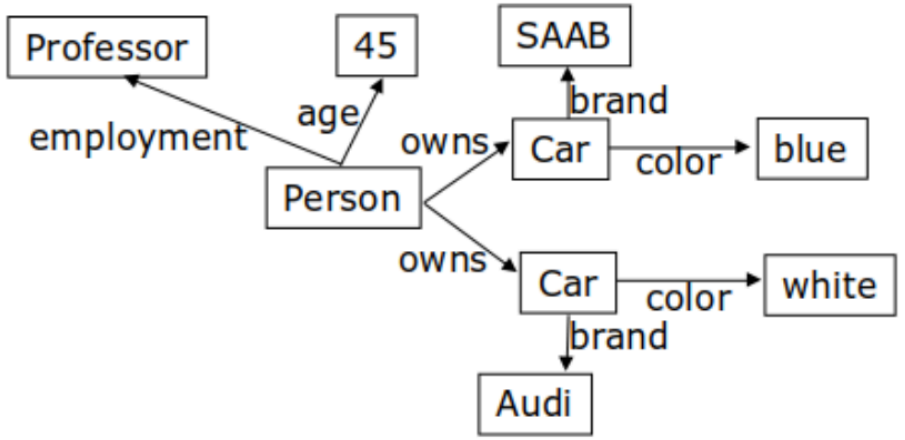
Data Preparation

02.04.2020

Data Preparation

- The instances need to be represented by fixed-length feature vectors
- For predictive modeling, labels have to be assigned to instances, and information from test instances should not affect choice of data preparation and learning algorithms
- There can be no missing, numerical or categorical values
- Numerical features have to be normalized
- The curse of dimensionality has to be remedied by limiting the number of features

Representation



Empl.	Age	Brand1	Color1	Brand2	Color2	Brand3	Color3	...
Prof.	45	SAAB	blue	Audi	white	-	-	...

Empl.	Age	No. SAABs	No. Audis	No. Volvos	...	No. blues	No. whites	No. greens	...
Prof.	45	1	1	0	...	1	1	0	...

Handling Missing Values

- Some techniques require missing values to be handled, by
  - removing them, i.e., removing rows and/or columns, or
  - replacing (imputing) them
- How to impute missing values is a research area of its own; with techniques ranging from replacing missing values with the mean or mode to more advanced methods relying on using values from nearest neighbors

Handling Missing Values in DataFrames

	id	grade	award
0	NaN	NaN	NaN
1	2.0	b	gold
2	3.0	NaN	silver
3	4.0	c	bronze
4	5.0	NaN	NaN

```
df.dropna(how='any')
```

	id	grade	award
1	2.0	b	gold
3	4.0	c	bronze

Handling Missing Values in DataFrames

	id	grade	award
0	NaN	NaN	NaN
1	2.0	b	gold
2	3.0	NaN	silver
3	4.0	c	bronze
4	5.0	NaN	NaN

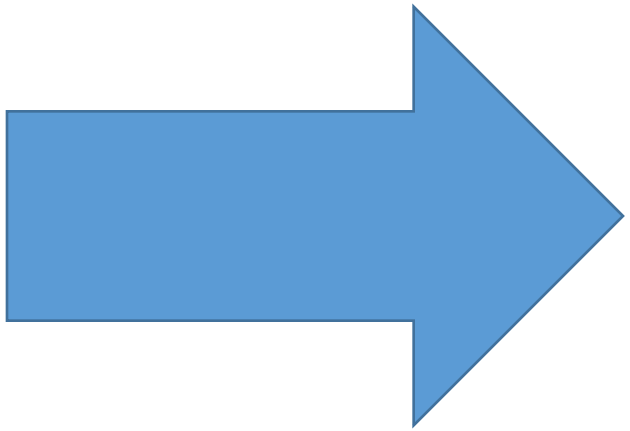
```
values = {'grade':'e','award':'iron'}
df.fillna(values)
```

	id	grade	award
0	NaN	e	iron
1	2.0	b	gold
2	3.0	e	silver
3	4.0	c	bronze
4	5.0	e	iron

```
df["id"].fillna(df["id"].mean(),inplace=True)
df["award"].fillna(df["award"].mode()[0],inplace=True)
```

Encoding Features

	value
0	a
1	b
2	c
3	a
4	b
5	c



	value_a	value_b	value_c
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	1	0
5	0	0	1

Encoding Features

min-max normalization

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

z-normalization

$$x_i' = \frac{x_i - \bar{x}}{s}$$