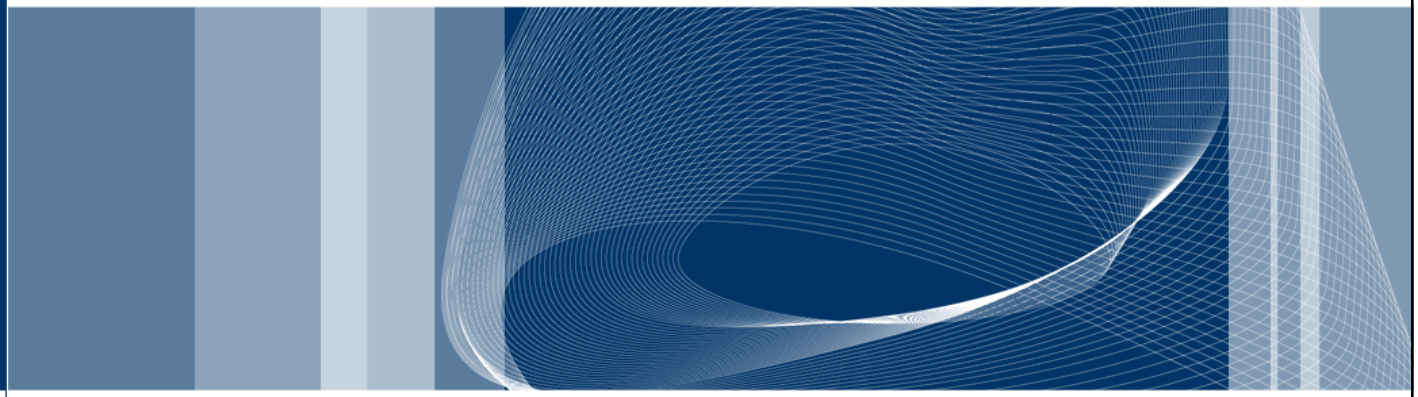
POLITECNICO DI MILANO



Data Representation

Data Mining and Text Mining

Prof. Pierluca Lanzi

describing data

The Weather Dataset

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

Instances, Attributes, Concepts

- Instances (aka observations, cases, records, items, examples)
  - The atomic elements of information from a dataset
  - Each row in previous table corresponds to an instance
- Attributes (aka variables, features, independent variables)
  - Measures aspects of an instance
  - Each instance is composed of a certain number of attributes
  - Each column in previous table contains values of an attribute
- Concept (aka class, target variable, dependent variable)
  - Special content inside the data
  - Kind of things that can be learned
  - Intelligible and operational concept description
  - Last column of previous table was the class

attribute types

Attributes

- Numeric Attributes
  - Real-valued or integer-valued domain
  - Interval-scaled when only differences are meaningful (e.g., temperature)
  - Ratio-scaled when differences and ratios are meaningful (e.g., Age)
- Categorical Attributes
  - Set-valued domain composed of a set of symbols
  - Nominal when only equality is meaningful (e.g., domain(Sex) = { M, F})
  - Ordinal when both equality (are two values the same?) and inequality (is one value less than another?) are meaningful (e.g., domain(Education) = { High School, BS, MS, PhD})

Numerical Attributes

- 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
- Sometimes they are divided into "discrete" and "continuous"

Nominal Attributes (or Categorical)

- Values are distinct symbols that 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

Other Types of Attributes

- Ordinal Attributes
  - Categorical attributes with an imposed order on values
  - No distance between values defined
  - For instance, temperature encoded as "hot","mild", and "cool"
  - Size encoded as "small", "medium", "large", and "jumbo"
- Ratio Attributes
  - Numerical attributes for which the measurement scheme defines a zero point (e.g., an attribute representing distance)
- Binary Attributes
  - Represented by just two values 0/1

example

Contraceptive Method Choice Data Set

https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

[2]:
```
df.head(20)
```

[2]:

	Age	WifeEducation	HusbandEducation	NumberOfChildren	WifeReligion	WifeNowWorking	HusbandOccupation	StandardOfLivingIndex	MediaExposure	MethodUsed
0	24	2	3	3	1	1	2	3	0	1
1	45	1	3	10	1	1	3	4	0	1
2	43	2	3	7	1	1	3	4	0	1
3	42	3	2	9	1	1	3	3	0	1
4	36	3	3	8	1	1	3	2	0	1
5	19	4	4	0	1	1	3	3	0	1
6	38	2	3	6	1	1	3	2	0	1
7	21	3	3	1	1	0	3	2	0	1
8	27	2	3	3	1	1	3	4	0	1
9	45	1	1	8	1	1	2	2	1	1
10	38	1	3	2	1	0	3	3	1	1
11	42	1	4	4	1	1	1	3	0	1
12	44	4	4	1	1	0	1	4	0	1
13	42	2	4	1	1	0	3	3	0	1
14	38	3	4	2	1	1	2	3	0	1
15	26	2	4	0	1	1	4	1	0	1
16	48	1	1	7	1	1	2	4	0	1
17	39	2	2	6	1	1	2	4	0	1
18	37	2	2	8	1	1	2	3	0	1
19	39	2	1	5	1	1	2	1	1	1

What are the attribute types of these variables?

Contraceptive Method Choice Data Set (Data Description)

- Wife's age (numerical)
- Wife's education (categorical) 1=low, 2, 3, 4=high
- Husband's education (categorical) 1=low, 2, 3, 4=high
- Number of children ever born (numerical)
- Wife's religion (binary) 0=Non-Islam, 1=Islam
- Wife's now working? (binary) 0=Yes, 1=No
- Husband's occupation (categorical) 1, 2, 3, 4
- Standard-of-living index (categorical) 1=low, 2, 3, 4=high
- Media exposure (binary) 0=Good, 1=Not good
- 10. Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term

What are the attribute types of these variables (now)?

another perspective

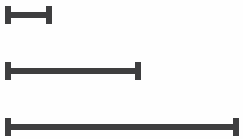
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

Attribute types

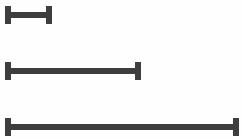
Attribute Types

- Categorical

e.g., gender, race, eye color

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

Attribute types

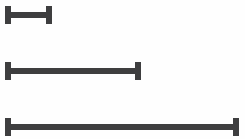
Attribute Types

- Categorical

- Ordered
  - *Ordinal*

e.g., edu level, ranking

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

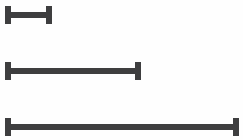
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



e.g., age, height, weight

Ordering Direction

- Sequential

- Diverging

- Cyclic

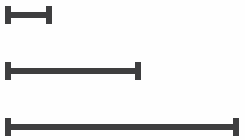
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

e.g., age, height, weight

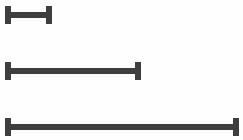
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

e.g., temperature, altitude

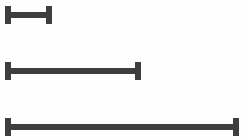
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

e.g., hour, week

hierarchies

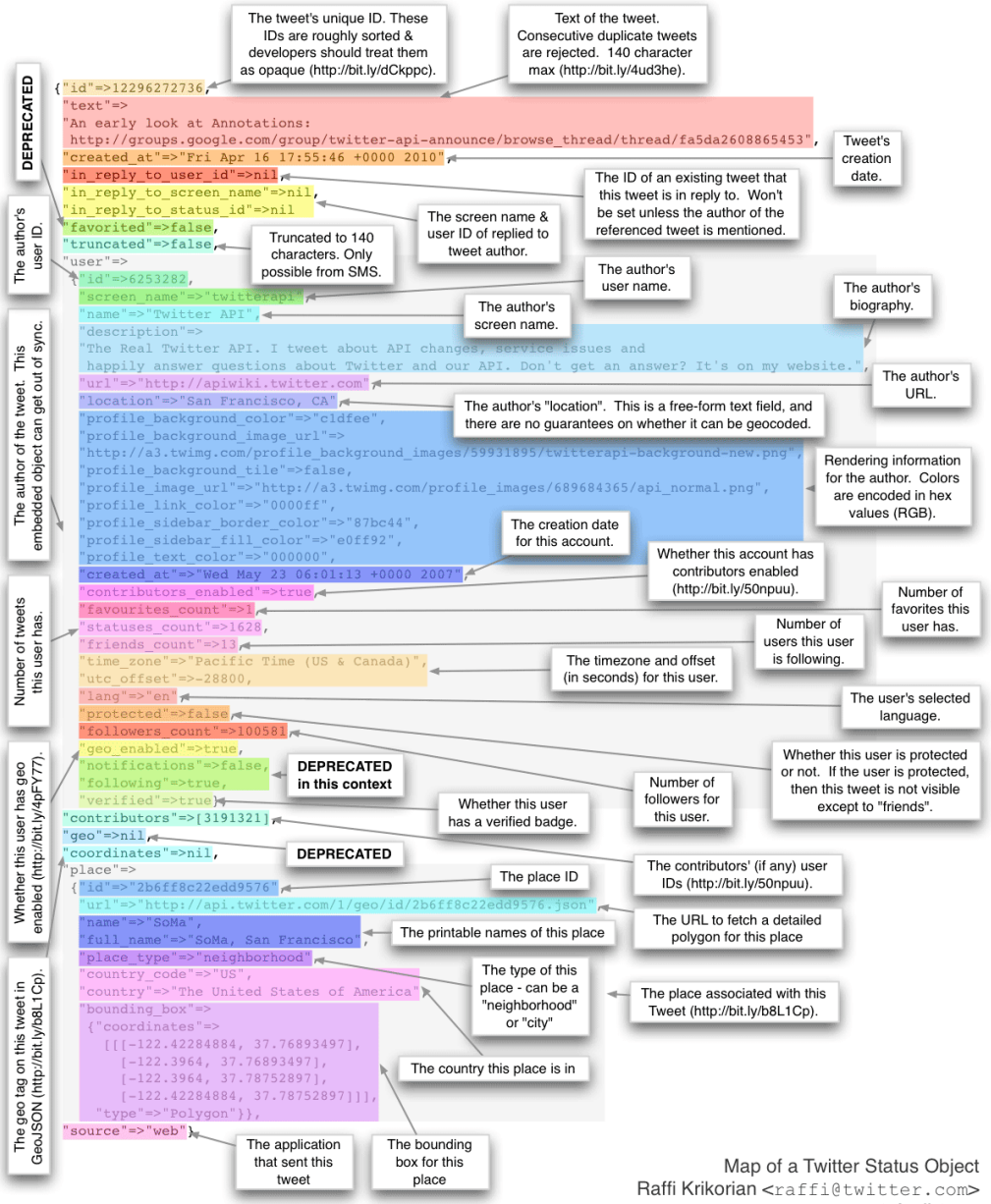
Some attributes may have an internal hierarchical structure

For example, dates, mail addresses, spatial regions, taxonomies, etc.

missing values

Why Missing Values Exist?

- Faulty equipment, incorrect measurements, missing cells in manual data entry, censored/anonymous data
- Review scores for movies, books, etc.
- Very frequent in questionnaires for medical scenarios
- Censored or anonymous data
- Data from rich representations for which many of the fields might not be used (e.g., Twitter data)



http://socialmedia-class.org/twittertutorial.html

Missing Values

- Missing value may have significance in itself
  - E.g. missing test in a medical examination or an empty field in a questionnaire
- They are frequently indicated by out-of-range entries (e.g. max/min float), NaN or special values (e.g., zero)
- Most schemes assume that is not the case and "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

What Types of Missing Values?

- Missing not at random (MNAR)
  - Distribution of missing values depends on missing value
  - E.g., respondents with high income less likely to report it
- Missing at random (MAR)
  - Distribution of missing values depends on observed attributes, but not missing value
  - E.g., men less likely than women to respond to question about mental health

What Types of Missing Values?

- Missing completely at random (MCAR)
  - Distribution of missing values does not depend on observed attributes or missing value
  - E.g., survey questions randomly sampled from larger set of possible questions
- Identifying MNAR and MAR can be difficult often requires domain knowledge

Dealing with Missing Values

- Use what you know
  - Why data is missing
  - Distribution of missing data
- Decide on the best strategy to yield the least biased estimates
  - Deletion Methods (listwise deletion, pairwise deletion)
  - Single Imputation Methods (mean/mode substitution, dummy variable method, single regression)
  - Model-Based Methods (maximum Likelihood, multiple imputation

Strategies for missing values handling

- The handling of missing data depends on the type
- Discarding all the examples with a 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
- Fill in the missing value manually ☺
- 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.

Do Not Impute (DNI)

- Simply use the default policy of the data mining method
- Works only if the policy exists
- Some methods can work around missing data

List-wise Deletion (Complete Case Analysis)

- Only analyze cases with available data on each variable
- Simple, but reduces the data
- Comparability across analyses
- Does not use all the information
- Estimates may be biased if data not MCAR

Gender	8th grade math test score	12th grade math score
~~F~~	~~45~~	.
~~M~~	.	~~99~~
F	55	86
F	85	88
F	80	75
~~.~~	~~81~~	~~82~~
F	75	80
~~M~~	~~95~~	.
M	86	90
F	70	75
~~F~~	~~85~~	.

Pairwise deletion (Available Case Analysis)

- Analysis with all cases in which the variables of interest are present
- Example
  - When using only the first two variables, the missing values of the third variable are not considered
- Advantage
  - Keeps as many cases as possible for each analysis
  - Uses all information possible with each analysis
- Disadvantage
  - Can't compare analyses because sample different each time

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Single Imputation Methods

- Mean/mode substitution (most common value)
  - Replace missing value with sample mean or mode
  - Run analyses as if all complete cases
  - Advantages: Can use complete case analysis methods
  - Disadvantages: Reduces variability
- Dummy variable control
  - Create an indicator for missing value (1=value is missing for observation; 0=value is observed for observation)
  - Impute missing values to a constant (such as the mean)
  - Include missing indicator in the algorithm
  - Advantage: uses all available information about missing observation
  - Disadvantage: results in biased estimates, not theoretically driven
- Regression Imputation
  - Replaces missing values with predicted score from a regression equation.

Imputation 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

inaccurate values

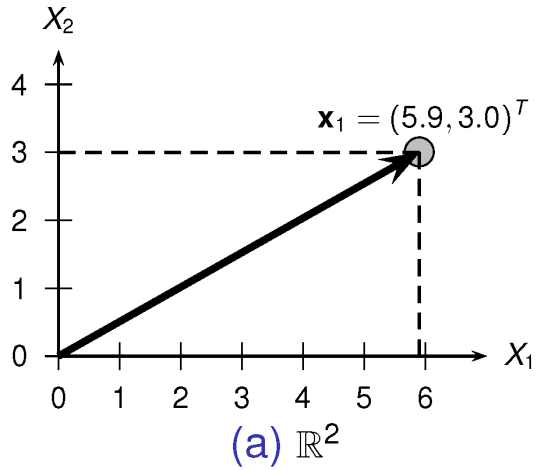
Inaccurate Values

- 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)

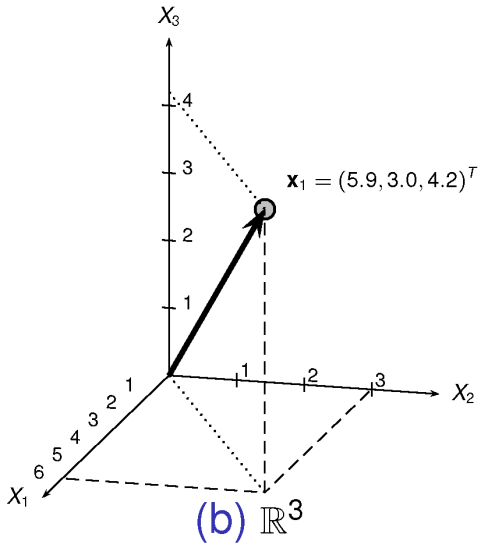
the geometric view

The Geometrical View of the Data

- When the data contains only numerical values
  - Every row can be viewed as a point in a d-dimension space
  - Every column as a point in a n-dimensional space



(a) $\mathbb{R}^2$



(b) $\mathbb{R}^3$

From categorical attributes to numerical ones and Vice-versa

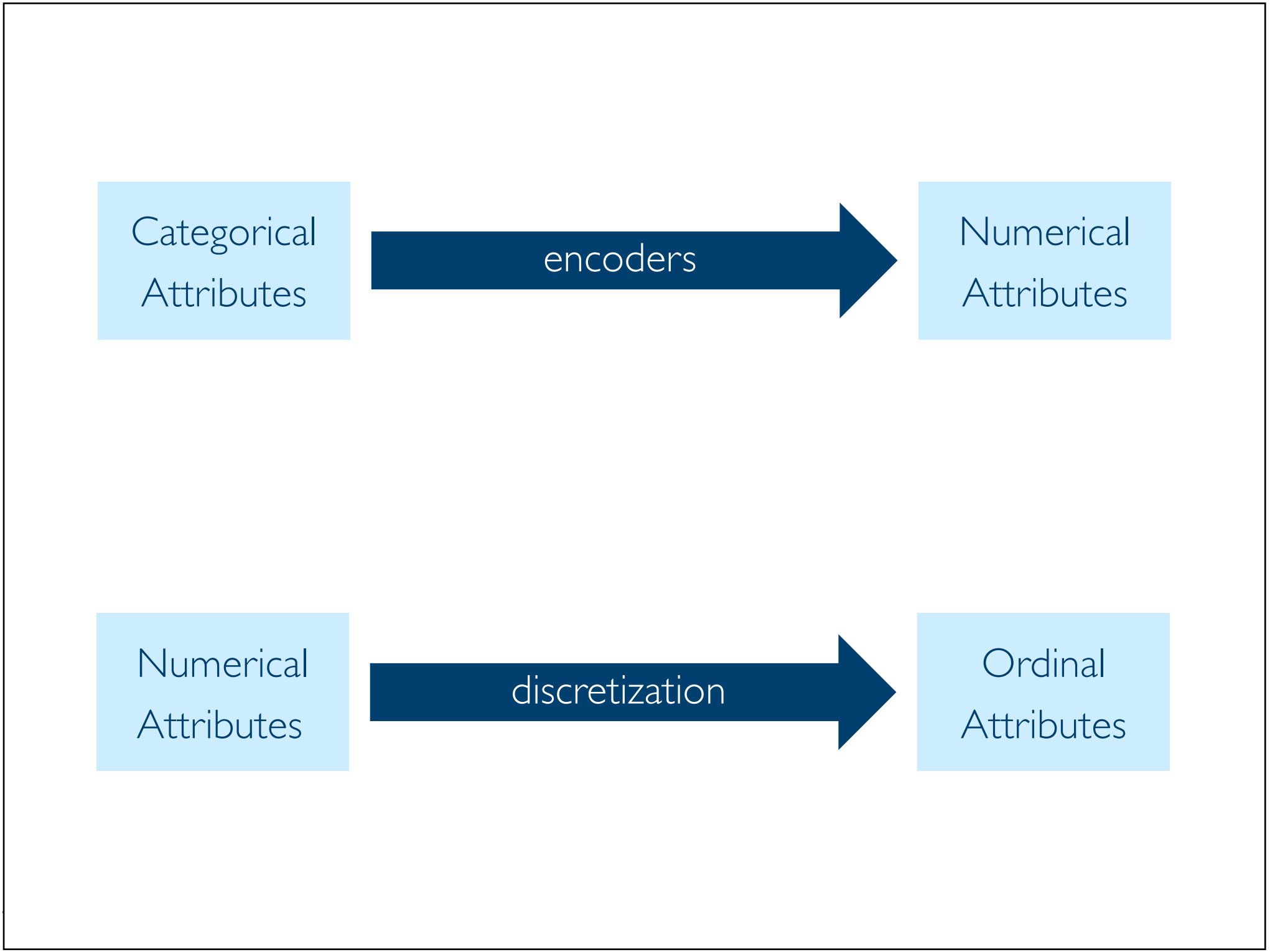
Why do we care about data types?

They influence the type of statistical analyses and visualization we can perform

Some algorithms and functions fit some specific data types best

Check for valid values

Deal with missing values, etc.


Categorical Attributes
encoders
Numerical Attributes
Numerical Attributes
discretization
Ordinal Attributes

Scikit-Learn Encoders

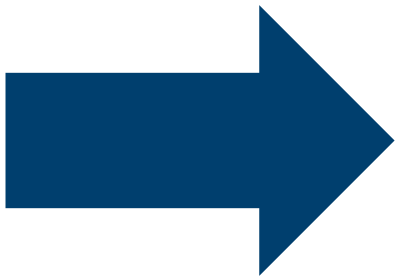
- LabelEncoder
  - Encodes target labels with values between 0 and n_labels-1
- OneHotEncoder
  - Performs a one-hot encoding of categorical features.
- OrdinalEncoder
  - Performs an ordinal (integer) encoding of the categorical features
- …

Label Encoder

Map a categorical variable described by n values
into a numerical variables with values from 0 to n-1

For example, attribute Outlook would be replaced by a
numerical variables with values 0, 1, and 2

Outlook
Sunny
Sunny
Overcast
Rainy
Rainy
Rainy
Overcast
Sunny
Sunny
Rainy
Sunny
Overcast
Overcast
Rainy



Outlook
2
2
0
1
1
1
0
2
2
1
2
0
0
1

Label Encoder for the Outlook attribute.

Warning

By replacing a label with a number might influence the process in unexpected ways

In the example, by assigning 0 to overcast and 2 to sunny we give a higher weight to the latter

What happens if we then apply a regression model?

Would the result change with different assigned values?

If we apply label encoding, we should store the mapping used for each attribute to be able to map the encoded data into the original ones

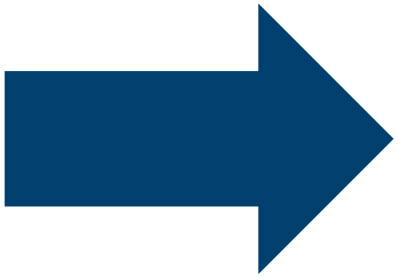
One Hot Encoding

Map each categorical attribute with
n values into n binary 0/1 variables

Each one describing one
specific attribute values

For example, attribute Outlook is replaced by three
binary variables Sunny, Overcast, and Rainy

Outlook
Sunny
Sunny
Overcast
Rainy
Rainy
Rainy
Overcast
Sunny
Sunny
Rainy
Sunny
Overcast
Overcast
Rainy



Outlook_Overcast	Outlook_Rainy	Outlook_Sunny
0	0	1
0	0	1
1	0	0
0	1	0
0	1	0
0	1	0
1	0	0
0	0	1
0	0	1
0	1	0
0	0	1
1	0	0
1	0	0
0	1	0

One Hot Encoding for the Outlook attribute.

Warning

One hot encoding assign the same numerical value (1) to all the labels

But it can generate a massive amount of variables when applied to categorical variables with many values

We will discuss discretization later …

data format

Data Format

- 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
```

Attribute-Relation File Format (ARFF)


```
%
% 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

Missing Values in ARFF


```
@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'
```

DSPL: Dataset Publishing Language

- Open format by Google available at
  http://code.google.com/apis/publicdata/
  https://code.google.com/archive/p/dspl/downloads
- 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

Predictive Model Markup Language

- 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 repositories

Publicly Available Datasets

- UCI repository
  - http://archive.ics.uci.edu/ml/
  - Probably the most famous collection of datasets
- Kaggle
  - http://www.kaggle.com/
  - 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/
- Google Data Search
  - https://datasetsearch.research.google.com

Readings

- "Data Mining and Analysis" – Chapter 1
- "Mining of Massive Datasets" – Chapter 1
- Survey of imputation methods
  http://sci2s.ugr.es/MVDM/index.php