NOVA IMS
Information Management School

4

DATA PREPARATION

Machine Learning for Marketing

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

NOVA IMS Information Management School

Summary

1. Introduction
2. Data cleaning
3. Data reduction
4. Data transformation
5. Data integration
6. Application exercise

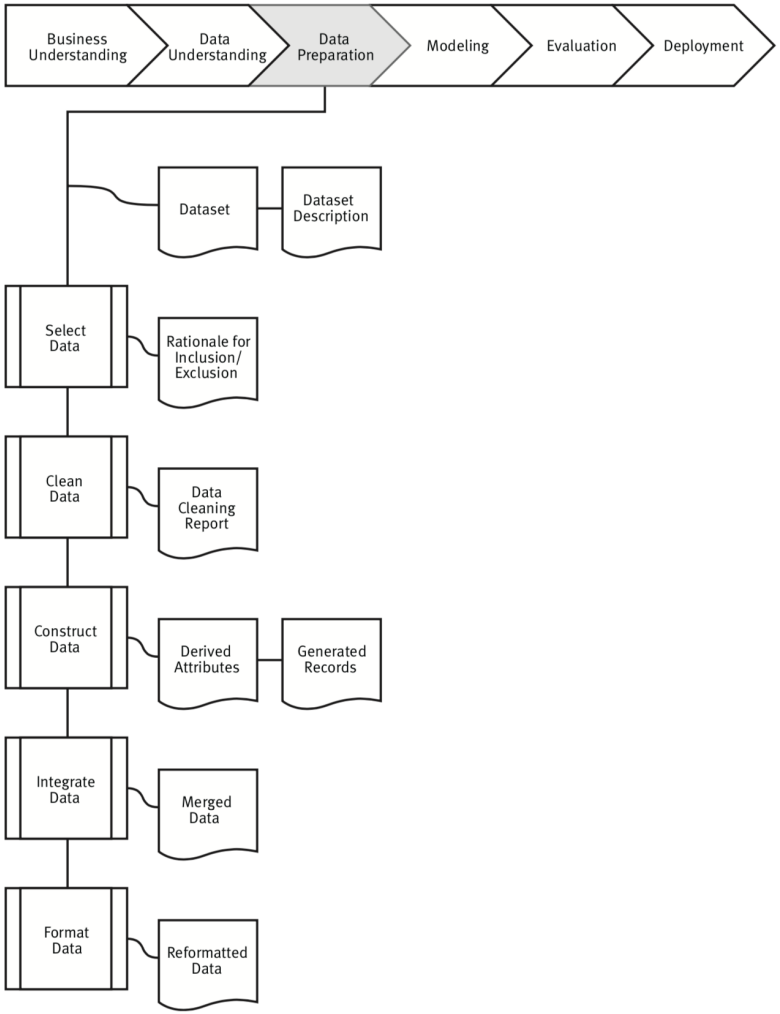
NOVA IMS Information Management School

4.1

Introduction

Data preparation

Data preparation



original dataset

modeling dataset aka Analytical Base Table (ABT)

Why data preparation

Due to their size and multiple, heterogenous sources, real-word databases commonly have:

- "Noise" (random error or variance)
- Missing data
- Inconsistent data

Low quality data → Low quality mining results

For this reason, data must be preprocessed and prepared to improve the efficiency and ease of the mining process.

NOVA IMS Information Management School

4.2

Data cleaning

Data preparation

Objective

Fix variables problems, such as:

- Duplicates
- Redundancy
- Incorrect or miscoded values
- Outliers
- Missing values

NOVA IMS Information Management School

Duplicates

It is common for real-world datasets to have duplicate instances, even when they should not exist (e.g., having two instances of the same customer profile)

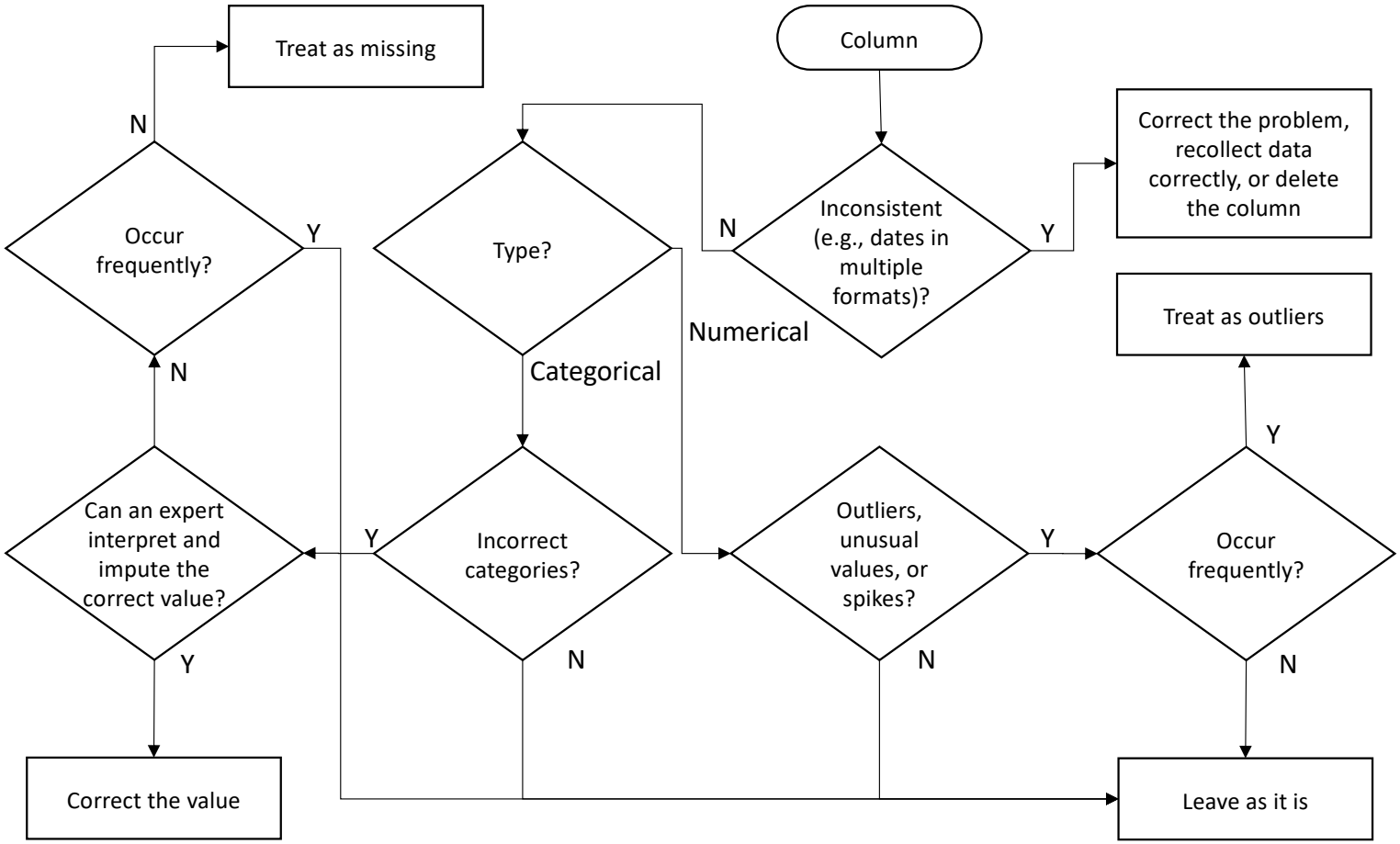
- If instances are exact-match duplicates, with all columns having the same values, most of the times all duplicated instances could be **deleted** (except one, off course)
- If some columns are not equal, (e.g., if there are two customer instances with the same name, telephone, and address but a different volume of purchases) **aggregations** may be required. In this example, sales could need to be summed up and one of the instances deleted after

Redundancy

When two attributes are redundant, one of them **should not be included** in the modeling dataset. Removing correlated attributes:

- Improves the model development speed
- Decrease harmful bias
- Increases interpretability

Incorrect or miscoded values



NOVA IMS Information Management School

Approaches to handling outliers

Data cleaning

Approaches to handling outliers (1/6)

Remove from the modeling data

When outliers distort the models more than they can help, in numeric algorithms (K-means clustering or Principal component analysis)

Risk of removing outliers:

⚠Model deployment can be compromised when outliers appear (produces unexpected scores)

Approaches to handling outliers (2/6)

Separate the outliers and create models just for them

- Relax the definition of outliers from two standard deviations from the mean to three standard deviations
- Create a separate model to identify outliers (e.g., linear regression)

⚠ Some algorithms, such as Decision trees-based algorithms already incorporate this approach in the algorithm design itself

Approaches to handling outliers (3/6)

Transform the outliers so they are no longer outliers

Apply skew transformation or normalization techniques to reduce the distance between the outliers and the main body of the distribution (e.g., $\log_{10}$)

Approaches to handling outliers (4/6)

Transform the outlier and create an indicator column

Apply skew transformation of normalization techniques as in the previous approach, but, additionally, create a dummy column indicating if the observation is an outlier (0: no; 1:yes)

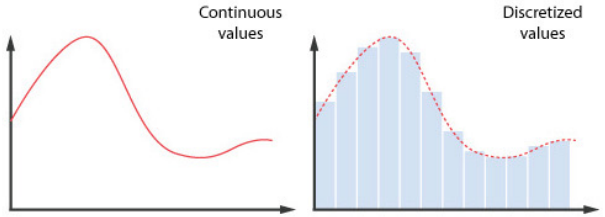
Approaches to handling outliers (5/6)

Bin the data (discretize the data)

Because transformations may not capture too extreme outliers, an alternative to transformations is to transform the numeric variable in categorical (e.g., instead of salary amount use *low*, *medium*, *high*)

Common binning options:

- **Equal-Frequency**: the number of unique values in all bins are similar
- **Equal-Width**: pre-defined or range-based width (dividing the range by the number of bins to define size)
- **Clustering**: dividing the data into discrete groups or clusters
- **Entropy**: boundaries are defined so that entropy is minimized in partitions



Approaches to handling outliers (6/6)

Leave in the data without modification

Employ on algorithms that are unaffected by outliers, such as Decision trees-based algorithms

NOVA IMS Information Management School

Approaches to handling missing values

Data cleaning

Approaches to handling missing values (1/6)

Listwise and column deletion

- If a small percentage of observations have columns with missing values, just remove those observations
- If a specific column has many missing values, consider removing it

Approaches to handling missing values (2/6)

Imputation with a constant

- For categorical variables, this is as simple as filling the missing values with a value indicating that is missing (e.g., “NULL”)
- For numeric variables, if the 0 (zero) makes sense (e.g., bank balance) then fill it with a 0 (zero). Otherwise, try other approach, like the “Mean or median imputation”

Approaches to handling missing values (3/6)

Mean and median imputation (for continuous variables)

One of the most common approaches in continuous variables is the imputation of the mean value. However, if the distribution is skewed, the median could be better

⚠ If the number of observations is large, this operation could be computationally expensive.

Approaches to handling missing values (4/6)

Imputations with distributions

In numeric variables, when a large percentage of values are missing, the summary statistics are affected by mean/median imputation. In these cases, the missing value should be replaced from a random number of a known distribution (based on the variable distribution)

Approaches to handling missing values (5/6)

Random imputation from own distributions

This approach involves for each missing value, randomly, select a value of one of the non-missing values existing on the column.

The advantage of this approach is that the distribution of imputed values matches the populated data.

Approaches to handling missing values (6/6)

Impute value from a model

This is the more complex approach. It involves developing a model to impute missing values.

⚠ This approach can take time

⚠ When deployed, requires that missing values in data to be also processed by the model

Additional consideration on missing values

Creation of dummy variables

In some cases, the existence of missing values can be informative for the model. In those cases, besides implementing one of the previous approaches, a dummy variable could be created to indicate if there is a missing value (0: no; 1:yes)

NOVA IMS Information Management School

4.3

Data reduction

Data preparation

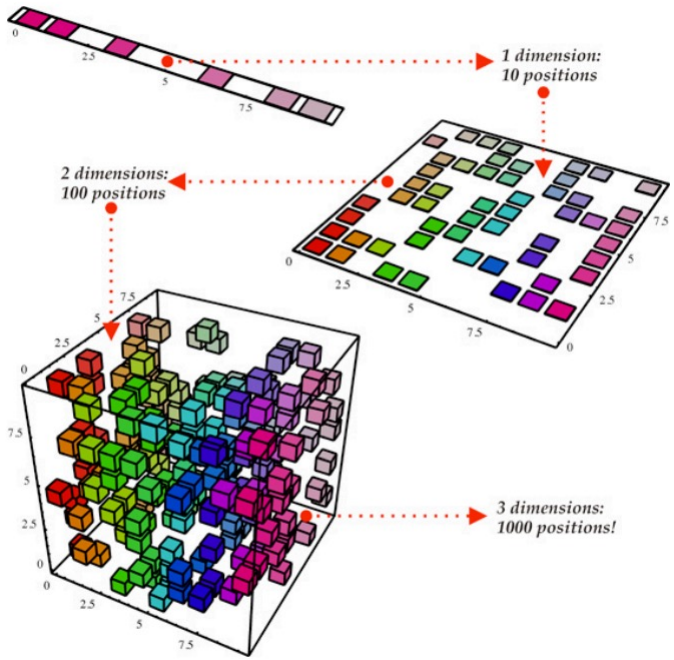
NOVA IMS Information Management School

Dimensionality reduction

Data reduction

The curse of dimensionality

As the number of candidate variables for modeling increase, the number of observations must also increase (exponentially) to be able to capture the high-dimensional patterns. One way to address this problem is to **reduce the number of dimensions**



source: http://www.turingfinance.com

Attribute subset selection

Datasets may contain hundreds of attributes, but many of which may be **irrelevant** to the mining task or **redundant**. For example, for segmenting customers, telephone number may be irrelevant

"The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes"

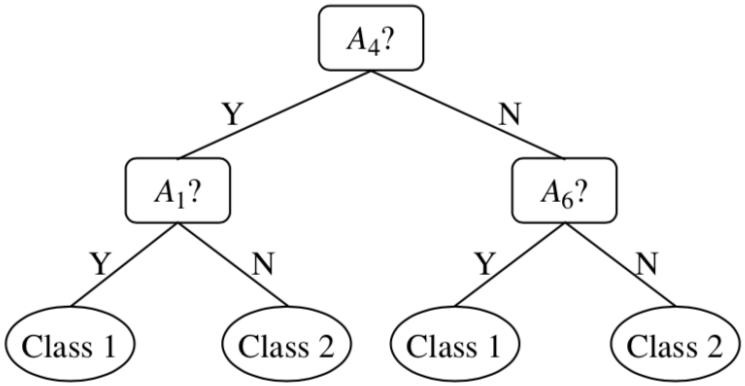
Han et al. (2012)

Attribute subset selection types

- **Wrapper**: uses ML to select features to use (forward selection, backward selection, or other method)
- **Filter**: uses statistics tests (Pearson correlation, Chi-squared, etc.)
- **Embedded**: included in the algorithm

Attribute subset selection

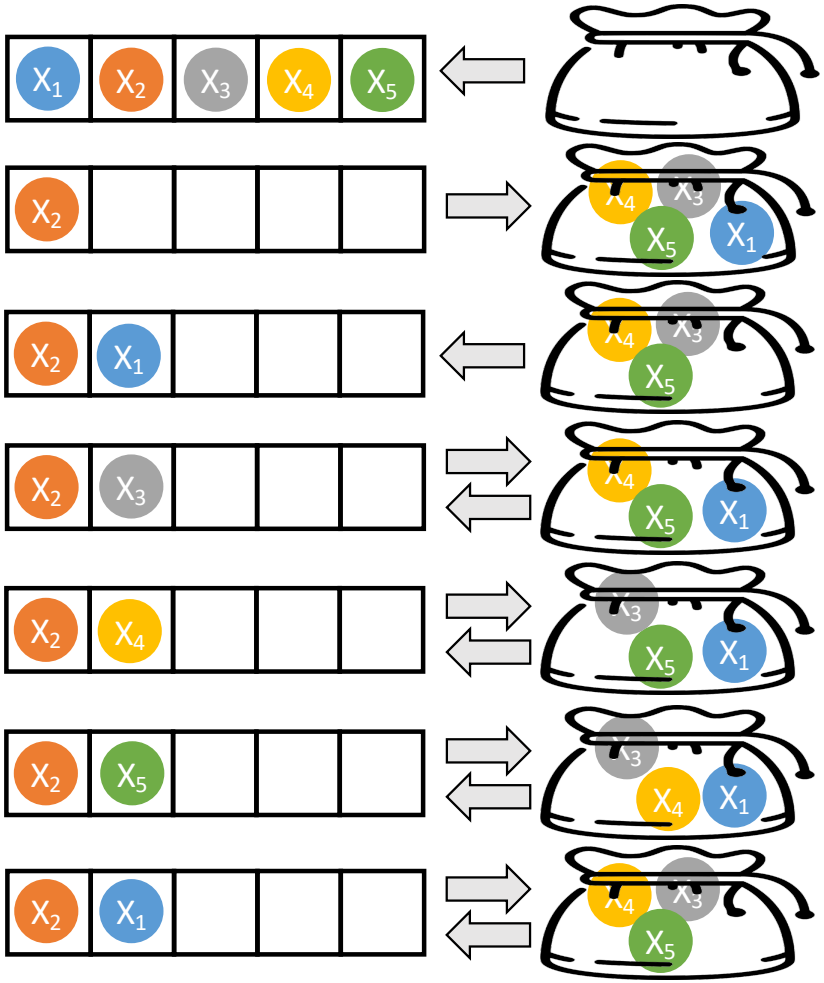
Greedy (heuristic) methods

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

Han et al. (2012)

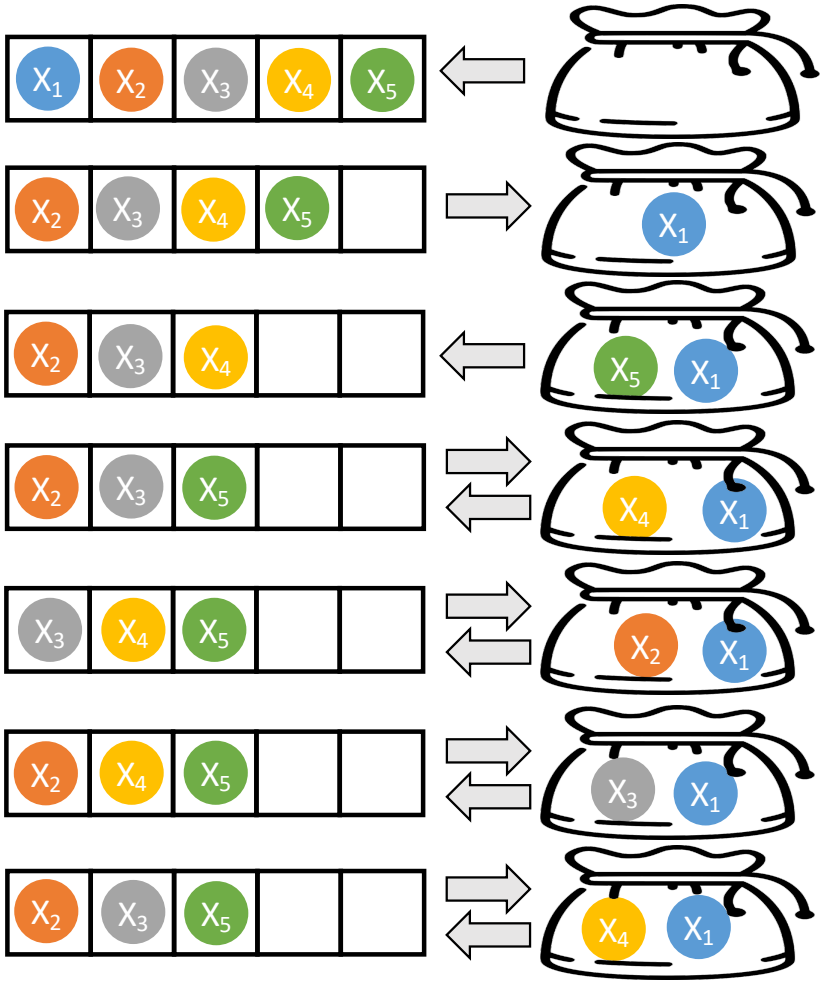
Forward selection

1. Build a model with all the input variables
2. Leave the variable with the smallest error
3. Add to the model each of the other variables, one at the time
4. Build a model for each of these combinations
5. The best two-variable model is kept
6. Repeat to add additional variables, until a predefined threshold is reached



Backward selection

1. Build a model with all the input variables
2. Remove the least significant of the variables (or verify the performance with all combinations), one at each time
3. Build a model
4. Repeat the process until a predefined threshold is reached



Selecting attributes prior to modeling

"The simplest way to select variables for predictive modeling when there are too many is to first run an assessment to score the predictive power for each variable in predicting the target variable, one at a time. " [Abbott, D. (2014)]

Technique	Comparison type	Inclusion metric
Chi-square test	Categorical input vs. categorical target	p-value or top N variables
CHAID tree stump using chi-square test	Continuous or categorial input vs. continuous or categorical target	p-value or top N variables
Association rules confidence, 1 antecedent	Categorical input vs. categorical target	Confidence, Support
ANOVA	Continuous input vs. categorical target	p-value or top N variables
Kolmogoroc-Smirnov (K-S) Distance, two sample test	Continuous input vs. continuous target	K-S test critical value, top N variables
Linear regression forward selection (1 step)	Continuous input (or dummy) vs. continuous target	p-value, AIC, MDL, top N variables
PCA (select top loader for each PC)	Continuous input vs. continuous target	Top N variables

NOVA IMS Information Management School

Other techniques for dimensionality reduction

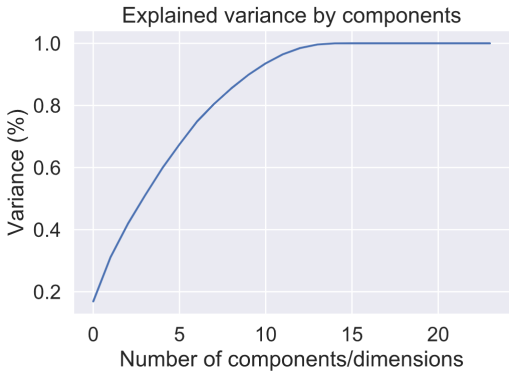
- **Principal Components Analysis (PCA)**: reduces dimensionality, while retaining as much variance in data as possible (finds a new set of variables that are a linear combination of the original variables)
- **Kernel PCA (KPCA)**: nonlinear variation of PCA
- **Linear Discriminant Analysis (LDA)**: unsupervised learning method that transforms a set of features to a new set
- **Singular Value Decomposition (SVD)**: extracts important features from data, while reconstructing the original dataset to a smaller dataset (e.g., transform a 1 024 pixels image to 66 pixels)
- Among others

Other techniques for dimensionality reduction
Principal Component Analysis (PCA)

Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, **PCA "combines" the essence of attributes by creating an alternative, smaller set of variables**

The basic procedure for PCA is:

1. Normalization of input data
2. PCA computes the k orthonormal vectors. These are unit vectors that each point in a direction perpendicular to the others
3. The principal components are sorted in order of decreasing "significance"
4. Because components are sorted in decreasing order of "significance", with the strongest principal components is possible to reconstruct a good approximation of the original data



NOVA IMS Information Management School

Numerosity reduction

Data reduction

Numerosity reduction (1/2)

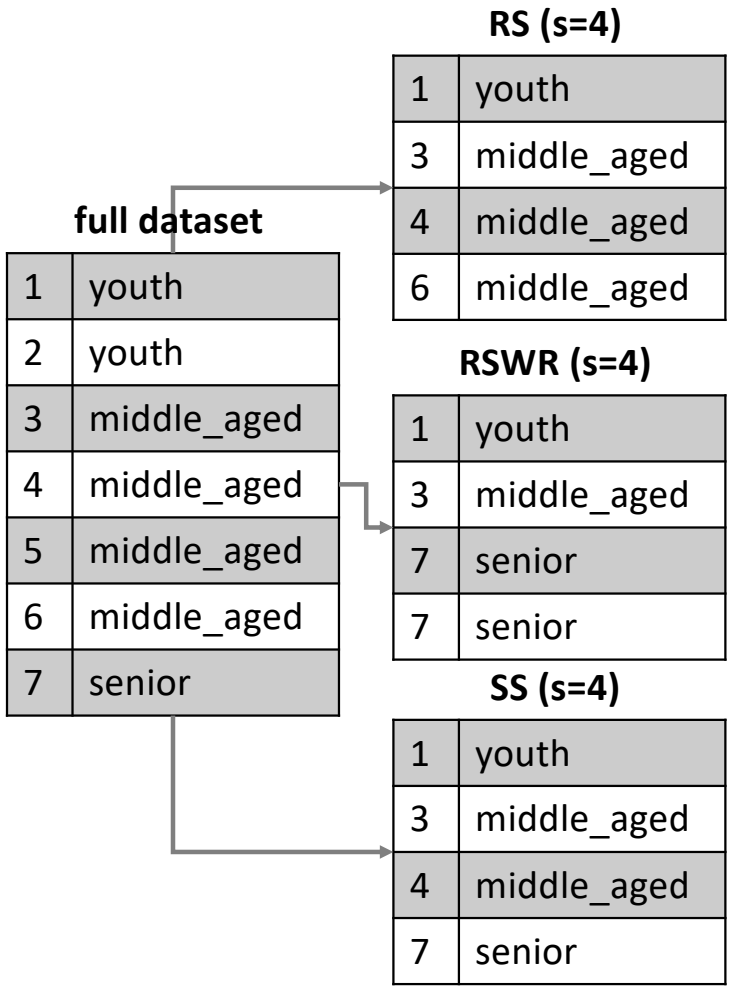
Methods:

- **Aggregations**: aggregate the data in a different unit of analysis (e.g., weekly data, instead of daily data)
- **Clustering**: cluster representations of the data are used to replace the actual data
- **Parametric data reduction**: regression and log-linear models are used to "predict" an output, based on a set of inputs (e.g., using multivariate linear regression to transform a set of variables in only one)

Numerosity reduction (2/2)

Methods (cont.):

- **Sampling**: allows a dataset to be represented by a smaller subset. Could be:
  - **Random sampling (RS)**: selects a random percentage of instances
  - **Random sampling with replacement (RSWR)**: similar to previous, but the same instance can be selected more than once
  - **Stratified sample (SS)**: selects instances accordingly the relative frequencies of the levels of a specified stratification feature. This selection ensures that the sample presents a distribution similar to the population



NOVA IMS Information Management School

4.4

Data transformation

Data preparation

NOVA IMS Information Management School

Normalization

Data transformation

Normalization

- Some algorithms, such as the K-MEANS algorithm, have difficulty in covering variables in very different ranges (e.g., age in the range of [15, 80] and salary in the range [30 000, 80 000]
- Linear regression coefficients are also influenced disproportionately by the large values of a skewed distribution
- Normalization can make a continuous variable fall within a specific range while maintaining the relative differences between the values for the variable

Common normalization techniques

Method	Formula	Range
Magnitude scaling	$x' = \frac{x}{\max(\lvert x \rvert)}$	[-1, 1]
Sigmoid	$x' = \frac{1}{(1+e^{-x})}$	[0, 1]
Min-max	$x' = \frac{(x-x_{min})}{(x_{max}-x_{min})}$	[0, 1]
Z-score	$x' = \frac{(x-\bar{x})}{\sigma_x}$	mostly [-3, 3]
Rank binning	$x' = \frac{(100 \times rank\ order)}{\#\ observations}$	[0, 100]

Measures scaling

Normalization techniques are also used to scale measurements in different scales to the same scale

Example

Rating A scale: [1, 5]

Rating B scale: [2.5, 10]

Min-max scale to convert an 8.1 rating in Rating B to 0-10 scale

$$\text{scale} = x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} = \frac{(8.1 - 2.5)}{(10 - 2.5)} = \frac{5.6}{7.5} = 0.7467 \times 10 = 7.5$$

Min-max scale to convert a 4 rating in Rating A to 0-10 scale

$$\text{scale} = x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} = \frac{(4 - 1)}{(5 - 1)} = \frac{3}{4} = 0.75 \times 10 = 7.5$$

NOVA IMS Information Management School

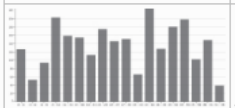
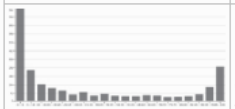
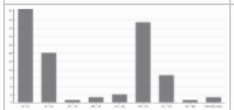
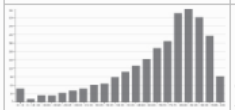
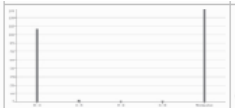
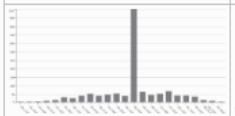
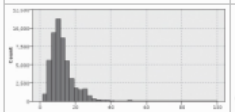
Feature engineering

Data transformation

Feature engineering

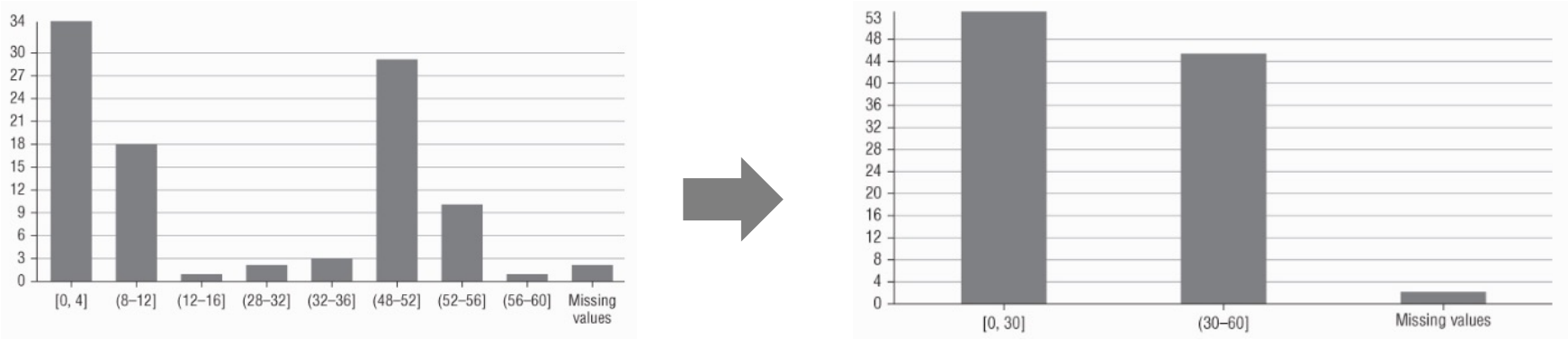
The creation of new features (also know as ”derived variables” or “derived attributes”) provides more value-added to the quality of data than any other modeling step

Distributions and possible "corrections"

Distribution	Possible Corrective Actions
	Maybe none; close enough to uniform.
	Primarily positive skew; consider log10 transform.
	Consider binning to capture four or five regions centered on spikes.
	Consider binning into two bins (for peaks) or four bins (two peak and two trough regions).
	Negative skew; consider power transform or flip transform and use log10.
	Consider dummy variables for spikes on left and right.
	Positive skew; consider a log transform.
	Spike in distribution, consider dummy indicator of the spike vs. non-spike. If spike generated through mean imputation, consider imputing with a distribution.
	Classic positive skew. Log10 transform.

Abbott (2014)

Binning (discretizing) variables



Abbott (2014)

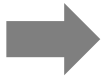
Other possible transformations

- Reciprocal transformation: $\frac{1}{x}$
- Square root transformation: $\sqrt{x}$
- Exponential transformation: e^x
- Box-cox transformation: $\begin{cases} \frac{x^\gamma - 1}{\gamma} & for\ \gamma \neq 0 \\ \log x & for\ \gamma = 0 \end{cases}$

Encode categorical variables (1/2)

Numerical algorithms such as Linear Regression or K-MEANS require inputs to be numerical. The most common approach is to create dummy variables (aka One-hot encoding)

CustomerID	Spent	Segment
1	€ 100	Corporate
2	€ 120	SME
3	€ 110	Individual



CustomerID	Spent	Corporate	SME	Individual
1	€ 100	1	0	0
2	€ 120	0	1	0
3	€ 110	0	0	1

Encode categorical variables (2/2)

Approach to handling high cardinality:

- Encode categorical variables using an encoder that does not generate a column for each value/level of the categorical variable (e.g., the count or probability of observations that have that value/level)
- If there is a hierarchy, consider using higher levels only . For example, if you have street, city, and region, consider using only city and region, or even just region
- For values/levels present in more than a predetermined threshold of observations (e.g., 2%) create dummy variables

CustomerID	Spent	Segment
1	€ 100	Corporate
2	€ 120	SME
3	€ 110	Individual
4	€ 105	Corporate

30%

CustomerID	Spent	Segment	Corporate
1	€ 100	2	1
2	€ 120	1	0
3	€ 110	1	0
4	€105	2	1

Date/time variables

Datasets are two-dimensional, so, when models require the introduction of time, transformations are necessary to include a third dimension (time).

Usually date/time variables are converted to numeric units related to the outcome to be predicted. For example:

- The date the customer was offered a quotation for a loan could be converted to the number of days before the mortgage was signed or the number of days since a certain date (e.g., 2000-01-01)
- The date of mortgage signature could be converted to the the day in the year or the week number

Multidimensional features

The most powerful of features. The two most common examples are:

- Interactions: multiplication of variables
- Ratios: division of variables

Usually, domain expertise is required to understand which interactions, and above all, which ratios may have modeling value.

Multidimensional features - ratios

Ratios are important because they are difficult for most algorithms to uncover. Ratios can:

- Provided a normalized version of a variable. For example, a percentage (e.g., a customer website purchase ratio = $\frac{number\ of\ purchases}{customer\ website\ visits}$)
- Can incorporate complex ideas. For example, the claims received to premiums paid in an insurance company = $\frac{claims\ received}{premiums\ paid}$
- Can make models to live "longer". For example, a model for real estate property value, instead of using each property price, due to prices increasing trend, could be = $\frac{property\ price\ (m2)}{average\ property\ price\ (m2)}$

NOVA IMS Information Management School

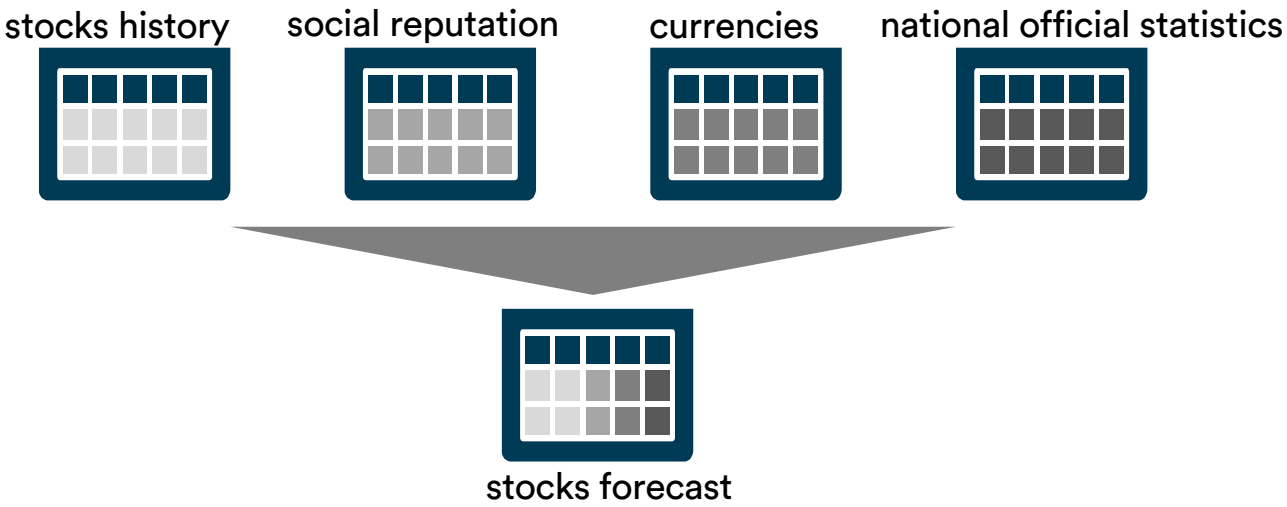
4.5

Data integration

Data preparation

Merge data

Joining data that comes from two or more databases about the unit of analysis under studied



Reformat data

Apply syntactic modifications that do not change data meaning, but are required for modeling, for example:

- Remove commas from text fields if the dataset is supposed to be saved as comma separated values
- Put the target/label in the first column (when there is one)
- Remove any ordering that might exist in the observations
- Trim some variables (e.g., text variables) to a certain maximum size

NOVA IMS Information Management School

4.6

Application exercise

Data preparation

Business problem

For an insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries. As a result, insurers invest a great deal of time and money to develop models that accurately forecast medical expenses.

Medical expenses are difficult to estimate because the costliest conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

[use case from Lantz, B (2013)]

Business objective

Estimate the medical care expenses per individual. These estimates could be used to create actuarial tables which set the price of yearly premiums higher or lower depending on the expected treatment costs

Understanding key drivers of the estimates

NOVA IMS Information Management School

Predict medical expenses

1. Copy from the datasets folder copy the dataset “medical_expenses.csv”
2. Copy and open the Jupyter notebook “PredictMedicalExpenses_DataPrep.ipynb”
3. Follow the presentation of the notebook, answer questions, and explore the challenges

Questions?

Machine Learning for Marketing

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa