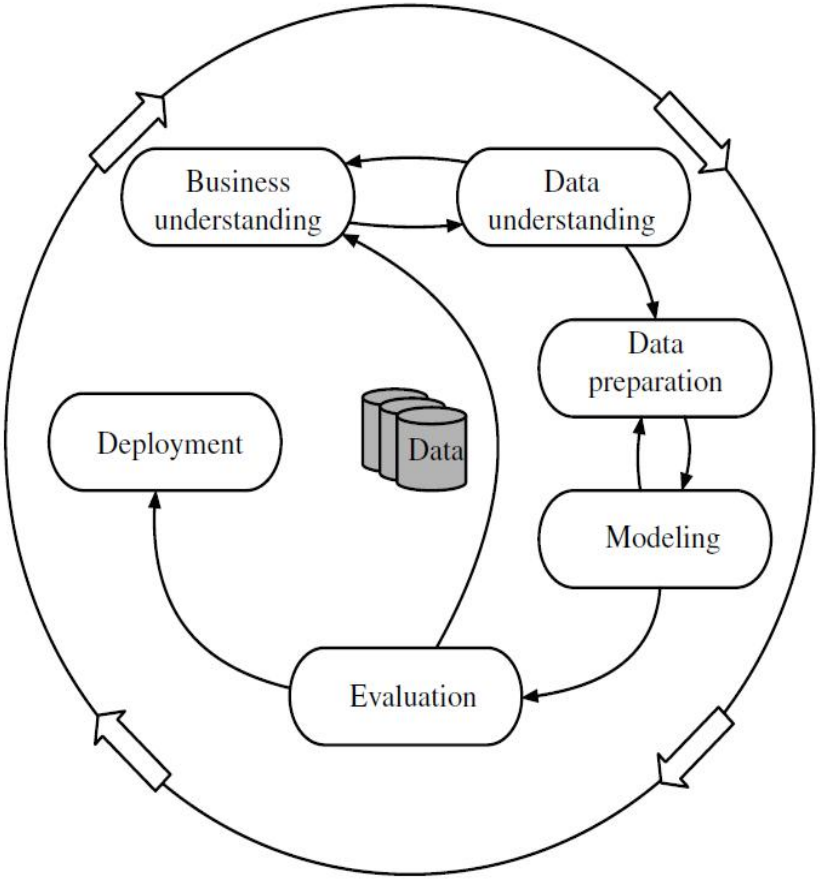
Introdução a Data mining (Practical machine learning)

Life cycle of a data mining (practical machine learning) project.



Fonte: Witten, I., Frank, E. **Data Mining: Practical Machine learning Tools and Techniques.** 4a.ed. Elsevier. 2014

Input: Problem, instances and Attributes?

- What is the Learning Problem ?
  - Classification,
  - Regression (numeric prediction)
  - Association,
  - Clustering,
- What's in an instance?
  - Relations
- What's in an variable ?
  - categorical, numeric
- Preparing the input
  - sparse data, tables, attributes, missing and inaccurate values, unbalanced data, getting to know your data

Components of the input

- Concepts: kinds of things that can be learned
  - Aim: intelligible and operational concept description
- Instances: the individual, independent examples of a concept to be learned
  - More complicated forms of input with dependencies between examples are possible
- Variables: measuring aspects of an instance
  - We will focus on categorical and numeric ones

Classification learning

- Example problems: weather data, contact lenses, irises, labor negotiations
- Classification learning is *supervised*
  - Scheme is provided with actual outcome
- Outcome is called the *class* of the example
- Measure success on fresh data for which class labels are known (*test data*)
- In practice success is often measured subjectively

Association learning

- Can be applied if no class is specified and any kind of structure is considered "interesting"
- Difference to classification learning:
  - Can predict any attribute's value, not just the class, and more than one attribute's value at a time
  - Hence: far more association rules than classification rules
  - Thus: constraints are necessary, such as minimum coverage and minimum accuracy

Not-supervised Learning(Clustering)

- Finding groups of items that are similar
- Clustering is *unsupervised*
  - The class of an example is not known
- Success often measured subjectively

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

Numeric prediction

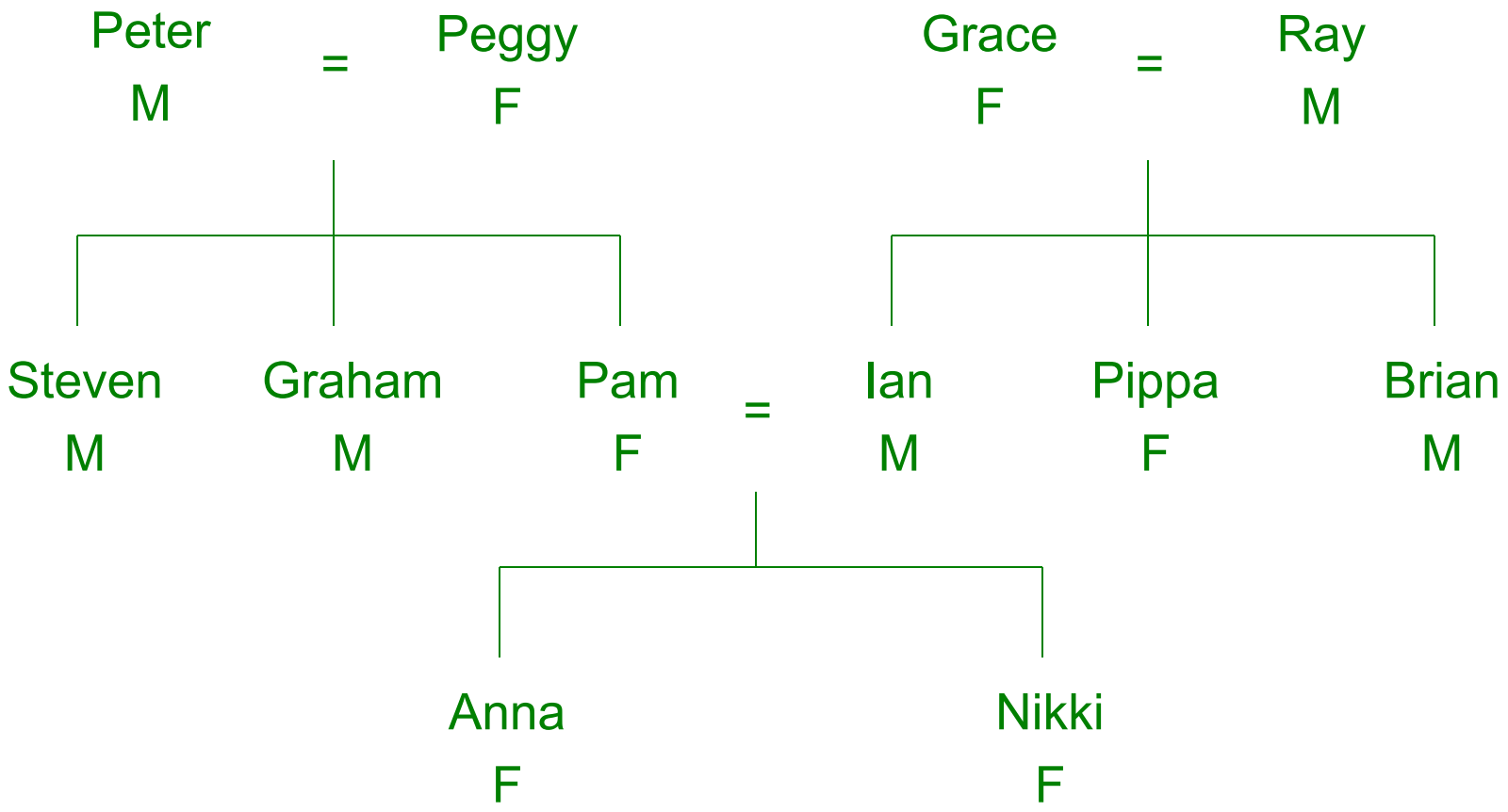
- Variant of classification learning where "class" is numeric (also called "regression")
- Learning is supervised
  - Scheme is being provided with target value
- Measure success on test data

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
...	...	...	...	...

What's in an example/instance?

- Instance: specific type of example
  - Thing to be classified, associated, or clustered
  - Individual, independent example of target concept
  - Characterized by a predetermined set of attributes
- Input to learning model: set of instances(or data points)/dataset
  - Uually represented as a single table
- Rather restricted form of input
  - No relationships between objects
- Most common form in practical data mining (and Machine learning)

Example: A family tree



How to represent it in a table ?

Family tree represented as a table

Name	Gender	Parent1	parent2
Peter	Male	?	?
Peggy	Female	?	?
Steven	Male	Peter	Peggy
Graham	Male	Peter	Peggy
Pam	Female	Peter	Peggy
Ian	Male	Grace	Ray
Pippa	Female	Grace	Ray
Brian	Male	Grace	Ray
Anna	Female	Pam	Ian
Nikki	Female	Pam	Ian

The "sister-of" relation

First person	Second person	Sister of?
Peter	Peggy	No
Peter	Steven	No
…	…	…
Steven	Peter	No
Steven	Graham	No
Steven	Pam	Yes
…	…	…
Ian	Pippa	Yes
…	…	…
Anna	Nikki	Yes
…	…	…
Nikki	Anna	yes

First person	Second person	Sister of?
Steven	Pam	Yes
Graham	Pam	Yes
Ian	Pippa	Yes
Brian	Pippa	Yes
Anna	Nikki	Yes
Nikki	Anna	Yes
All the rest		No

Closed-world assumption

A full representation in one table

First person				Second person				Sister of?
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Steven	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Graham	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Ian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Brian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Anna	Female	Pam	Ian	Nikki	Female	Pam	Ian	Yes
Nikki	Female	Pam	Ian	Anna	Female	Pam	Ian	Yes
All the rest								No

```
If second person's gender = female
   and first person's parent = second person's parent
   then sister-of = yes
```

What's in an variable/attribute?

- Each instance is described by a fixed predefined set of features, its variables or attributes
- But: number of variables may vary in practice
  - Possible solution: "irrelevant value" flag
- Related problem: existence of a variable may depend of value of another one
- Main Possible attribute types
  - *categorical, numerical*

Categorical levels of measurement

- Values are distinct symbols
  - Values themselves serve only as labels or names
  - Also called nominal or discrete
  - *Nominal* comes from the Latin word for name
- Example: attribute "outlook" from weather data
  - Values: "sunny","overcast", and "rainy"
- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

Variable types used in practice

- Many data mining models accommodate just two levels of measurement: nominal and ordinal
- Others deal exclusively with ratio quantities

- Nominal attributes are also called "**categorical**", "enumerated", or "**discrete**"

- Special case: dichotomy ("boolean" attribute)

- **numeric** or "**continuous**" variables deal with number (reals or integers)

Metadata

- Information about the data that encodes background knowledge
- This information can be used to restrict the search space of the learning algorithm
- Examples:
  - Dimensional considerations
    (i.e., expressions must be dimensionally correct)
  - Circular orderings
    (e.g., degrees in compass)
  - Partial orderings
    (e.g., generalization/specialization relations)

Preparing the input

- Denormalization is not the only issue when data is prepared for learning
- Problem: different data sources (e.g., sales department, customer billing department, ...)
  - Differences: styles of record keeping, coding conventions, time periods, data aggregation, primary keys, types of errors
  - Data must be assembled, integrated, cleaned up
  - "Data warehouse": consistent point of access
- External data may be required ("overlay data")
- Critical: type and level of data aggregation

Unbalanced data

- Unbalanced data is a well-known problem in classification problems
  - One class is often far more prevalent than the rest
  - Example: detecting a rare disease
- Main problem: simply predicting the majority class yields high accuracy but is not useful
  - Predicting that no patient has the rare disease gives high classification accuracy
- Unbalanced data requires techniques that can deal with unequal misclassification costs
  - Misclassifying an afflicted patient may be much more costly than misclassifying a healthy one

Getting to know your data

- Simple visualization tools may be very useful
  - Nominal attributes: histograms (Is the distribution consistent with background knowledge?)
  - Numeric attributes: graphs (Any obvious outliers?)
- 2-D and 3-D plots show dependencies
- May need to consult domain experts
- Too much data to inspect manually? Take a sample!

Inteligência Artificial

Introdução a Aprendizado Não-supervisonado

Aprendizado - paradigmas

- Aprendizado supervisionado
  - O crítico comunica a EA o erro relativo entre a ação que deve ser tomada idealmente pelo EE e a ação efetivamente escolhida pelo agente. Pares (corretos) de entrada/saída podem ser observados (ou demonstrados por um supervisor).
- Aprendizado por reforço
  - O crítico comunica apenas uma indicação de desempenho (indicação de quão bom ou ruim é o estado resultante), por vezes de modo intermitente e apenas quando situações dramáticas são atingidas (*feedback* indireto, com retardo).
- Aprendizado não-supervisionado
  - O crítico não envia nenhum tipo de informação ao EA, não há "pistas" sobre as saídas corretas (geralmente utiliza-se regularidades, propriedades estatísticas dos dados sensoriais)
    - Busca-se encontrar padrões ou estruturas / **agrupamentos** nos dados. técnicas

Clustering

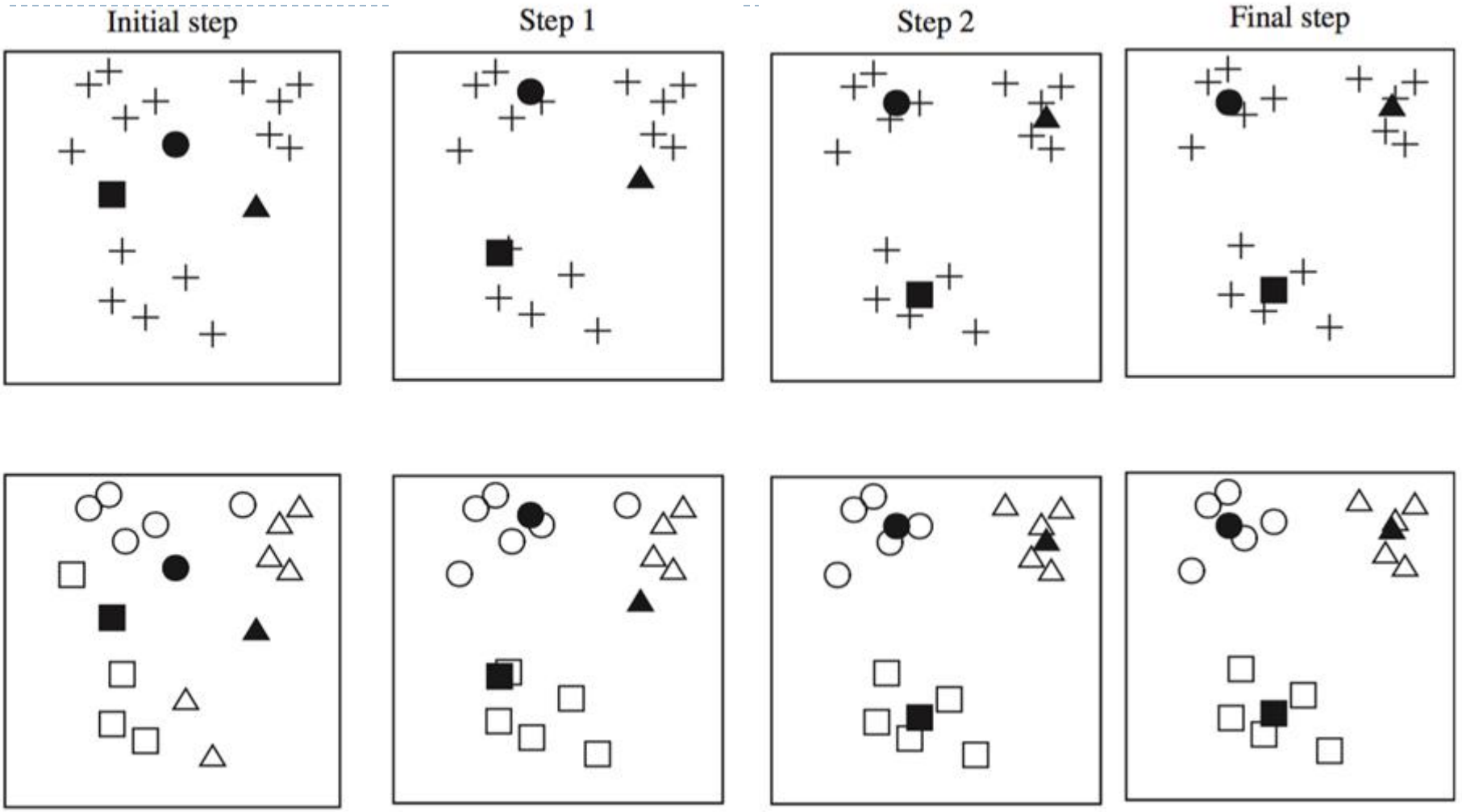
- Clustering techniques apply when there is no class to be predicted: they perform unsupervised learning
- Aim: divide instances into "natural" groups
- We will look at a classic clustering algorithm called *k-means*
- *k-means* clusters are disjoint, deterministic, and flat
- The basic idea is grouping instances (data points) that are "closer" to the others. In another, one group is formed by the instances with smaller **distance** among themselves
- Most algorithms use Eudcliden distance. The distance between an instance $a^{(1)}$ and instance $a^{(2)}$ each with k attributes is given by:

$$\sqrt{\left(a_1^{(1)}-a_1^{(2)}\right)^2+\left(a_2^{(1)}-a_2^{(2)}\right)^2+\ldots+\left(a_k^{(1)}-a_k^{(2)}\right)^2}.$$

The k-means algorithm

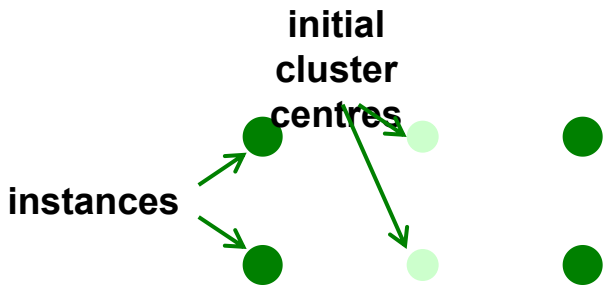
- Step 1: Choose *k* random cluster centers
- Step 2: Assign each instance to its closest cluster center based on Euclidean distance
- Step 3: Recompute cluster centers by computing the average (aka *centroid*) of the instances pertaining to each cluster
- Step 4: If cluster centers have moved, go back to Step 2
- This algorithm minimizes the squared Euclidean distance of the instances from their corresponding cluster centers
  - Determines a solution that achieves a *local* minimum of the squared Euclidean distance
- Equivalent termination criterion: stop when assignment of instances to cluster centers has not changed

The k-means algorithm: example



Discussion

- Algorithm minimizes squared distance to cluster centers
- Result can vary significantly
  - based on initial choice of seeds
- Can get trapped in local minimum
  - Example:



- To increase chance of finding global optimum: restart with different random seeds
- Can we applied recursively with $k = 2$

Resumo

- Vimos aqui o algoritmo básico para aprendizado não-supervisionado…Mas há diversas técnicas para agrupamento (k-means, hieráquico, …)

- Outras técnicas exploram o uso de redes complexas e teoria de grafos para fazer associação entre instâncias.

- Nos últimos anos, tem sido tratado também o chamado aprendizado “semi-supervisionado”, onde existem algumas poucas instâncias classificadas e uma grande quantidade de instâncias com classes desconhecidas. (Russel, cap. 18)