UNIVERSITÉ DE NEUCHÂTEL

MASTER IN COMPUTER SCIENCE

Course 32032: Machine Learning and Data Mining

Chapter 2

Input: concepts, instances, attributes

Prof. Marcelo Pasin

UNINE / University of Neuchatel
HES-SO / University of Applied Sciences and Arts Western Switzerland

Fall 2020

Contents

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
- Attributes: measuring aspects of an instance
  - We will focus on nominal and numeric ones

Concept

- Concept: thing to be learned
- Concept description: output of learning scheme
- Styles of learning:
  - Classification
  - Association
  - Clustering
  - Numeric prediction

Classification learning

- Example problems
  - Weather data, contact lenses, irises, labour 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
  - Far more association rules than classification rules
  - Constraints are necessary, such as minimum coverage and minimum accuracy

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 predictions

- 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

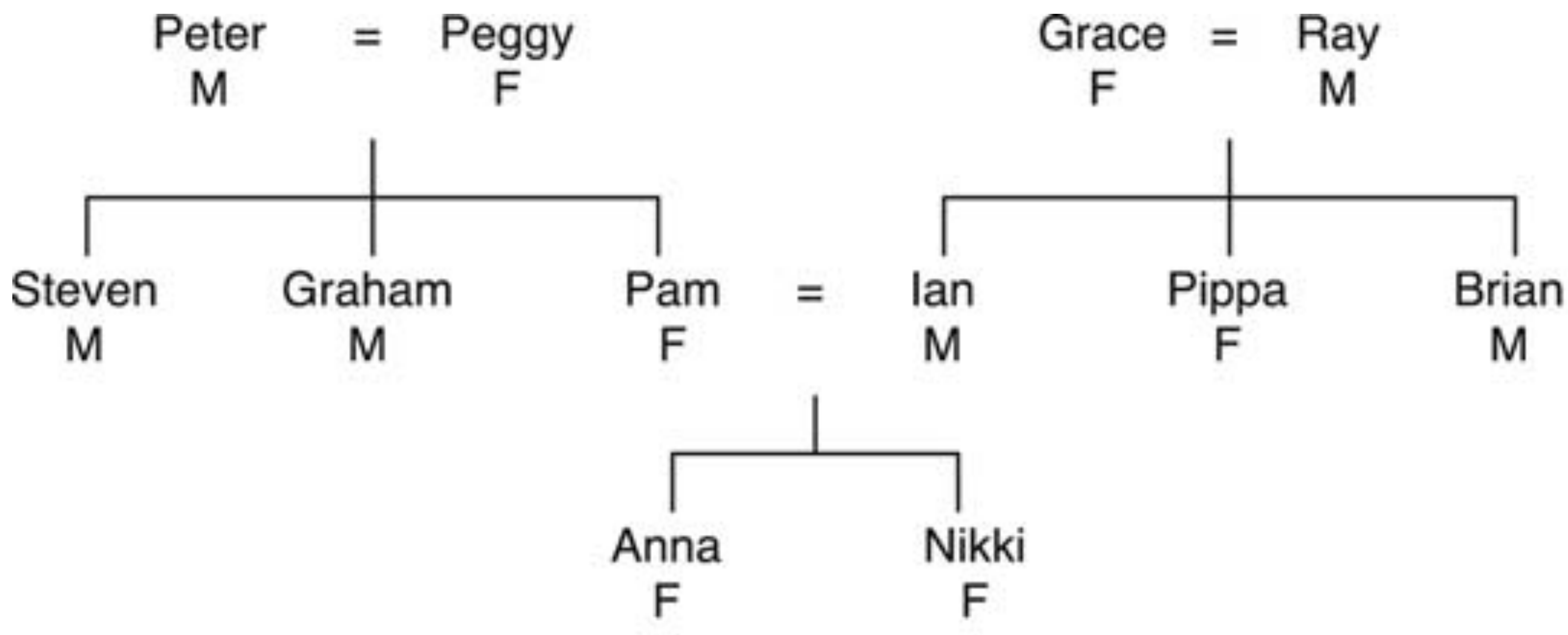
Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
…	…	…	…	…

Instances (examples of data)

- 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 scheme:
  set of instances/dataset
  - Represented as a single relation/flat file
- Rather restricted form of input
  - No relationships between objects
- Most common form in practical data mining

Family Tree

- Given a family tree, determine concept “sister of”



Siste-of relations

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

Family tree 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

Single table representation

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
```

Generating the single table

- Process of flattening called “denormalization”
  - Several relations are joined together to make one
  - Possible with any finite set of finite relations
  - Problematic: relationships without a pre-specified number of objects (ex. all siblings)

- May also produce spurious regularities reflecting the structure of the database
  - Example: “supplier” predicts “supplier address”

The “ancestor-of” relation

```
If person1 is a parent of person2
   then person1 is an ancestor of person2

If person1 is a parent of person2
   and person2 is an ancestor of person3
   then person1 is an ancestor of person3
```

- Infinite relations require recursion
- Appropriate techniques are known as “inductive logic programming” (ILP) methods
- Not covered in this course

Attribute

- Each instance is described by a fixed predefined set of features, its “attributes”
  - Columns in the dataset
- Attribute types (“levels of measurement”):
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal values

- Values are distinct symbols
  - Values themselves serve only as labels or names
  - Nominal comes from the Latin word for name
- Ex.: attribute "outlook" from weather data
  - Values: "sunny", "overcast", and "rainy"
- No relation is implied among nominal values
  - No ordering nor distance measure
- Only equality tests can be performed

Ordinal values

- Impose order on values
  - No distance between values defined
- Example: “temperature” in weather data
  - Values: “hot” > “mild” > “cool”
- Note: addition and subtraction don’t make sense
- Example rule:

  if **temperature < hot** then **play = yes**
- Distinction between nominal and ordinal not always clear (e.g., “sunny”, “overcast”, “rainy”)

Interval values

- Interval quantities are not only ordered but measured in fixed and equal units
- Examples
  - "temperature" expressed in degrees Fahrenheit
  - "year"
- Difference of two values makes sense
- Sum or product doesn't make sense
  - Zero point is not defined

Ratio values

- Ratio quantities are ones for which the measurement scheme defines a zero point
  - Example: attribute “distance”
  - Distance between an object and itself is zero
- Ratio quantities are treated as real numbers
  - All mathematical operations are allowed
- Is there an “inherently” defined zero point?
  - Answer depends on scientific knowledge
    - Fahrenheit knew no lower limit to temperature

Attributes, in practice

- Many data mining schemes accommodate just two levels of measurement: nominal and ordinal
- Others deal exclusively with ratio quantities
- Nominal attributes are also called “categorical”, “enumerated”, or “discrete”
  - But: “enumerated” and “discrete” imply order
  - Special case: dichotomy (“boolean” attribute)
- Ordinal attributes are sometimes coded as “numeric” or “continuous”
  - But: “continuous” implies mathematical continuity

Sparse data

- In some applications most attribute values are zero and storage requirements can be reduced
  - E.g.: word counts in a text categorization problem
- This also works for nominal attributes
  - The first value of the attribute corresponds to “zero”
- Some learning algorithms work very efficiently with sparse data
- File formats as ARFF support sparse data storage

```
0, 26, 0,  0, 0 ,0, 63, 0, 0, 0, “class A”
0,  0, 0, 42, 0, 0,  0, 0, 0, 0, “class B”
```

```
{1 26, 6 63, 10 “class A”}
{3 42, 10 “class B”}
```

Nominal x ordinal

- Attribute “age” nominal

```
If age = young and astigmatic = no
    and tear production rate = normal
    then recommendation = soft
If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal
    then recommendation = soft
```

- Attribute “age” ordinal
  “young” < “pre-presbyopic” < “presbyopic”

```
If age ≤ pre-presbyopic and astigmatic = no
   and tear production rate = normal
   then recommendation = soft
```

Missing values

- Missing values are frequently indicated by out-of-range entries for an attribute
  - There are different types of missing values: unknown, unrecorded, irrelevant
  - Reasons:
    - Malfunctioning equipment
    - Changes in experimental design
    - Collation of different datasets
    - Measurement not possible
- Missing value may have significance in itself (e.g., missing test in a medical examination)
  - Most schemes assume that is not the case and “missing” may need to be coded as a separate attribute value

Innacurate values

- Input data may contain errors
  - Reason: data has not been collected for mining it
  - Some errors may not affect the original purpose of the data (e.g., age of customer)
  - Errors may be deliberate (e.g., wrong zip codes)
  - Result: errors and omissions that affect the accuracy of data mining
- Typographical errors in nominal attributes: values need to be checked for consistency
- Typographical and measurement errors in numeric attributes: outliers need to be identified
- Other problems: duplicates, stale data

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