Quantitative and multi-level patterns**
- Knowledge incorporation
- Final remarks

Quantitative and multi-level associations

- Major problems of classic pattern mining?
  - unable to deal with **numeric data**
    - solution: discretization
    - problems?
      - discretization is susceptible to item-boundary problems
      - loss of expressivity
  - unable to handle with the **multi-level** and **multi-dimensional** structure of data
    – e.g. different types of potatoes, different milk brands – generalize or specialize?
- Handling these problems
  - **biclustering** (*next class*!) and **quantitative associations**
  - multi-level and multi-dimensional associations

Quantitative associations

- Simplistic preprocessing strategies
  - **static discretization** based on predefined concept hierarchies (manual)
  - **dynamic discretization**
    - based on variable distribution (simple)
    - such that the confidence or compactness of the rules mined is maximized (advanced)
  - **one-dimensional clustering**: distance-based association
- Under these strategies, classic pattern mining is applied on the found numeric ranges
  - problems of discretization?
  - what about non-constant patterns?
  - to be covered in the next class

Quantitative associations

- Two major forms of association rules:
  - categorical ⇒ quantitative rules **or** quantitative ⇒ quantitative rules
  - e.g. education in [14-18] (yrs) ⇒ mean wage = $11.64/h
- Finding extraordinary and therefore interesting phenomena, e.g.

$$\text{sex} = female \Rightarrow \text{wage: } mean = \$7 \ (overall\ mean = \$9)$$

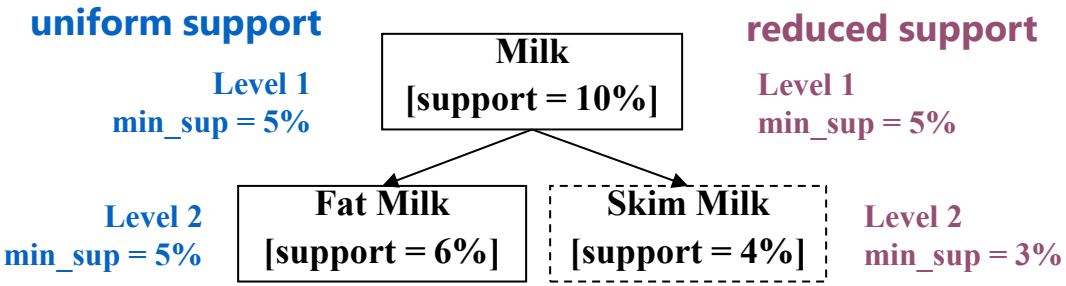
  - LHS: a subset of the population
  - RHS: an extraordinary behavior of this subset
- The rule is accepted if a statistical test (e.g. Z-test) confirms inference with high confidence
- Subrule: highlights the extraordinary behavior of a pop. subset of the super rule, e.g.

$$(\text{sex} = female) \wedge (\text{south} = yes) \Rightarrow \text{mean wage} = \$6.3/\text{h}$$

- Open problem: efficient methods for LHS containing two or more quantitative attributes

Multi-level associations

- **Items** in transactional databases or **categories** in multivariate data often form **hierarchies**
- Flexible support settings: items at the lower level are expected to have lower support



- some rules may be redundant due to "**ancestor**" relationships between items. Example
  - milk ⇒ wheat bread [support = 8%, confidence = 70%]
  - fat milk ⇒ wheat bread [support = 2%, confidence = 72%]
- we say the first rule is an **ancestor** of the second rule
  - a rule is **redundant** if its support is close to the "*expected*" value based on rule's ancestor

Multi-dimensional associations

- Single-dimensional rules:
  - $\text{buys}(X, milk) \Rightarrow \text{buys}(X, bread)$
- **Multi-dimensional rules**: ≥2 dimensions or predicates
  - *Inter-dimension* association rules (no repeated predicates)
    - $\text{age}(X, [19,25]) \wedge \text{occupation}(X, student) \Rightarrow \text{buys}(X, coke)$
  - *Hybrid-dimension* association rules (repeated predicates)
    - $\text{age}(X, [19,25]) \wedge \text{buys}(X, popcorn) \Rightarrow \text{buys}(X, coke)$
- How to? Data cube approaches

Outline

- Pattern discovery
- Statistically significant patterns
- Positive *vs* negative patterns
- Association rules
- Efficient pattern mining
- Quantitative and multi-level patterns
- **Knowledge incorporation**
- Final remarks

Constrained pattern mining

- Finding all the patterns in a database autonomously? Unrealistic!
  - the patterns could be too many but not focused
  - what makes a truly interesting pattern?
- Available **domain knowledge** can be used to **guide the pattern discovery**
  - ideally we can formalize background knowledge using constraints
- A **constraint** is a predicate in the data space
- Given a dataset and a set of constraints C, the problem of **constrained pattern mining** is the discovery of all relevant patterns satisfying C
- Simplistic examples:
  - **metric** constraints, e.g. $\text{lift} > \theta$
  - **value/item** constraints, e.g. specific features to be included/excluded from patterns

Constrained pattern mining

- **Meta-rule constraints**
  - partially instantiated predicates and constants, example:
    - $P1(x, Y) \wedge P2(x, W) \Rightarrow \text{buys}(x, iPad)$
    - pattern satisfying this constraint: $age(x, [15,25]) \wedge job(x, student) \Rightarrow \text{buys}(x, iPad)$
  - *How*? push constants deeply when possible into the mining process
- Other constraints:
  - knowledge type constraint: known classification, association, etc.
  - query constraint: e.g. “find product pairs sold together in stores in Chicago”
  - dimension/level constraint: in relevance to region, price, brand, customer category
- Recall: data mining should be an **interactive process**
  - user directs what to be mined using a data mining query language

Constrained pattern mining

- Domain-driven constraints:
  - *user flexibility*: provides constraints on what to be mined
  - *system optimization*: explores such constraints for efficient mining. *How?* Below
- **Pattern space pruning constraints**
  - *anti-monotonic*: if constraint c is violated, its further mining can be terminated
  - *monotonic*: if c is satisfied, no need to check c again
  - *succinct*: c must be satisfied, so one can start with the data sets satisfying c
  - *convertible*: c can be mapped into (anti-)monotonic c when values are properly ordered
- **Data space pruning constraints**
  - *data succinct*: data space can be pruned at the initial pattern mining process
  - *data anti-monotonic*: if an observation does not satisfy c, it can be pruned

Outline

- Pattern discovery
- Statistically significant patterns
- Positive *vs* negative patterns
- Association rules
- Efficient pattern mining
- Quantitative and multi-level patterns
- Knowledge incorporation
- **Final remarks**

Remarks

- Until here… we have tapped into patterns in multivariate and transactional data structures
- Two notes of care
  - check if pattern discovery offers a way to answer your problem
    - remember that pattern discovery is primarily used for **descriptive ends** ⇒ check next slide
    - although it can be used for **predictive ends** ⇒ check next class (*associative learning*)
  - methods can be hard to scale for high-dimensional data with extensive feature dependencies
    - practical principles to boost the search
      - reduce dimensionality
      - handle redundant variables (remove, combine…)
      - remove highly frequent categories
      - when discretizing data opt for *qcut* (equal frequency) over *cut* (equal width)
      - …

Remarks

Check if *explicit* patterns are the best solution to our problem! An illustrative **example**:

- **Recommendation** problems span our daily lives: shopping, media, webpage, document suggestions
  … and pattern discovery can be used to retrieved frequently co-accessed, co-bought, co-liked items ☺
- Yet… not all recommendation tasks need to be answered using pattern discovery
  - we need to separate *description* (knowledge acquisition) from *prediction*
  - user-specific recommendations can be answered using predictive tasks
    - predict a set of outputs from a set of input features that characterize user behavior
      - each output corresponds to a measure interest (binary/numeric) of a user on a given item
      - items/outputs are then ranked based on the estimated interest to deliver recommendations
  - what about recommendations not custom to single users (e.g. market shelf display, privacy-aware web navigation rules…)? Yes, patterns offer here important rules and descriptive insights!
- Yet our preferences change over time, so…

Complex data structures

Details in our **next class!**

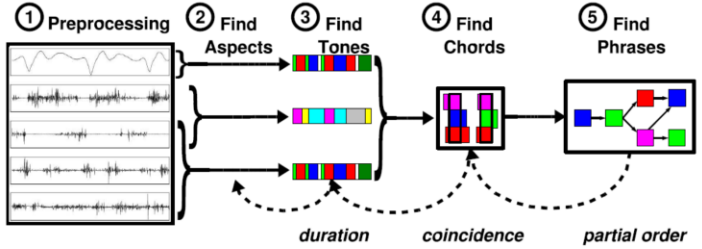
- Moving beyond multivariate and transactional data…
- Recall: different data structures ⇒ different patterns
  - **temporal** patterns: trends, motifs, cycles, anomalies
  - **vision** patterns: edges, textures, shapes, object parts
  - **language** patterns: syntactic structures, semantic associations, phrase usage
  - **graph** patterns: subgraphs, e.g. social communities
  - **relational** patterns: associations spanning multiple tables, e.g. $\text{buys}(y, x) \wedge \text{retailer}(z, y)$
  - **spatiotemporal** patterns: flocks, trajectories, colocation
  - **multi-event** patterns: episodes, dependency graphs

Complex data: a snippet

Details in our **next class!**

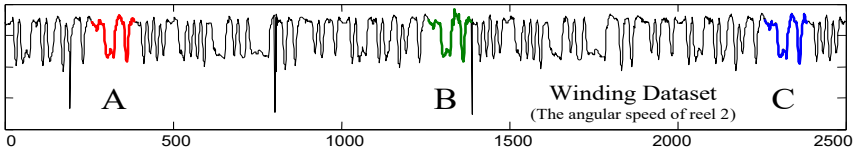
A *snippet* view on patterns in temporal data structures where **frequency** can be considered…

- *across* time series observations
  - **sequential pattern mining** (frequent orders)
  - **biclustering** (univariate time series data)
  - **triclustering** (multivariate time series data)
  - **temporal association rules**



- *within* a single time series
  - **motif discovery**
  - **predictive rule mining** A $\Rightarrow^{\Delta t}$ B

    once antecedent is observed, consequent expected within interval Δt



Thank you!

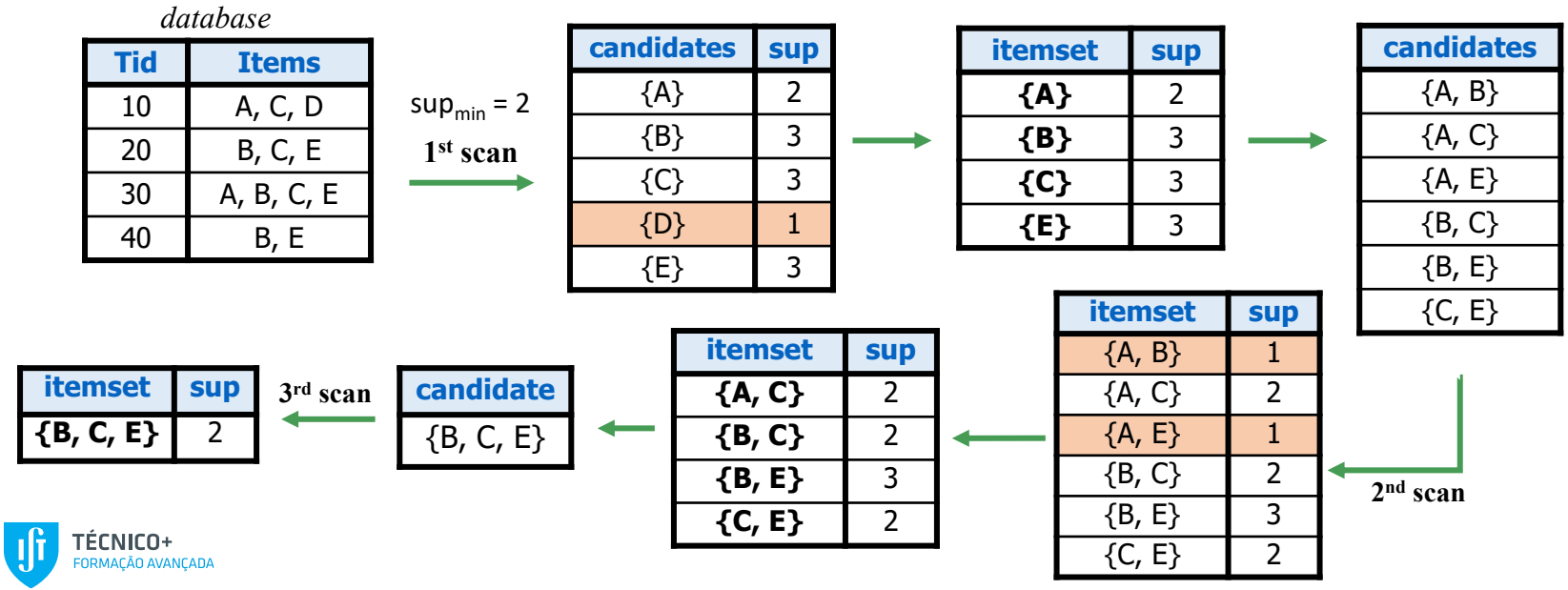
Rui Henriques

rmch@tecnico.ulisboa.pt

APPENDIX

Apriori

- **Principle**: if a pattern is infrequent, its superset is infrequent (should not be generated)
- **Method**: iteratively increase (k+1)-length candidate patterns from k-length frequent patterns



Improving Apriori

- **Challenges**
  - multiple scans of transaction database
  - huge number of candidates
  - workload of support counting for candidates
- **Improving** Apriori: general ideas
  - reduce passes of transaction database scans
  - shrink number of candidates
  - facilitate support counting of candidates
  - implementation available for relational databases using object-relational extensions
  - distribution and parallelization

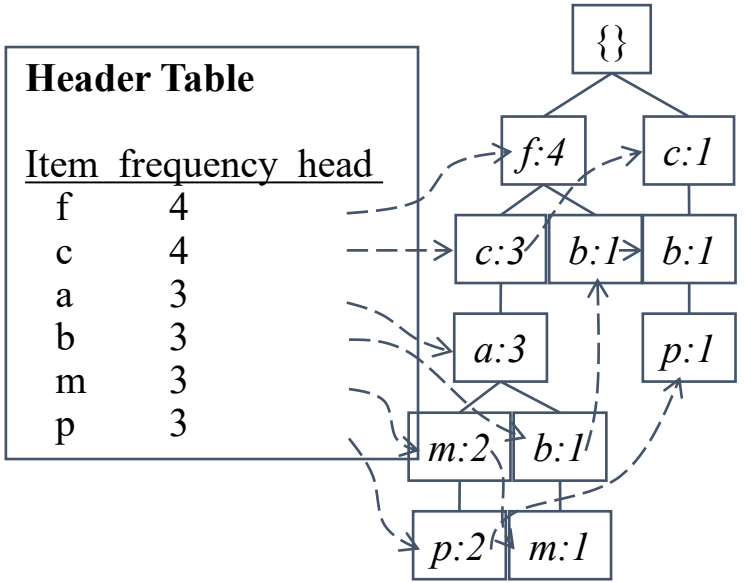
FP-Growth

- Bottlenecks of Apriori
  - multiple database scans are costly
  - long patterns generate lots of candidates
    - recall $\varphi = \{y_{i1}, y_{i100}\}$ contains 1.27×10^{30} patterns
- Can we avoid candidate generation?
  - Yes! using Frequent Pattern (FP) tree structure (avoid further data scans) and pattern-conditional trees to efficiently grow patterns
- FP-Growth approach
  - for each frequent itemset, construct its conditional pattern-base and FP-tree
  - repeat the process on each newly created conditional FP-tree

FP-Growth

TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

- divide-and-conquer
  - decompose the mining task according to the frequent patterns obtained so far
  - leads to focused search of smaller databases
- no candidate generation, no candidate test
- no repeated scans of entire database
- compressed database: FP-tree structure
- efficient counting operations on the FP-tree



Vertical approaches

- **Problem**: high-dimensionality
  - Apriori and FP-Growth are computationally heavy for spaces with thousands of variables
- Solutions:
  - **scalability** extensions
    - **vertical partitioning** of the dataset, followed by efficient pattern fusion
    - **parallelization** of operations
  - **vertical** pattern mining **approaches** (e.g. ECLAT) – flip/transpose the view:
    - items and categoric features have a list of observation IDs
    - pattern growth by intersecting observation IDs (instead of large sets of items)

Mining maximal patterns

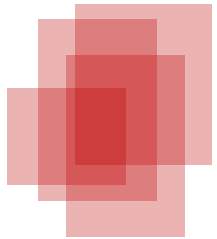
- 1st scan: find frequent items
  - A, B, C, D, E
- 2nd scan: find support for
  - AB, AC, AD, AE, ABCDE
  - BC, BD, BE, BCDE
  - CD, CE, DE, CDE

ID	items
10	A,B,C,D,E
20	B,C,D,E,
30	A,C,D,F

- Since BCDE is a max-pattern, no need to check BCD, BDE, CDE in later scan
- Principles combined in Apriori and FP-Growth approaches

Mining closed patterns (CLOSET)

- Recall: closed representations are lossless
  - useful to remove subspaces contained in large ones
  - relevant for large patterns and spaces with high homogeneity
- List of all frequent items in support ascending order, e.g. *d-a-f-e-c*
- Divide search space
  - patterns having *d*
  - patterns having *d* but no *a*, etc.
- Find frequent closed pattern recursively
  - every transaction having *d* has *cfa,* hence *cfad* is a closed pattern



min support =2

ID	items
10	a, c, d, e, f
20	a, b, e
30	c, e, f
40	a, c, d, f
50	c, e, f