# Data Mining and Knowledge Discovery

Petra Kralj Novak

January 8, 2020

http://kt.ijs.si/petra_kralj/dmkd3.html

# So far …

- Nov. 11. 2019
  - Basic classification
  - Orange hands on data visualization and classification
- Dec. 11 2019
  - Fitting and overfitting
  - Data leakage
  - Decision boundary
  - Evaluation methods
  - Classification evaluation metrics: confusion matrix, TP, FP, TN, FN, accuracy, precision, recall, F1, ROC
  - Imbalanced data and unequal misclassification costs
  - Probabilistic classification
  - Naïve Bayes classifier

# So far …

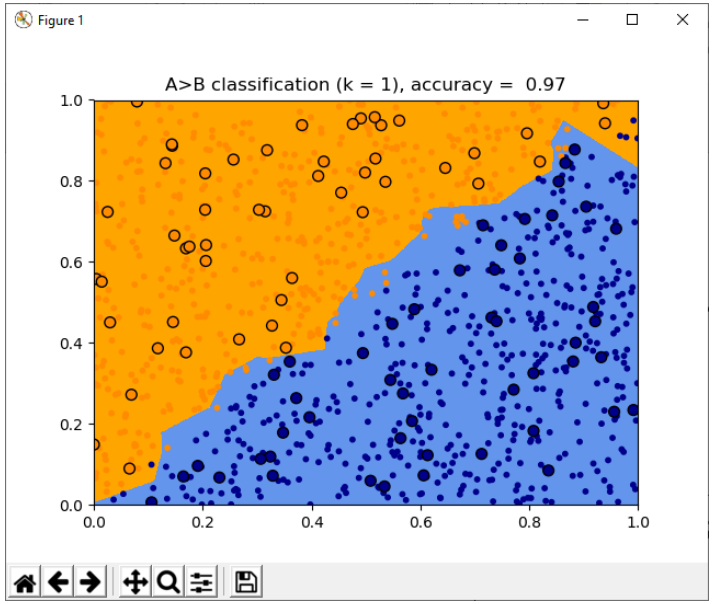
- Dec. 18 2019
  - Naive Bayes classifier
  - Laplace estimate
  - Regression (numeric prediction) and its evaluation

# From previous lessions

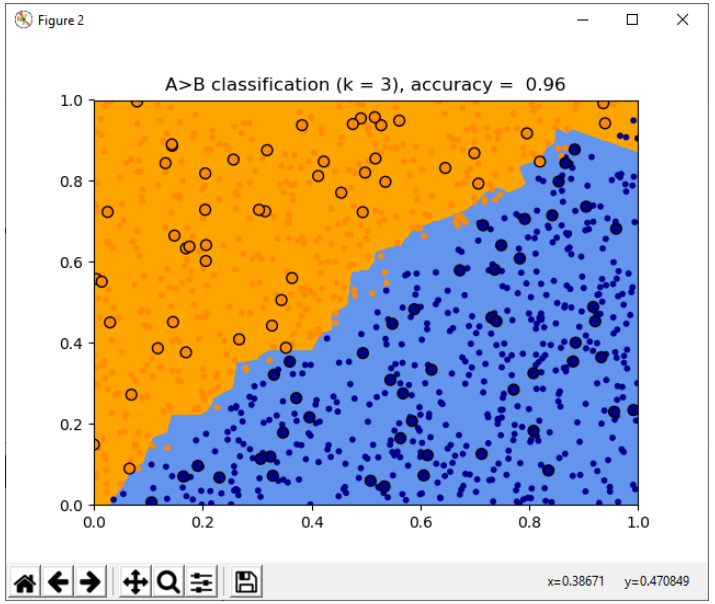
- Decision boundary for KNN

# What is a decision boundary like for KNN

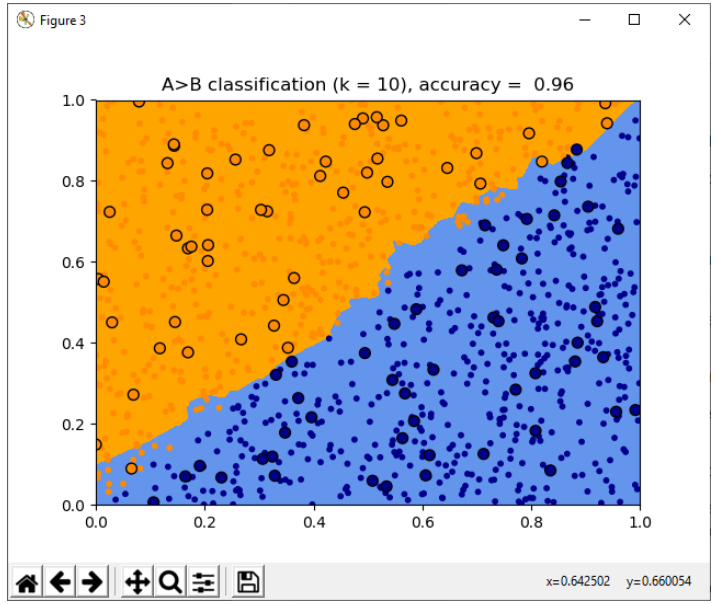
**K=1**




**K=3**



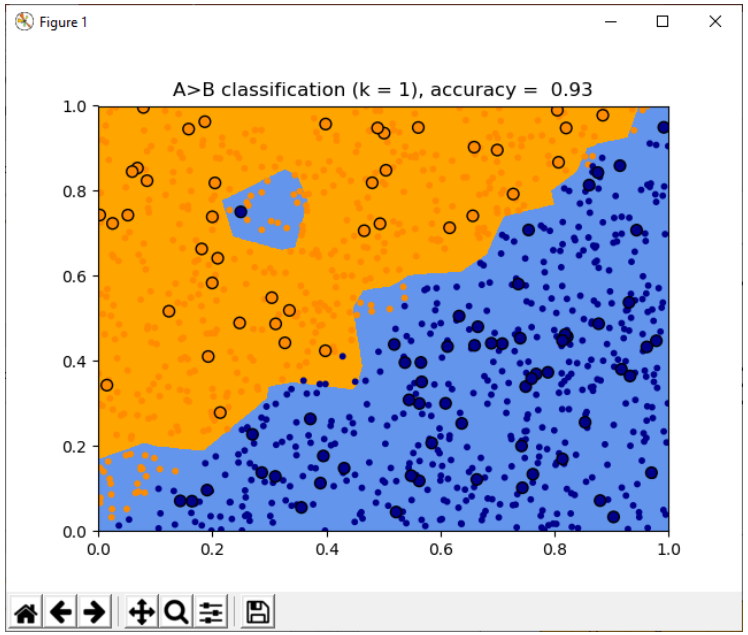

**K=10**



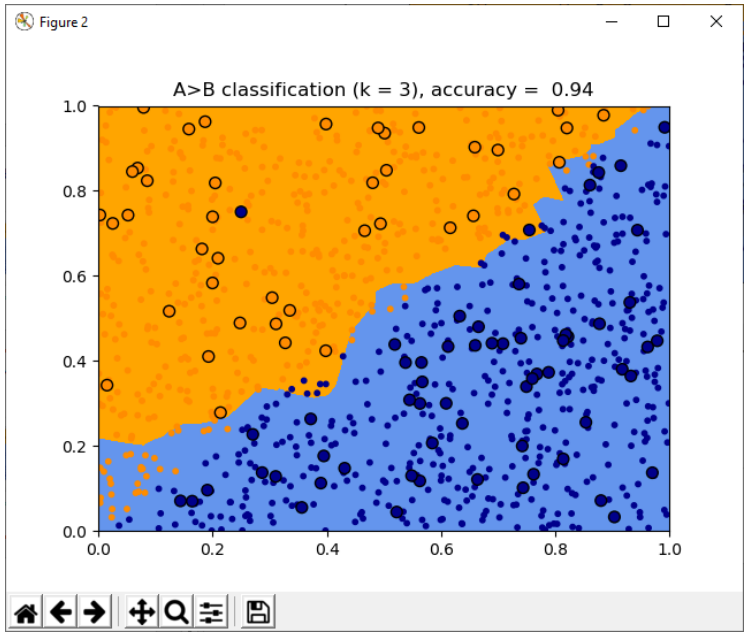

The large circles are the training set, the small ones are the test set – colored by the real labels. The background colors represent the decision boundary.
The source code for this is available at http://source.ijs.si/pkraljnovak/DM_course

# Decision boundary for KNN: one noisy example

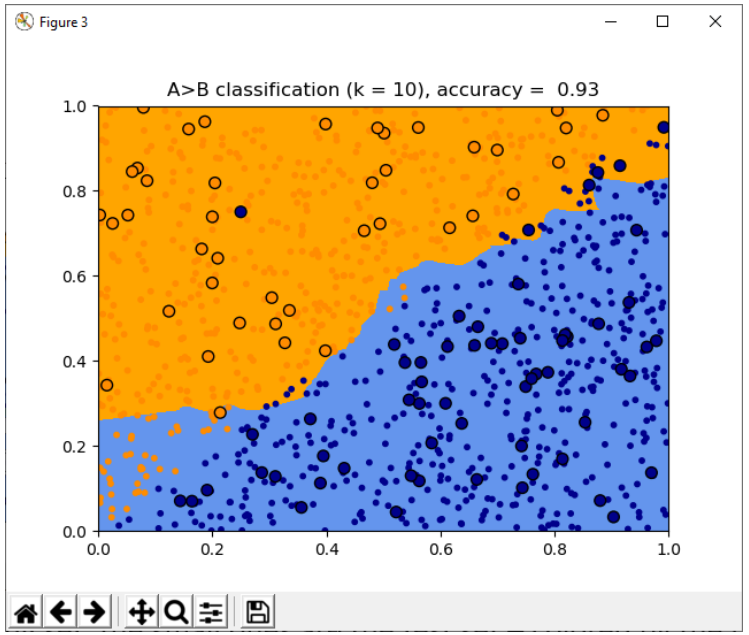
**K=1**



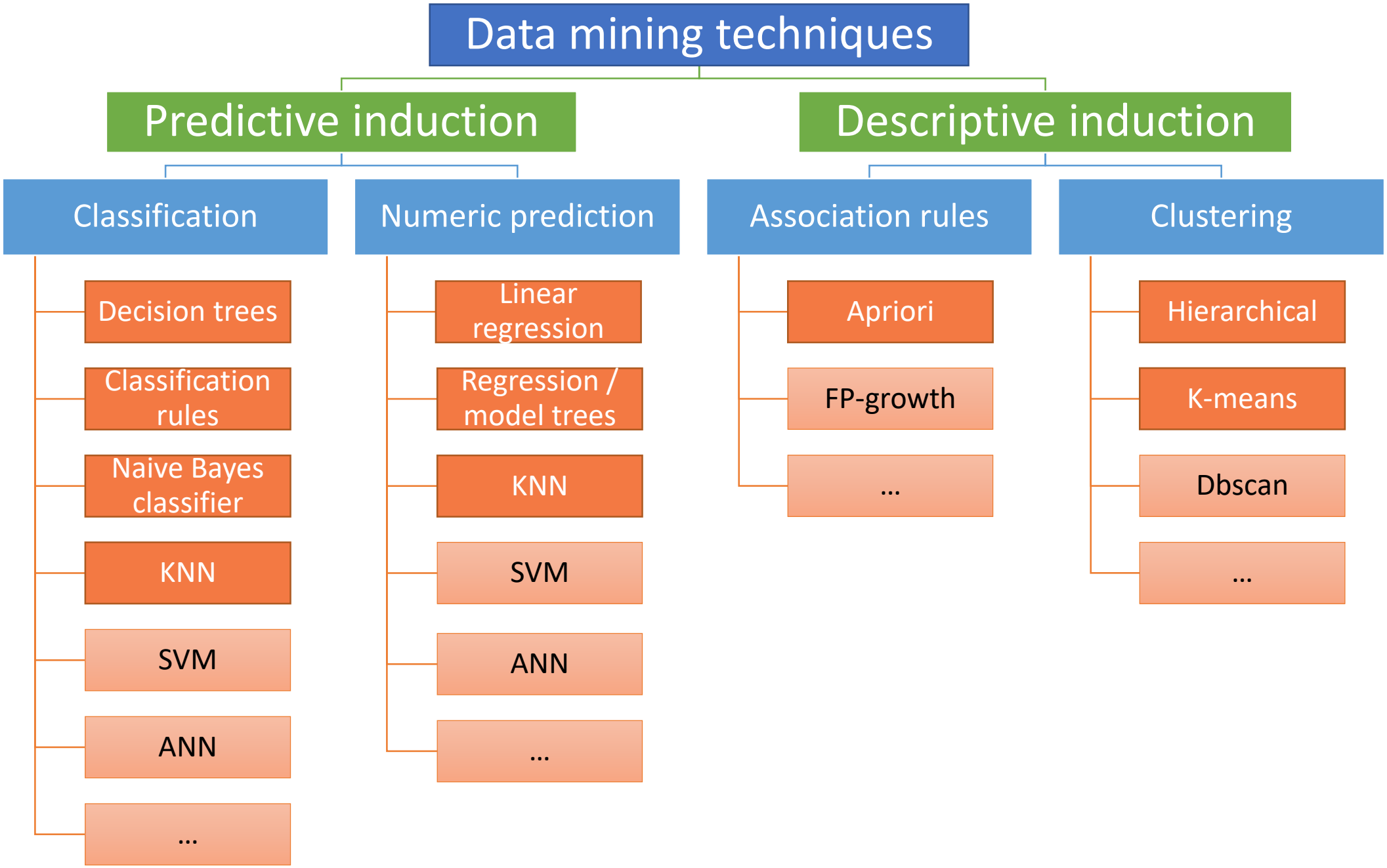

**K=3**




**K=10**




The large circles are the training set, the small ones are the test set – colored by the real labels. The background colors represent the decision boundary.
The source code for this is available at http://source.ijs.si/pkraljnovak/DM_course


Data mining techniques
Predictive induction
Descriptive induction
Classification
Numeric prediction
Association rules
Clustering
Decision trees
Classification rules
Naive Bayes classifier
KNN
SVM
ANN
…
Linear regression
Regression / model trees
KNN
SVM
ANN
…
Apriori
FP-growth
…
Hierarchical
K-means
Dbscan
…

## Or harried dads rewarding themselves with impulse buys

# Motivation


What people buy in a given shopping experience.

- 25 Osco Drug stores
- 1.2 million market baskets

(A market basket is the stuff you put in
the physical cart and check out at the register.)

- An unexpected pattern

Between 5p.m. and 7p.m. **diapers → beer**

http://www.dssresources.com/newsletters/66.php

# Association rules

# Association rules

- Determine associations between groups of items bought by customers.
- No predefined target variable(s).
- Find interesting, useful patterns and relationships.
- Data mining, business intelligence.

* Terminology from market basket analysis (transactions, items, itemsets, …)

# Confidence and support

- The dataset consists of **n** transactions
- We have an association rule A→ B

The **support** of an itemset A is defined as the fraction of the transactions in the database T = {T1 . . . Tn} that contain A as a subset.

$$supp(A) = \frac{|A|}{n}$$

$$supp(A \rightarrow B) = \frac{|A \wedge B|}{n}$$

The **confidence** of the rule A→ B is the conditional probability of A and B occurring in a transaction, given that the transaction contains A.

$$conf(A \rightarrow B) = \frac{|A \wedge B|}{|A|} = P(B|A)$$

# Example of a snapshot of a market basket data set

| tid | Set of items | Binary representation |
|---|---|---|
| 1 | $\{Bread, Butter, Milk\}$ | 110010 |
| 2 | $\{Eggs, Milk, Yogurt\}$ | 000111 |
| 3 | $\{Bread, Cheese, Eggs, Milk\}$ | 101110 |
| 4 | $\{Eggs, Milk, Yogurt\}$ | 000111 |
| 5 | $\{Cheese, Milk, Yogurt\}$ | 001011 |

- What do customers buy together?
- Which items imply the purchase of other items?

Charu C. Aggarwal : Data Mining: The Textbook (2015), pg. 95

# Exercise

| tid | Set of items | Binary representation |
|---|---|---|
| 1 | $\{Bread, Butter, Milk\}$ | 110010 |
| 2 | $\{Eggs, Milk, Yogurt\}$ | 000111 |
| 3 | $\{Bread, Cheese, Eggs, Milk\}$ | 101110 |
| 4 | $\{Eggs, Milk, Yogurt\}$ | 000111 |
| 5 | $\{Cheese, Milk, Yogurt\}$ | 001011 |

$$supp(A) = \frac{|A|}{n}$$

$$supp(A \rightarrow B) = \frac{|A \wedge B|}{n}$$

$$conf(A \rightarrow B) = \frac{|A \wedge B|}{|A|} = P(B|A)$$

Supp ({Bread}) =

Supp ({Milk, Yogurt}) =

Conf ({Milk} → {Yogurt}) =

Conf ({Yogurt} → {Milk}) =

Charu C. Aggarwal : Data Mining: The Textbook (2015, pg. 95)

# Association rules

- Rules **A → B**, where A and B are conjunctions of items
- Task: Find **all** association rules that satisfy the minimum support and minimum confidence constraints

- **Support**:

$$supp(A \rightarrow B) = \frac{|A \wedge B|}{n}$$

- **Confidence**:

$$conf(A \rightarrow B) = \frac{|A \wedge B|}{|A|} = P(B|A)$$

# Hard problem

- In practice
  - Millions of transactions
  - Many (thousands) of items
- Too many possible combinations
  - 1000 items for sale → $2^{1000}$ - 1 candidate market baskets
- Solution
  - *Apriori algorithm*

# Frequent itemsets: intuition

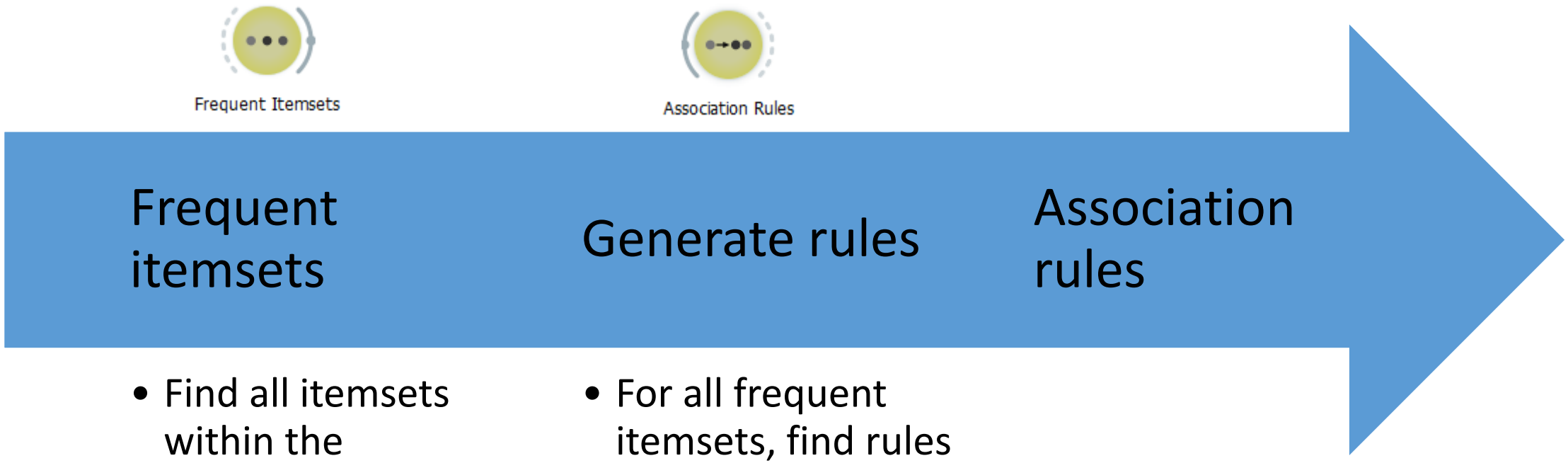
- We have ***n*** transactions containing (at least) {gloves, scarf and hat}
- What can we say about the number of transaction containing {gloves and scarf}?

*At least **n**.*

- The **anti-monotone property of support:** if we drop out an item from an itemset, support value of new itemset generated will either be the same or will increase.

$$\forall A, B: \ A \subseteq B \Rightarrow supp(A) \geq supp(B)$$

# Apriori




- Find all itemsets within the *minSupport* constraint
- For all frequent itemsets, find rules which satisfy the *minConfidence* constraint

*Frequent itemsets = large itemsets, sometimes also frequent patterns

# Apriori

Create $L_1$ = set of supported itemsets of cardinality one
Set $k$ to 2
while $(L_{k-1} \neq \emptyset)$ {
    Create $C_k$ from $L_{k-1}$
    Prune all the itemsets in $C_k$ that are not
        supported, to create $L_k$
    Increase $k$ by 1
}
The set of all supported itemsets is $L_1 \cup L_2 \cup \cdots \cup L_k$

(Generates $C_k$ from $L_{k-1}$)
Join Step
Compare each member of $L_{k-1}$, say $A$, with every other member, say $B$, in turn. If the first $k-2$ items in $A$ and $B$ (i.e. all but the rightmost elements of the two itemsets) are identical, place set $A \cup B$ into $C_k$.
Prune Step
For each member $c$ of $C_k$ in turn {
Examine all subsets of $c$ with $k-1$ elements
Delete $c$ from $C_k$ if any of the subsets is not a member of $L_{k-1}$
}

# Apriori: Frequent itemset mining

Create $L_1$ = set of supported itemsets of cardinality one
Set $k$ to 2
while $(L_{k-1} \neq \emptyset)$ {
    Create $C_k$ from $L_{k-1}$
    Prune all the itemsets in $C_k$ that are not
        supported, to create $L_k$
    Increase $k$ by 1
}
The set of all supported itemsets is $L_1 \cup L_2 \cup \cdots \cup L_k$

- The items in the sets should be ordered (alphabetically, …)

# Apriori: constructing the next level from the previous one

- Since items in the sets are ordered (alphabetically, …)
- Join Step:
  - Merge sets that have all the elements the same except for the rightmost one
- Prune Step:
  - Remove the set if any of its subsets are not on the previous level

(Generates $C_k$ from $L_{k-1}$)

Join Step

Compare each member of $L_{k-1}$, say $A$, with every other member, say $B$, in turn. If the first $k-2$ items in $A$ and $B$ (i.e. all but the rightmost elements of the two itemsets) are identical, place set $A \cup B$ into $C_k$.

Prune Step

For each member $c$ of $C_k$ in turn {
Examine all subsets of $c$ with $k-1$ elements
Delete $c$ from $C_k$ if any of the subsets is not a member of $L_{k-1}$
}

# Rules from frequent itesmets

- Generate rules with a certain confidence
- All the counts we need are in the lattice (no database scanning)
- Confidence of rules generated from the same itemset has an  anti-monotone property
- No need to check all the rules, since

Conf ( {A,B} → {C} ) >= Conf ( {A} → {B,C} )
Conf( {A,B,C} → {D} ) ≥ Conf( {A,B} → {C,D} ) ≥ Conf( {A} →  {B,C,D} )

*In general, confidence does not have an anti-monotone property: Conf(ABC → D) can be larger or smaller than Conf(AB → D)

Charu C. Aggarwal : Data Mining: The Textbook (2015), pg. 98

# Exercise: Association rules

Generate frequent itemsets with support at least 2/6 and confidence at least 75%.

---

Items: **A**=apple, **B**=banana, **C**=coca-cola, **D**=doughnut

- Client 1 bought: A, B, C, D
- Client 2 bought: B, C
- Client 3 bought: B, D
- Client 4 bought: A, C
- Client 5 bought: A, B, D
- Client 6 bought: A, B, C

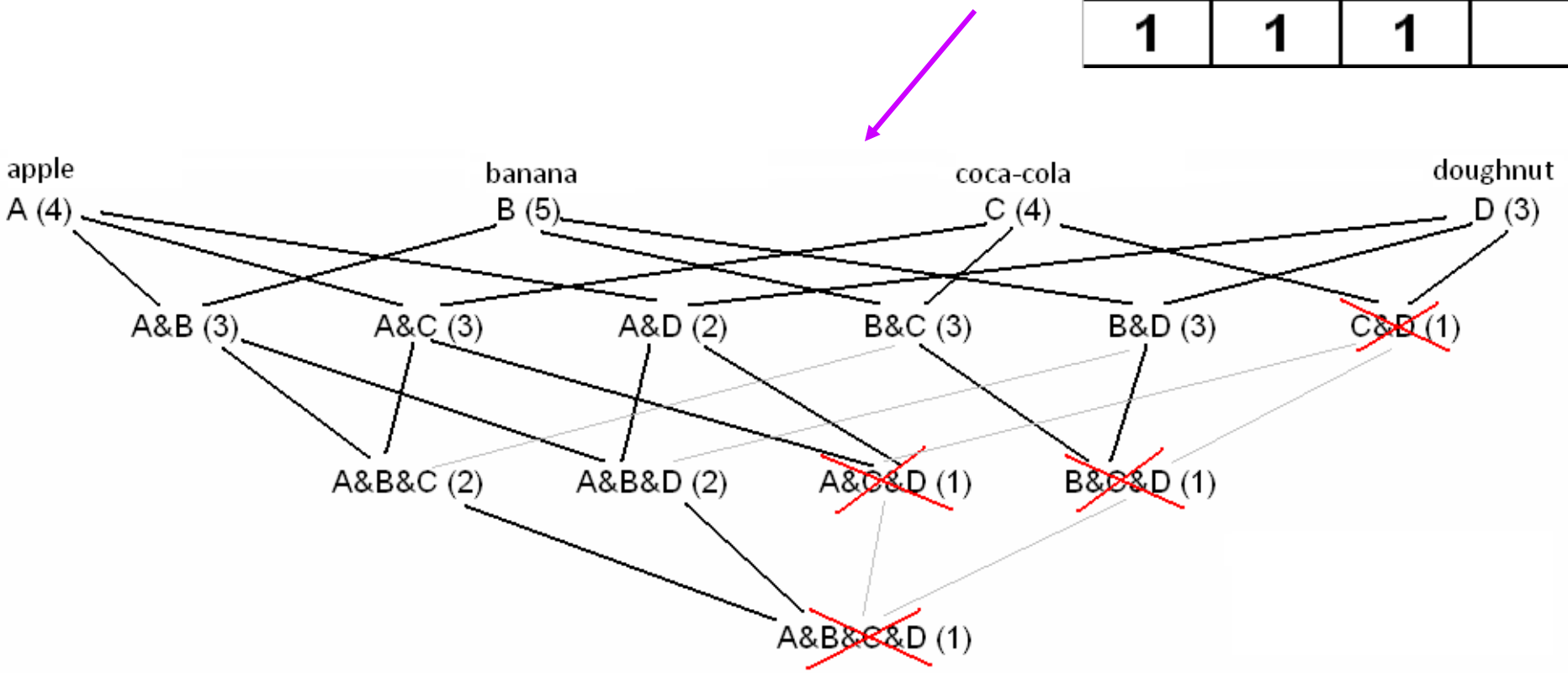
# Exercise: Frequent itemsets

To ease the counting, we transcribe into a binary representation.

| A | B | C | D |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
|  | 1 | 1 |  |
|  | 1 |  | 1 |
| 1 |  | 1 |  |
| 1 | 1 |  | 1 |
| 1 | 1 | 1 |  |

# Pogoste množice

| A | B | C | D |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| | 1 | 1 | |
| | 1 | | 1 |
| 1 | | 1 | |
| 1 | 1 | | 1 |
| 1 | 1 | 1 | |



# Association rules …

| | |
|---|---|
| **A&B** | **Support = 3/6** |
| A → B | Confidence = 2/4 = 50% |
| B → A | Conficence = 2/5 = 40% |
| **A&C** | **Support = 3/6** |
| A → C | Conficence = 3/4 = 75% |
| C → A | Conficence = 3/4 = 75% |
| **A&D** | **Support = 2/6** |
| A → D | Conficence = 2/4 = 50% |
| D → A | Conficence = 2/3 = 67% |
| **B&C** | **Support = 3/6** |
| B → C | Conficence = 3/5 = 60% |
| C → B | Conficence = 3/4 = 75% |
| **B&D** | **Support = 3/6** |
| B → D | Conficence = 3/5 = 60% |
| D → B | Conficence = 3/3 = 100% |

# … association rules

No need to generate rules {A}→{B,C}, {B}→{C,A} , {C}→{A,B} because the rules with two items on the left side from the same itemset do not satisfy the minimum confidence constraint.

Similar for the itemset {A, B, D}

| **A&B&C** | **Support = 2/6** |
|---|---|
| A&B → C | Confidence = 2/3 = 67% |
| A&C → B | Confidence = 2/3 = 67% |
| B&C → A | Confidence = 2/3 = 67% |
| **A&B&D** | **Support = 2/6** |
| A&B → D | Confidence = 2/3 = 67% |
| A&D → B | Confidence = 2/2 = 100% |
| B&D → A | Confidence = 2/3 = 67% |
| B → A&D | Confidence = 2/5 = 40% |

# Lift

- The lift of rule **L → R** measures how many more times the items in L and R occur together in transactions than would be expected if the itemsets L and R were statistically independent.

$$\text{lift}(L \to R) = \frac{\text{support}(L \cup R)}{\text{support}(L) \times \text{support}(R)}$$

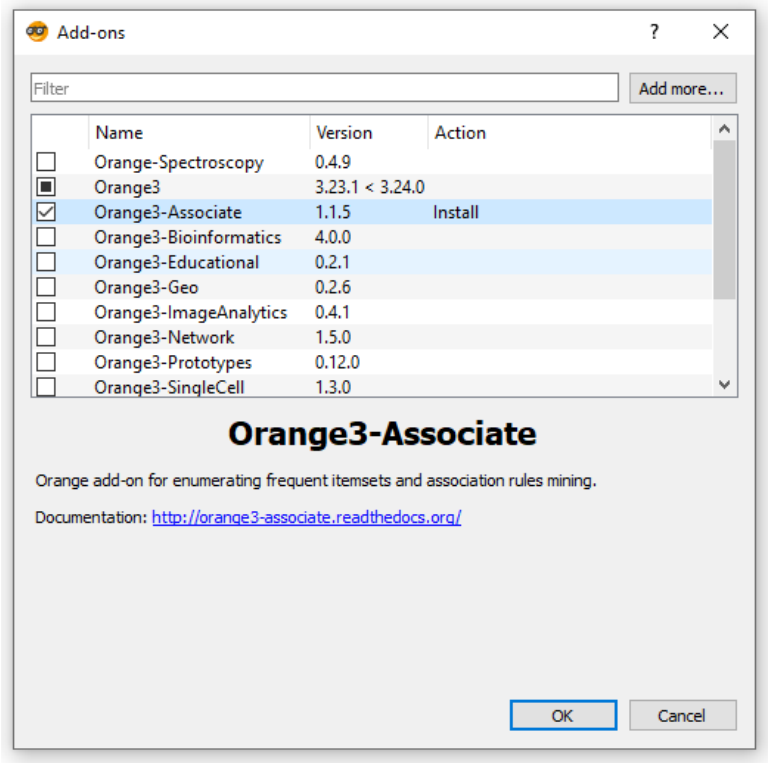
$$\text{lift}(L \to R) = \text{lift}(R \to L)$$

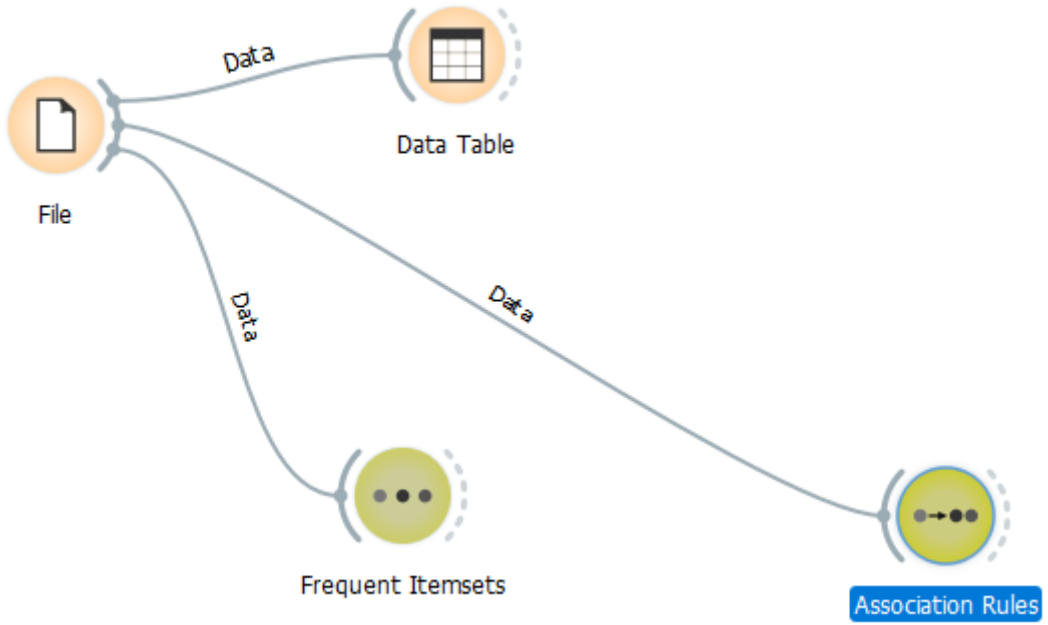
# Leverage

- The leverage of rule **L → R** is the difference between the support for L ∪ R (i.e. the items in L and R occurring together in the database) and the support that would be expected if L and R were independent.

$$\text{leverage}(L \rightarrow R) = \text{support}(L \cup R) - \text{support}(L) \times \text{support}(R)$$

# Association rules: Orange workflow




* Start with a small minSupport and we increase it gradually (to avoid running out of memory)

# Association rules quality measures in Orange

| Supp | Conf | Covr | Strg | Lift | Levr | Antecedent | | Consequent |
|---|---|---|---|---|---|---|---|---|
| 0.050 | 0.178 | 0.283 | 0.618 | 1.017 | 0.001 | Fresh Vegetables | → | Fresh Fruit |
| 0.050 | 0.287 | 0.175 | 1.619 | 1.017 | 0.001 | Fresh Fruit | → | Fresh Vegetables |

- **support, confidence, lift**, **leverage**
- **coverage**: how often antecedent items are found in the data set (support of antecedent/data)
- **strength**:  (support of consequent/support of antecedent)

# Lab exercise

## Datasets

- https://biolab.si/core/foodmart.basket
- https://github.com/digizeph/data_mining/blob/master/data/FoodMart.csv
- http://file.biolab.si/datasets/voting.tab

1. Compare the two datasets (files)
2. Generate frequent itemsets and association rules for both datasets. What is the difference?
3. Frequent itemsets and association rules for „Voting.tab“ $\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$
4. Compute „conviction“ for a few rules

Conviction: the ratio of the expected frequency that X occurs without Y if X and Y were independent divided by the observed frequency of incorrect predictions.

# Beyond the basics

- Summarization:
  - finding maximal itemsets
  - closed itemsets
  - nonredundant rules
- *Constraint incorporation:*
  - application-specific constraints into the itemset generation process → much lower support-levels
- Multilevel association rules: considering item classes
  - (e.g., chips, peanuts, bretzels, etc., belonging to class “snacks”)

# (Non) redundant patterns

**Maximal Frequent Itemset**

- *A frequent itemset is maximal at a given minimum support level minsup if it is frequent and no superset of it is frequent.*

**Closed Itemsets**

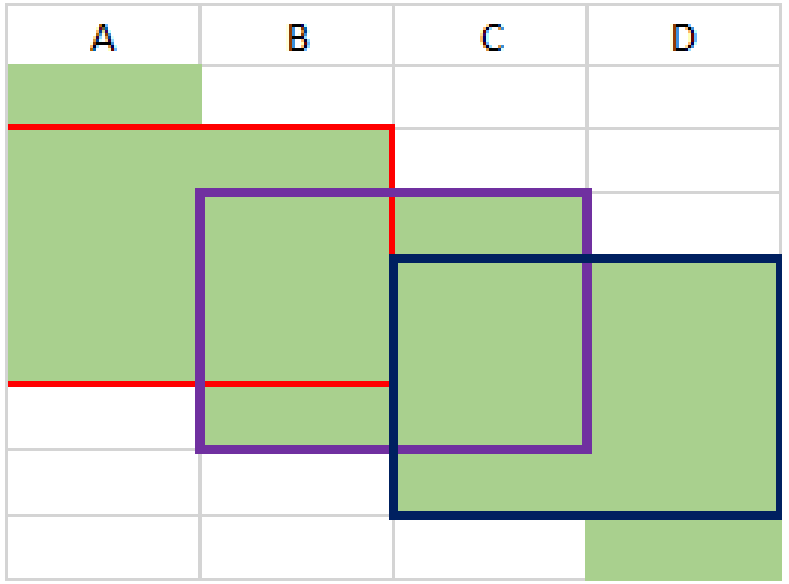
- *An itemset X is closed, if none of its supersets have exactly the same support count as X.*

**Closed Frequent Itemsets**

- *An itemset X is a closed frequent itemset at minimum support minsup, if it is both closed and frequent.*

# Example

| A | B | C | D |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |




Find closed frequent itemsets with the minSupport level 5/8 and 4/8.

# Constraint-based patter mining

- Constraint-based pattern mining systems are systems that with minimal effort can be programmed to find different types of patterns satisfying constraints.
- They achieve this genericity by providing
  - (1) high-level languages for specifying constraints;
  - (2) generic search algorithms that find patterns for any task

# Applications

- Market Basket Analysis
- Classification: rule based classifiers
- Clustering: CLIQUE algorithm for finding dense regions of data
- Outlier detection: the outliers are defined as transactions that are not “covered” by most of the association patterns in the data
- Recommendations and Collaborative Filtering: In collaborative filtering, the idea is to make recommendations to users on the basis of the buying behavior of other similar users. The problem of localized pattern mining is much more challenging because of the need to simultaneously determine the clustered segments and the association rules.

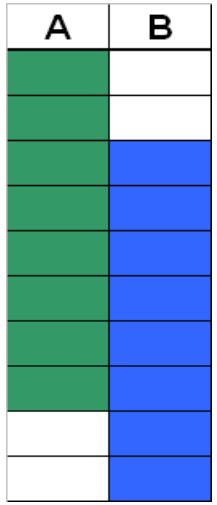
# Literature

- Max Bramer: Principles of data mining (2007)
  - 13. Association Rule Mining II
- Charu C. Aggarwal : Data Mining: The Textbook (2015)
  - 4. Association pattern mining
  - 5. Association pattern mining: advanced concepts

- What is the "true story" about using data mining to identify a relation between sales of beer and diapers? http://www.dssresources.com/newsletters/66.php

# Homework

1. Transformation of an attribute-value dataset to a transaction dataset.
2. What would be the association rules for a dataset with two items A and B, each of them with support 80% and appearing in the same transactions as rarely as possible?
   a. minSupport = 50%, min conf = 70%
   b. minSupport = 20%, min conf = 70%
3. What if we had 4 items: A, ¬A, B, ¬ B




4. Compare decision trees and association rules regarding handling an attribute like “PersonID”. What about attributes that have many values (eg. Month of year)