

Data Representation



Universiteit Leiden



Universiteit Leiden

The popular table

- Table (relation)
 - propositional, attribute-value
- Example
 - record, row, instance, case
 - independent, identically distributed
- Table represents a sample from a larger population
- Attribute
 - variable, column, feature, item
- Target attribute, class
- Sometimes rows and columns are swapped
 - bioinformatics

A	B	C	D	E	F
...
...
...



Example: symbolic weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

← attributes

→ examples



Example: symbolic weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

← attributes

↙ ↘ ↙ ↘ examples

↙ target attribute



Example: symbolic weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



Example: symbolic weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



three examples covered,
100% correct



Example: symbolic weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



three examples covered,
100% correct

if Outlook = sunny and Humidity = high then play = no

...

if Outlook = overcast then play = yes



Example: symbolic weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



three examples covered,
100% correct

if Outlook = sunny and Humidity = high then play = no

...

if Outlook = overcast then play = yes

...



Numeric weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

numeric attributes



Numeric weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	85 (hot)	85	false	no
sunny	80 (hot)	90	true	no
overcast	83 (hot)	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

numeric attributes



Numeric weather data

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

if Outlook = sunny and Humidity > 83 then play = no

if Temperature < Humidity then play = no



UCI Machine Learning Repository

The screenshot shows the UCI Machine Learning Repository website. The browser window title is "UCI Machine Learning Repository - Windows Internet Explorer". The address bar shows "http://archive.ics.uci.edu/ml/". The page features the UCI logo (University of California, Irvine) and the text "Machine Learning Repository" and "Center for Machine Learning and Intelligent Systems". A search bar is visible with a "Search" button. Below the header, there is a "Welcome to the UC Irvine Machine Learning Repository!" message. A paragraph of text provides information about the repository's data sets and links to "all data sets", "about page", "citation policy", and "donation policy". Below this, there are logos for "Supported By:" (UC Irvine) and "In Collaboration With:" (Rexa.info). The main content area is divided into three columns: "Latest News", "Newest Data Sets", and "Most Popular Data Sets (hits since 2007)". The "Most Popular Data Sets" column lists various data sets, with the "Iris" data set (ID 147513) circled in red. The "Featured Data Set" section highlights "Chess (King-Rook vs. King)" with a small image of a chessboard and details: Task: Classification, Data Type: Multivariate, # Attributes: 6, # Instances: 2056. The footer of the browser window shows "Internet" and "100%" zoom.

UCI Machine Learning Repository - Windows Internet Explorer

http://archive.ics.uci.edu/ml/

UCI Machine Learning Repository

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 190 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. Our [old web site](#) is still available, for those who prefer the old format. For a general overview of the Repository, please visit our [about page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#). We have also set up a [mirror site](#) for the Repository.

Supported By: In Collaboration With:

Latest News:

- 03-01-2010: [Note](#) from donor regarding Netflix data
- 10-16-2009: Two new data sets have been added.
- 09-14-2009: Several data sets have been added.
- 07-23-2008: [Repository mirror](#) has been set up.
- 03-24-2008: New data sets have been added!
- 06-25-2007: Two new data sets have been added: UCI Pen Characters, MAGIC Gamma Telescope
- 04-13-2007: Research papers that cite the repository have been associated to specific data sets.

Featured Data Set: [Chess \(King-Rook vs. King\)](#)

Task: Classification
Data Type: Multivariate
Attributes: 6
Instances: 2056

Chess Endgame Database for White King and Rook against Black

Newest Data Sets:

- 07-06-2010: [Opinion/Review](#)
- 02-09-2010: [p53 Mutants](#)
- 01-21-2010: [Demospongiae](#)
- 10-29-2009: [Parkinsons Telemonitoring](#)
- 10-15-2009: [URL Reputation](#)
- 10-07-2009: [Wine Quality](#)
- 08-17-2009: [Libras Movement](#)
- 07-13-2009: [Communities and Crime](#)

Most Popular Data Sets (hits since 2007):

- 147513: [Iris](#)
- 110235: [Adult](#)
- 97079: [Wine](#)
- 80049: [Breast Cancer Wisconsin \(Diagnostic\)](#)
- 63330: [Abalone](#)
- 60389: [Poker Hand](#)
- 54767: [Car Evaluation](#)
- 48819: [Forest Fires](#)

Internet 100%



CPU performance data (regression)

MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
125	256	6000	256	16	128	198	199
29	8000	32000	32	8	32	269	253
29	8000	32000	32	8	32	220	253
26	8000	32000	64	8	32	318	290
23	16000	64000	64	16	32	636	749
23	32000	64000	128	32	64	1144	1238
400	1000	3000	0	1	2	38	23
400	512	3500	4	1	6	40	24
60	2000	8000	65	1	8	92	70
350	64	6	0	1	4	10	15
200	512	16000	0	4	32	35	64
...

MYCT: machine cycle time in nanoseconds

MMIN: minimum main memory in kilobytes

MMAX: maximum main memory in kilobytes

CACH: cache memory in kilobytes

CHMIN: minimum channels in units

CHMAX: maximum channels in units

PRP: published relative performance

ERP: estimated relative performance from the original article

numeric target
attributes
(Regression,
numeric prediction)



CPU performance data (regression)

MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
125	256	6000	256	16	128	198	199
29	8000	32000	32	8	32	269	253
29	8000	32000	32	8	32	220	253
26	8000	32000	64	8	32	318	290
23	16000	64000	64	16	32	636	749
23	32000	64000	128	32	64	1144	1238
400	1000	3000	0	1	2	38	23
400	512	3500	4	1	6	40	24
60	2000	8000	65	1	8	92	70
350	64	6	0	1	4	10	15
200	512	16000	0	4	32	35	64
...

Linear model of Published Relative Performance:

$$\text{PRP} = -55.9 + 0.0489 \cdot \text{MYCT} + 0.0153 \cdot \text{MMIN} + 0.0056 \cdot \text{MMAX} + 0.641 \cdot \text{CACH} - 0.27 \cdot \text{CHMIN} + 1.48 \cdot \text{CHMAX}$$



Soybean disease data

- Michalski and Chilausky, 1980
- 'Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis.'
- 680 examples, 35 attributes, 19 categories
- Two methods:
 - rules induced from 300 selected examples
 - rules acquired from plant pathologist
- Scores:
 - induced model 97.5%
 - expert 72%



Soybean data

1. date: april,may,june,july,august,september,october,?.
2. plant-stand: normal,lt-normal,?.
3. precip: lt-norm,norm,gt-norm,?.
4. temp: lt-norm,norm,gt-norm,?.
5. hail: yes,no,?.
6. crop-hist: diff-1st-year,same-1st-yr,same-1st-two-yrs, same-1st-sev-yrs,?.
7. area-damaged: scattered,low-areas,upper-areas,whole-field,?.
8. severity: minor,pot-severe,severe,?.
9. seed-tmt: none,fungicide,other,?.
10. germination: 90-100%,80-89%,lt-80%,?.
- ...
32. seed-discolor: absent,present,?.
33. seed-size: norm,lt-norm,?.
34. shriveling: absent,present,?.
35. roots: norm,rotted,galls-cysts,?.



Soybean data

1. date: **april,may,june,july,august,september,october,?**.
2. plant-stand: normal,lt-normal,?.
3. precip: lt-norm,norm,gt-norm,?.
4. temp: lt-norm,norm,gt-norm,?.
5. hail: yes,no,?.
6. crop-hist: **diff-1st-year,same-1st-yr,same-1st-two-yrs, same-1st-sev-yrs,?**.
7. area-damaged: scattered,low-areas,upper-areas,whole-field,?.
8. severity: minor,pot-severe,severe,?.
9. seed-tmt: none,fungicide,other,?.
10. germination: **90-100%,80-89%,lt-80%,?**.
- ...
32. seed-discolor: absent,present,?.
33. seed-size: norm,lt-norm,?.
34. shriveling: absent,present,?.
35. roots: norm,rotted,galls-cysts,?.



Types

■ Nominal, categorical, symbolic, discrete

- only equality (=)
- no distance measure

■ Numeric

- inequalities ($<$, $>$, \leq , \geq)
- arithmetic
- distance measure

■ Ordinal

- inequalities
- no arithmetic or distance measure

■ Binary

- like nominal, but only two values, and True (1, yes, y) plays special role.



ARFF files

```
%  
% 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  
...
```



Other data representations

- time series

- uni-variate
- multi-variate



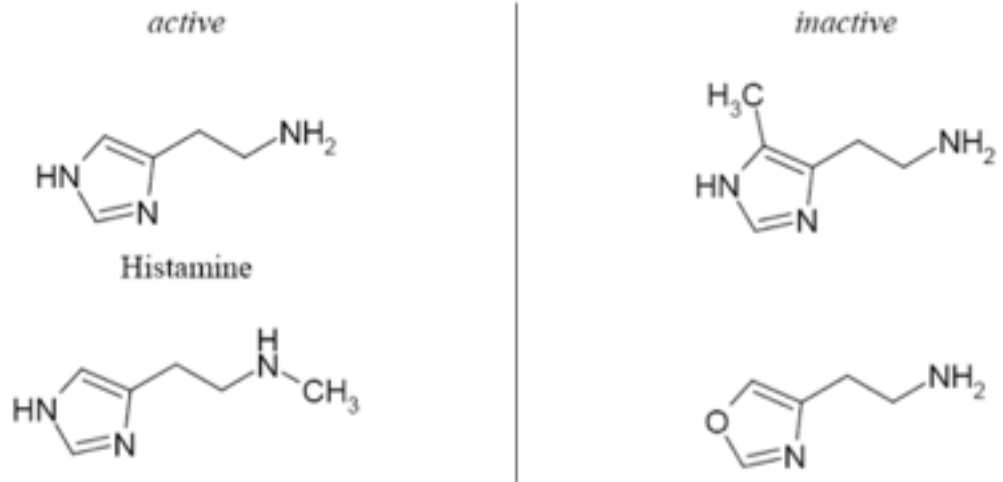
- Data streams

- stream of discrete events, with time-stamp
- e.g. shopping baskets, network traffic, webpage hits



Other representations

- Database of graphs



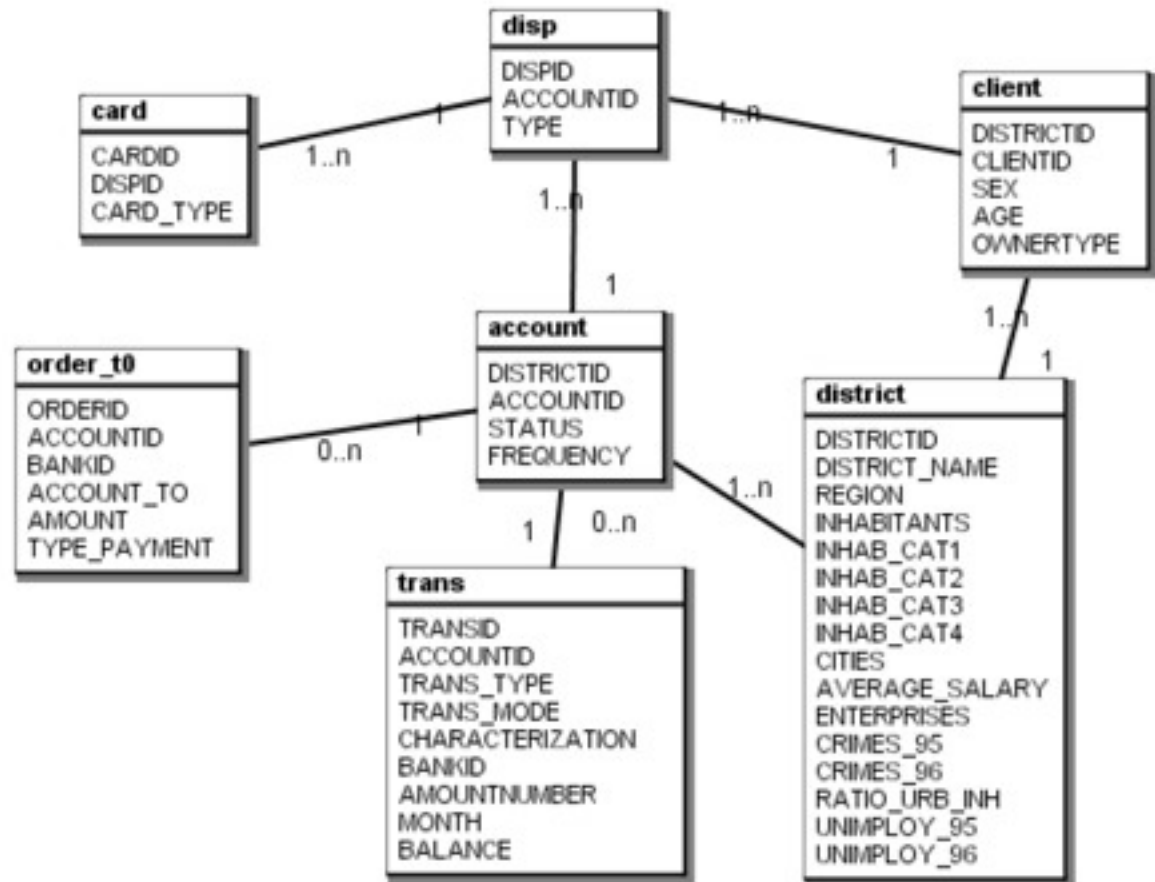
- Large graphs
 - social networks

vizster



Other representations

■ Multi-relational data



Assignment

- Direct Marketing in holiday park
 - Campaign for new offer uses data of previous booking:
 - customer id
 - price
 - number of guests
 - class of house
 - arrival date
 - departure date
 - positive response? (target)
- } data from previous booking
-
- Question: what alternative representations for the 2 dates can you suggest? The (multiple) new attributes should make explicit those features of a booking that are relevant (such as holidays etc).

