



Data Representation

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

- Describing Data
 - Concepts
 - Instances
 - Attributes
- Attributes Types
 - Nominal
 - Ordinal
 - Interval
 - Ratio
- Missing values and inaccurate values

Describing the Data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

- Instances
 - The atomic elements of information from a dataset
 - Also known as records, prototypes, or examples
- Attributes
 - Measures aspects of an instance
 - Also known as features or variables
 - Each instance is composed of a certain number of attributes
- Concepts
 - Special content inside the data
 - Kind of things that can be learned

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Two Versions of the Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...

Attribute Types

- Values are distinct symbols
- Values themselves serve only as labels or names
- Example
 - Attribute “outlook” from weather data
 - Values: “sunny”, “overcast”, and “rainy”
- Characteristics
 - No relation is implied among nominal values
 - No ordering
 - No distance measure
 - Only equality tests can be performed

- Impose order on values
- No distance between values defined
- Example
 - The attribute “temperature” in weather data
 - Values: “hot” > “mild” > “cool”
- Characteristics
 - Addition and subtraction don’t make sense
 - Distinction between nominal and ordinal not always clear (e.g. attribute “outlook”)

- Attribute “age” nominal
 - If age = young and astigmatic = no and tear production rate = normal then recommendation = soft
- Attribute “age” ordinal (e.g. “young” < “pre-presbyopic” < “presbyopic”)
 - If age \leq pre-presbyopic and astigmatic = no and tear production rate = normal then recommendation = soft

- Not only ordered but measured in fixed and equal units
- Examples
 - Attribute “temperature” expressed in degrees
 - Attribute “year”
- Characteristics
 - Difference of two values makes sense
 - Sum or product doesn’t make sense
 - Zero point is not defined

- Ratio quantities are ones for which the measurement scheme defines a zero point
- Example
 - Attribute “distance”
- Characteristics
 - 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?
 - It depends on scientific knowledge

- Check for valid values
- Express the best possible patterns into data
- Make the most adequate comparisons

- Example
 - Outlook > “sunny” does not make sense, while
 - Temperature > “cool” or
 - Humidity > 70 does

- Additional uses of attribute type
 - Check for valid values
 - Deal with missing values, etc.

- Most schemes accommodate just two or three levels of measurement: nominal, ordinal and numerical
- Boolean are a special case of nominal attribute

Missing Values

```
@relation labor
@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1,5,?, ?, ?,40, ?, ?,2, ?,11,'average', ?, ?, 'yes',?, 'good'
2,4.5,5.8, ?, ?,35,'ret_allw', ?, ?, 'yes',11,'below_average', ?, 'full', ?, 'full', 'good'
?, ?, ?, ?, ?,38,'empl_contr', ?,5, ?,11,'generous','yes','half','yes','half','good'
3,3.7,4,5,'tc', ?, ?, ?, ?, 'yes', ?, ?, ?, ?, 'yes', ?, 'good'
```

- Faulty equipment, incorrect measurements, missing cells in manual data entry, censored/anonymous data
- Very frequent in questionnaires for medical scenarios
- Censored/anonymous data
- In practice, a low rate of missing values may be suspicious
- Interview data provide many examples
 - For whom will you cast your vote in the next election?
 - What is your salary? Did you ever ...?

- Frequently indicated by out-of-range entries
- Missing value may have significance in itself
 - E.g. missing test in a medical examination
- Most schemes assume that is not the case
 - “missing” may need to be coded as additional value
- Does absence of value have some significance?
 - If it does, “missing” is a separate value
 - If it does not, “missing” must be treated in a special way

- Missing completely at random (MCAR): when the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data
- Missing at random (MAR), when the distribution of an example having a missing value for an attribute depends on the observed data, but does not depend on the missing data
- Not missing at random (NMAR), when the distribution of an example having a missing value for an attribute depends on the missing values.
- The handling of missing data depends on the type

- Discarding examples with missing values
 - Simplest approach
 - Allows the use of unmodified data mining methods
 - Only practical if there are few examples with missing values. Otherwise, it can introduce bias
- Convert the missing values into a new value
 - Use a special value for it
 - Add an attribute that indicates if value is missing or not
 - Greatly increases the difficulty of the data mining process
- Imputation methods
 - Assign a value to the missing one, based on the rest of the dataset. Use the unmodified data mining methods.

- Extract a model from the dataset to perform the imputation
- Suitable for MCAR and, to a lesser extent, for MAR
- Not suitable for NMAR type of missing data
- For NMAR we need to go back to the source of the data to obtain more information
- Survey of imputation methods available at
<http://sci2s.ugr.es/MVDM/index.php>
<http://sci2s.ugr.es/MVDM/biblio.php>

- Simply use the default policy of the data mining method
- Works only if the policy exists

- Replaces the missing value with the most common value available in the dataset
- Nominal Attributes: replace the missing value with the most frequent value of the attribute for the dataset
- Interval/Numerical Attributes: replace a missing value with the mean value of the attribute for the dataset (simple, but assumes that each attribute presents a normal distribution)

Concept Most Common (CMC) value

- Refinement of the MC policy
- The MV is replaced with the mean/most frequent value computed from the instances *belonging to the same class*
- Assumes that the distribution for an attribute of all instances from the same class is normal

Inaccurate Values

- Data has not been collected for mining it
- Errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes, thus values need to be checked for consistency
- Typographical and measurement errors in numeric attributes, thus outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)

Data Formats

- Most commercial tools have their own proprietary format
- Most tools import excel files and comma-separated value files

```
Year , Make , Model , Length  
1997 , Ford , E350 , 2 . 34  
2000 , Mercury , Cougar , 2 . 38
```

```
Year ; Make ; Model ; Length  
1997 ; Ford ; E350 ; 2 , 34  
2000 ; Mercury ; Cougar ; 2 , 38
```

```
%  
% ARFF file for weather data with some numeric features  
%  
@relation weather  
  
@attribute outlook {sunny, overcast, rainy}  
@attribute temperature numeric  
@attribute humidity numeric  
@attribute windy {true, false}  
@attribute play? {yes, no}  
  
@data  
sunny, 85, 85, false, no  
sunny, 80, 90, true, no  
overcast, 83, 86, false, yes  
...
```

<http://www.cs.waikato.ac.nz/~ml/weka/arff.html>

- ARFF supports string attributes:

```
@attribute description string
```

- Similar to nominal attributes but list of values is not pre-specified

- ARFF also supports date attributes:

```
@attribute today date
```

- Uses the ISO-8601 combined date and time format `yyyy-MM-dd-TTHH:mm:ss`

- Interpretation of attribute types in ARFF depends on the mining scheme
- Numeric attributes are interpreted as
 - Ordinal scales if less-than and greater-than are used
 - Ratio scales if distance calculations are performed (normalization/standardization may be required)
- Instance-based schemes define distance between nominal values (0 if values are equal, 1 otherwise)
- Integers in some given data file: nominal, ordinal, or ratio scale?

- Open format by Google available at <http://code.google.com/apis/publicdata/>
- Use existing data: add an XML metadata file existing CSV
- Read by the Google Public Data Explorer, which includes animated bar chart, motion chart, and map visualization
- Allow linking to concepts in other datasets
- Geo-enabled: allows adding latitude and longitude data to your concept definitions

Model Representation

- XML-based markup language developed by the Data Mining Group (DMG) to provide a way for applications to define models related to predictive analytics and data mining
- The goal is to share models between applications
- Vendor-independent method of defining models
- Allow to exchange of models between applications.
- PMML Components: data dictionary, data transformations, model, mining schema, targets, output

Data Repository

- UCI repository
 - <http://archive.ics.uci.edu/ml/>
 - Probably the most famous collection of datasets
- Kaggle
 - It is not a static repository of datasets, but a site that manages Data Mining competitions
 - Example of the modern concept of crowdsourcing

- KDNuggets
 - <http://www.kdnuggets.com/datasets/>
- PSPbenchmarks
 - <http://www.infobiotic.net/PSPbenchmarks/>
 - Datasets derived from Protein Structure Prediction problems
 - Interesting benchmarks because they can be parametrised in a very large variety of ways
- Pascal Large Scale Learning Challenge
 - <http://largescale.ml.tu-berlin.de/about/>

Summary

- Instances or examples are the atomic elements of a dataset
- Instances are described by a set of attributes (or variables)
- Attributes can be of different types (nominal, ordinal, interval, ratio)
- Values can be missing
- Missing values can be discarded, converted to a special value, or imputed