Universidade do Porto

Faculdade de Engenharia

FEUP

Data Mining

a gentle introduction

Data Mining is ...

Sources of data are increasing everyday:

- Sensors
- Internet, e.g. social networks
- Data warehousing systems

Data Mining

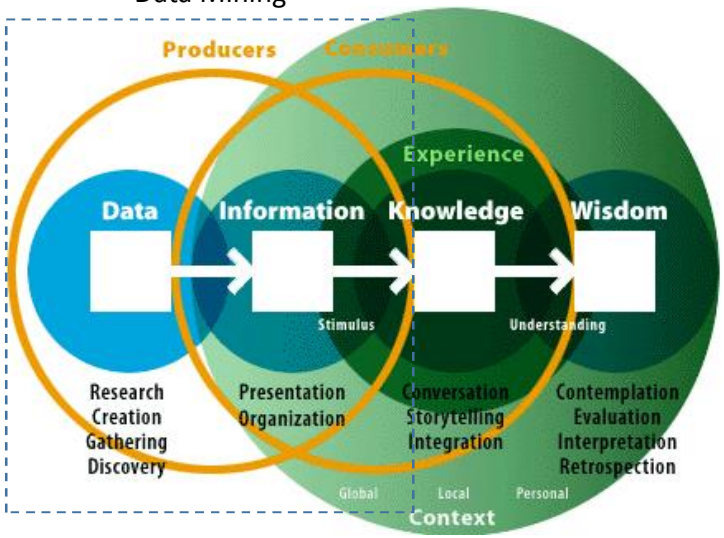


Image obtained in 31-5-2014 from
http://www.nathan.com/thoughts/course.html

Extracting information from data:

- Without doing assumptions about data distribution
- Discovering unknown information
- Using computational resources

Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

Classification: predicting a categorical variable; **Regression**: predicting a quantitative variable

Example: a regression problem

Age	Height (m)	Sex	Weight (kg)
25	1.90	M	70
37	1.75	F	60
75	1.68	M	83
55	1.55	F	65
49	1.62	M	80
42	1.76	M	93
35	1.69	F	62
66	1.48	F	57
71	1.73	F	???

What we know

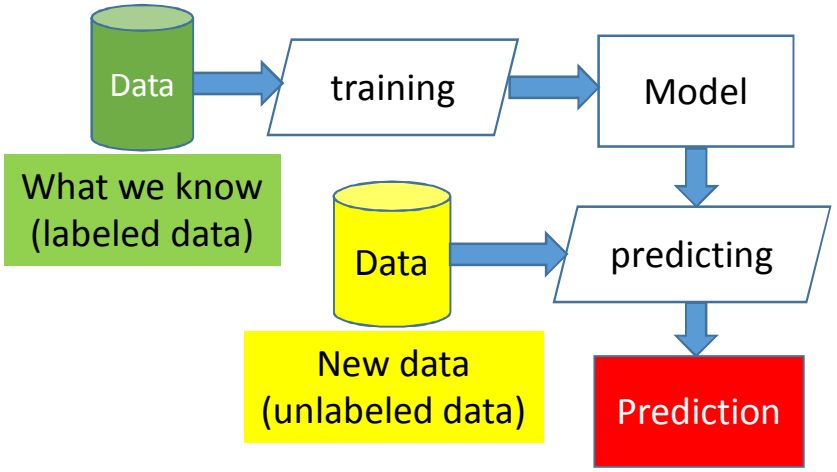
New data

What we want to know

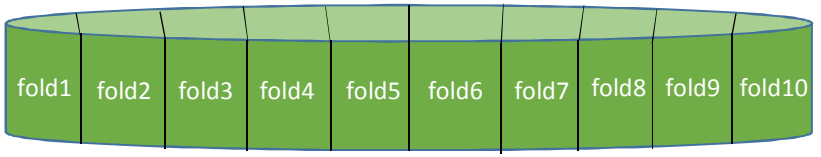
Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

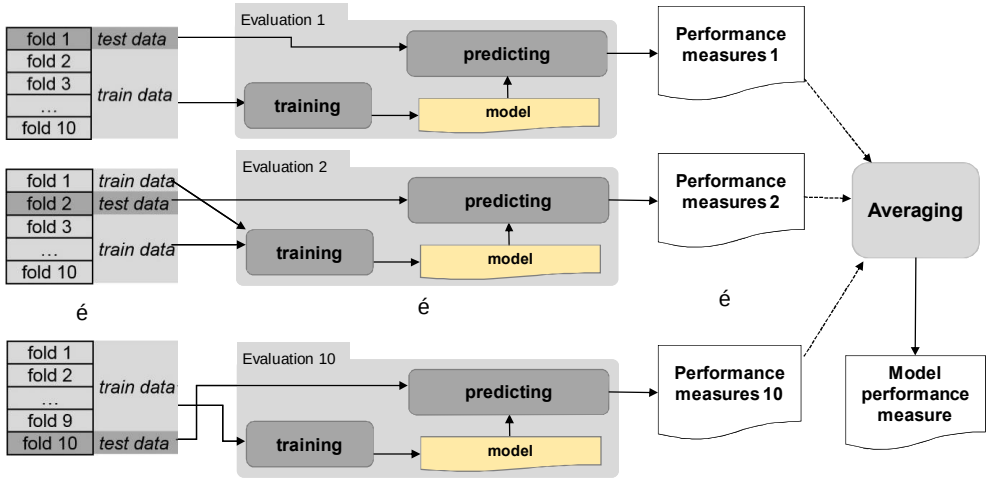
How to solve it



How to evaluate different models



Split data into k-folds (k=10 in the example) of identical size



Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

Algorithms (for training)

- Naïve Bayes (Classification)
- Logistic Regression (classification)
- Local Regression (regression)
- MARS (regression)
- Decision Trees
- Bagging
- AdaBoost (Classification)
- Random Forest
- Support Vector Machines
- Artificial Neural Networks
- …

Performance measures

For regression

- Root mean squared error = $\sqrt{\sum \frac{(\hat{y}_i - y_i)^2}{n}}$
- Relative squared error = $\frac{\sum (\hat{y}_i - y_i)^2}{\sum (\bar{y} - y_i)^2}$

$\hat{y}$: predicted value; y: true value; $\bar{y}$: *represents the average of the true unknown model;* n: number of instances

For classification

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp - fn}$$

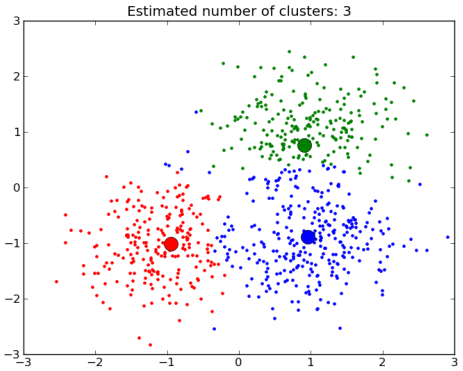
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

tp: true positive; tn: true negative; fp: false positive; fn: false negative

Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …



Organization of data into groups so that there is:

- High similarity between the objects belonging to the same group
- Low similarity between objects of different groups

What it's like to be similar?

There are similarity measures for different areas of applications:

- Text
- Images
- Sound
- Time series
- …

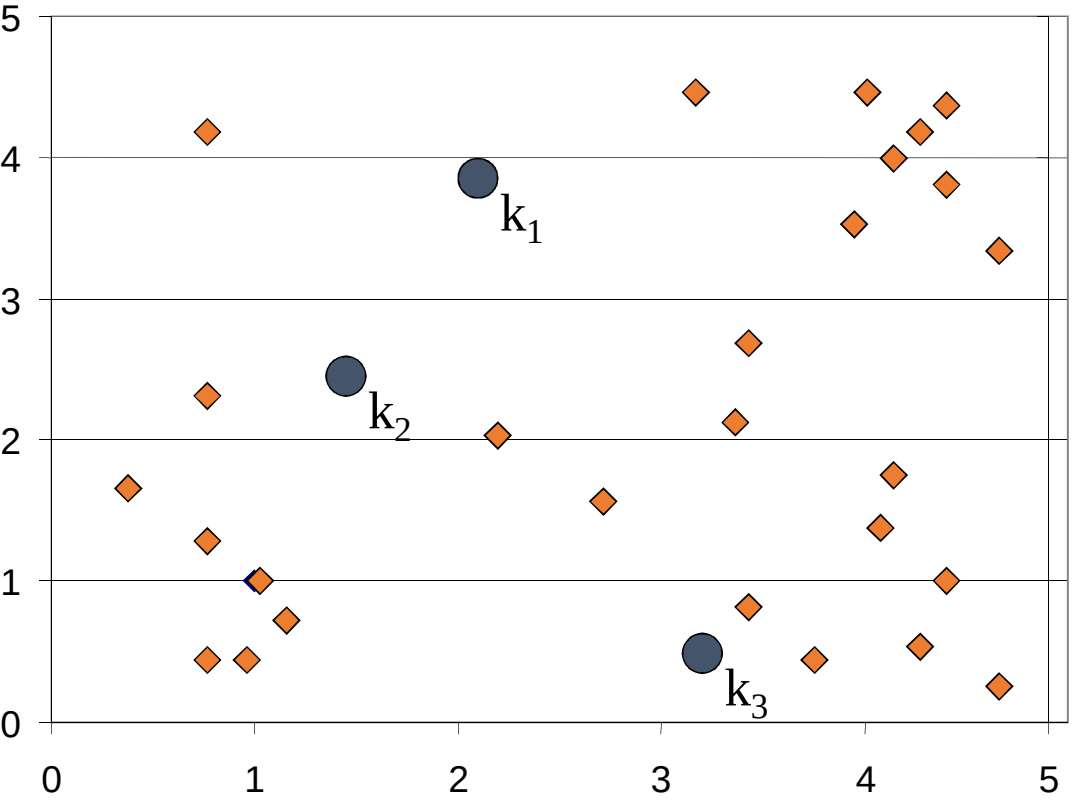
The most popular clustering algorithm is the **k-means** (it uses the Euclidean distance as similarity measure):

1. Decide a value for k (the number of groups).
2. Initializes the centers of the k clusters (randomly, if needed).
3. Decide the cluster to which each of the N objects belong by assigning them the cluster with the nearest center.
4. Re-estimate the k cluster centers, assuming the assignment made in 3 is correct.
5. If none of the N objects change cluster in the last iteration, leave. Otherwise go to 3.

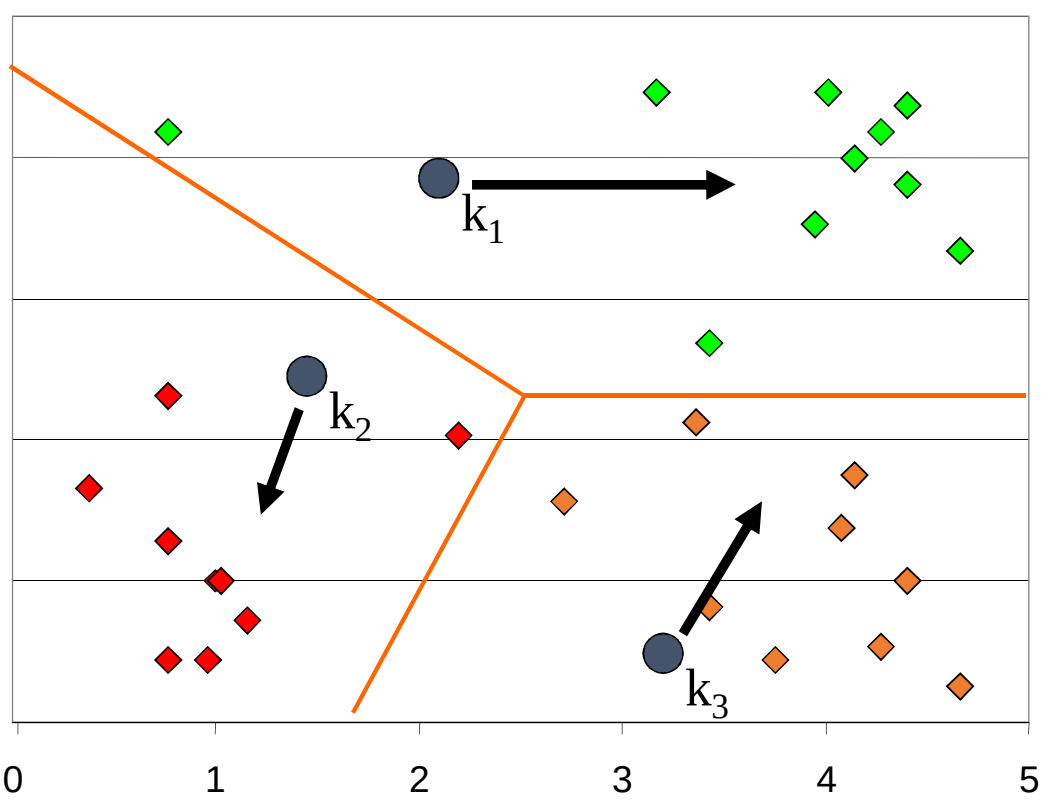
Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

K-means: step 1



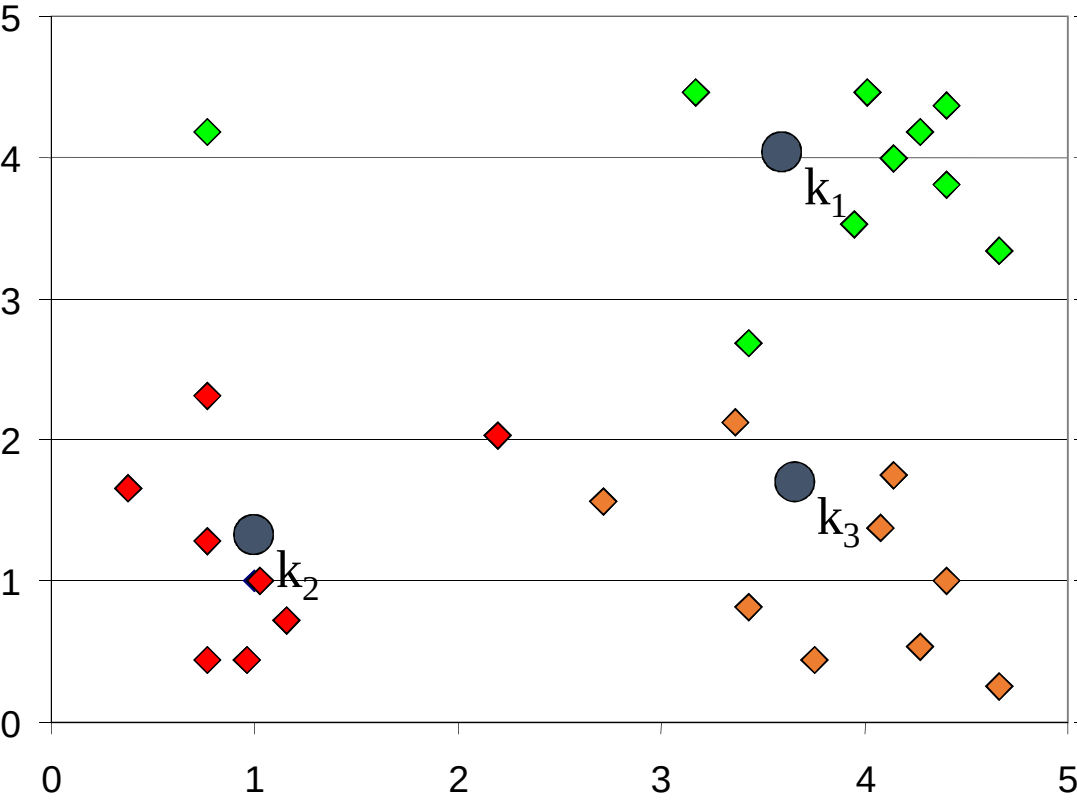
K-means: step 2



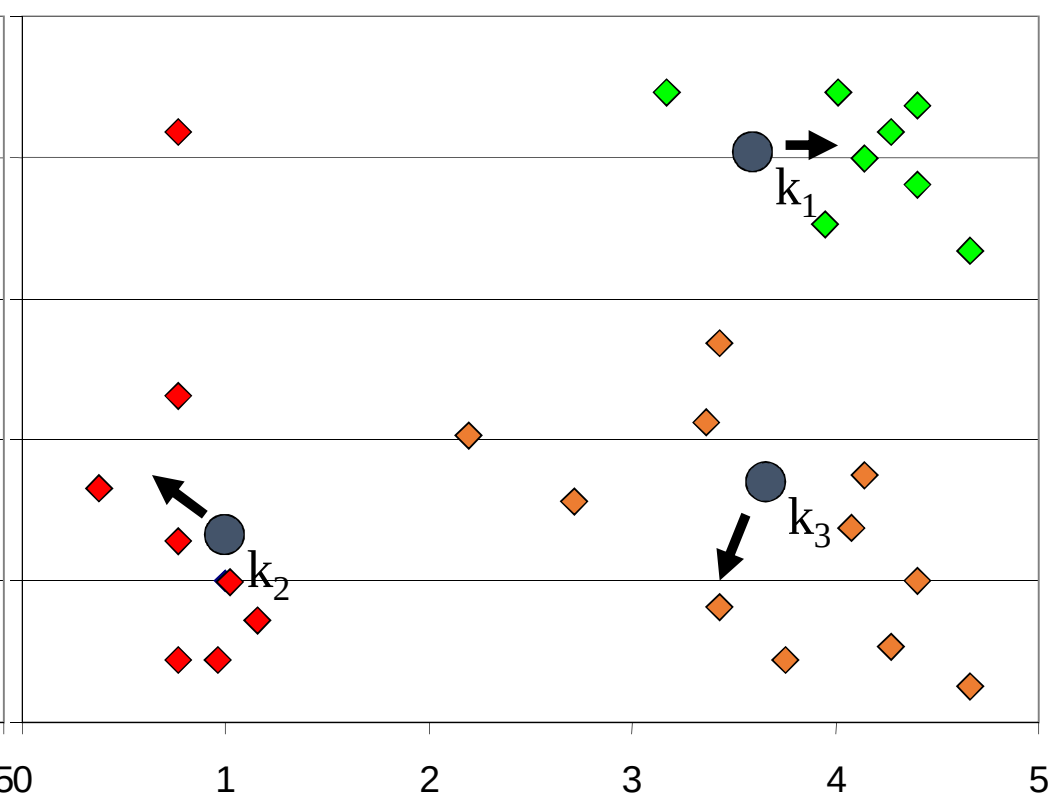
Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

K-means: step 3



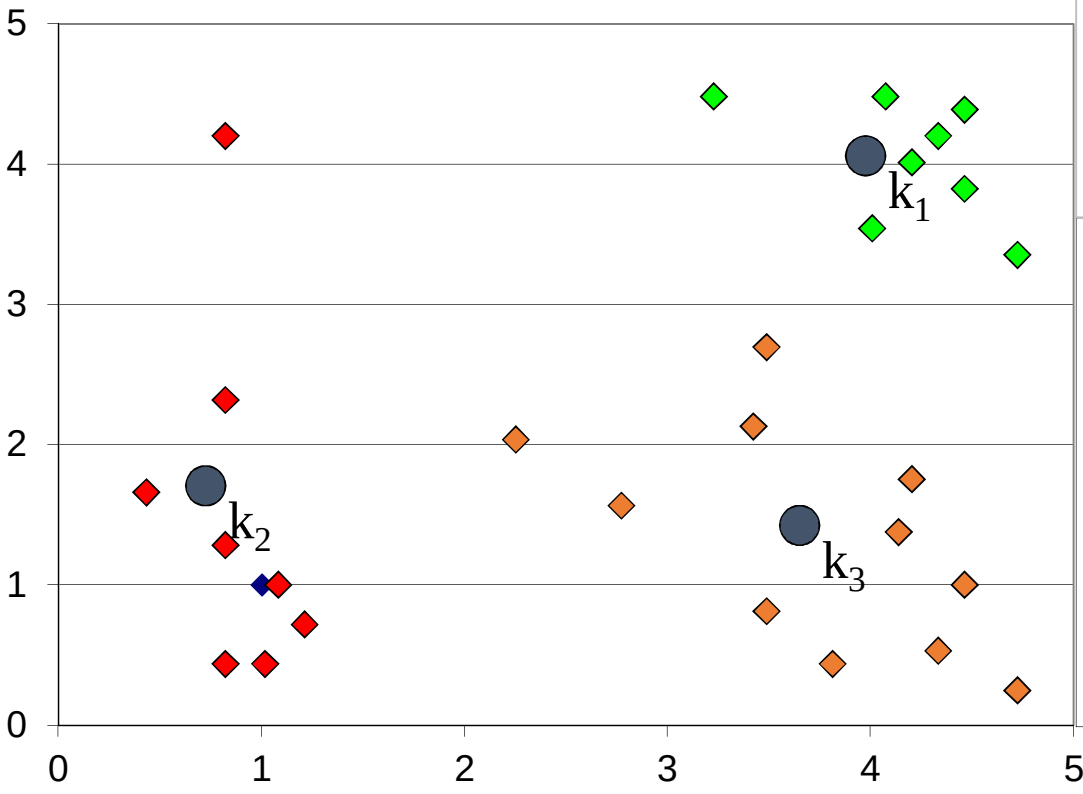
K-means: step 4



Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, ...

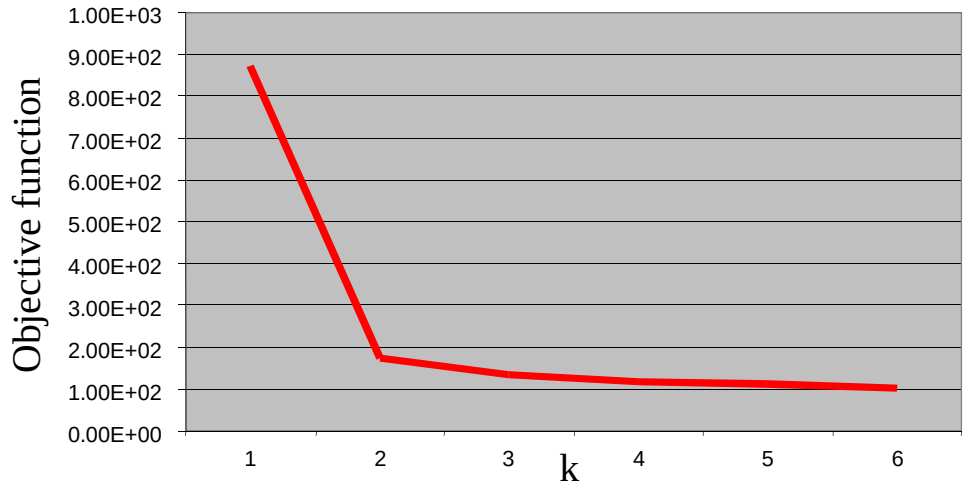
K-means: step 5



Evaluation: which is the right value for k?

We can make a chart with the values of the objective function for k from 1 to 6.

The sudden change when k = 2, is a strong indicator of the existence of 2 clusters in the data. This technique to determine the number of clusters is known as "**knee finding**ñ or "**elbow finding**".



Objective function: $\sum_{\square}^{\square}\left(\sum_{\square}^{\#\square}(\square_{\square}-\square_{\square})^{\square}\right)$, □ □□□□ #□□□□

□□□ □□□□ □□ □□□□□□□ □ □□□ □□ □□□ □□□□□□ □□ □□□□□□□ □

Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

Example: market basket analysis

TID	Production
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CORNFLAKES
4	MILK, BREAD, SUGAR
5	MILK, CORNFLAKES
6	BREAD, CORNFLAKES
7	MILK, CORNFLAKES
8	MILK, BREAD, CORNFLAKES, EGGS
9	MILK, BREAD, CORNFLAKES

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

ITEMS:

A = MILK
B= BREAD
C= CORNFLAKES
D= SUGAR
E= EGGS

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

The attributes are converted in binary attributes

The goal: to identify rules in the data, such as,
"*who buys products A and B also buys product C*"

Association rule example:
{ A1, A2 } => { A3, A4 }

Properties:
- support
  - Perportion of cases with the co-occurrence
  - Estimation of Prob(A1 & A2 & A3 & A4)
- confidence
  - Perportion of cases with {A3, A4} when {A1,A2} occurs
  - Estimation of Prob(A3 & A4 | A1 & A2)

Data Mining tasks

Prediction, Clustering, Association rules, Recommender systems, Ranking, Outlier/novelty detection, Social network analysis, …

The APRIORI algorithm (by Agrawal & Srikant)

Start by searching itemsets (sets of items) of size 1 (easy) by counting each item frequency

Idea: to use itemsets of size one to generate itemsets of size two, to use itemsets of size two to generate itemsets of size three, …

If (A B) is a frequent itemset, then both (A) and (B) are also frequent itemsets. For sure!

In general: if X is a frequent itemset with k-items, then all subsets of X with (k-1)-items are also frequent.

⇒Obtaining k-itemsets adding an item to (k-1)-itemsets.

For each frequent itemset I

For each subset J from I

To obtain all association rules of the form: I-J => J

?: Given the following item list, which association rules have min support = 40% and min confidence= 60% ?

TID	Item list
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Items	Sup
A	6/9
B	7/9
C	6/9
D	2/9
E	2/9
AB	4/9
AC	4/9
BC	4/9
ABC	2/9

AR	Conf
=>A	6/9
=>B	7/9
=>C	6/9
=>AB	4/9
=>AC	4/9
=>BC	4/9
A=>B	4/6
B=>A	4/7
A=>C	4/6
C=>A	4/6
B=>C	4/7
C=>B	4/6

Data Mining tasks

New trends

90% of world's data generated over last two years

http://www.sciencedaily.com/releases/2013/05/130522085217.htm

Big data

- **Volume: e.g., hadoop**
- **Velocity: knowledge discovery from data streams**
- **Variety: Information fusion, pre-processing for data streams**