Topics in Business Intelligence

Lecture 2: Data reduction

Tommi Tervonen

Econometric Institute, Erasmus School of Economics

ERASMUS UNIVERSITEIT ROTTERDAM

Data mining: preliminary steps

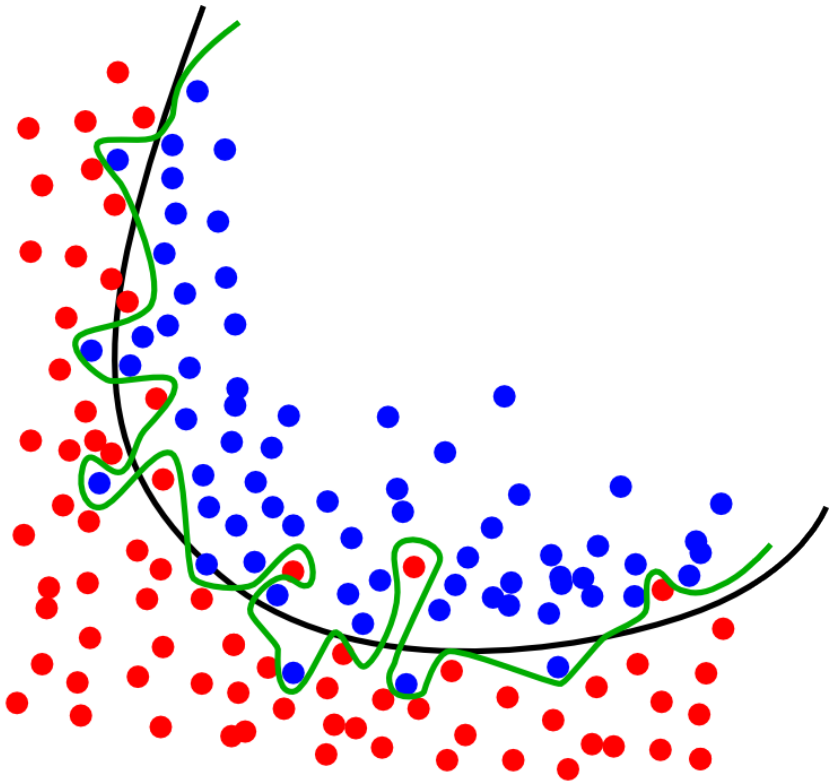
- Data organization
  - Variables in columns, observations in rows
  - In supervised learning, one variable as the response
- Sampling from a database
  - In case of rare events (e.g. customer purchasing a product in response to a mailing), oversample the rare events with or without replacement
- Preprocessing and cleaning the data
- Partitioning the data

Preprocessing and cleaning the data

- Classify variables as continuous, integer or nominal
  - Possibly convert numerical variables to nominal (most often response, e.g. credit score above a certain level $\rightarrow$ grant credit)
  - Possibly convert polynomial variables (student, employed, retired) to binomial (student=yes/no, employed=yes/no)
  - Last value of polynomial variables is redundant and **should not** be used when mapping to binomial

Preprocessing and cleaning the data

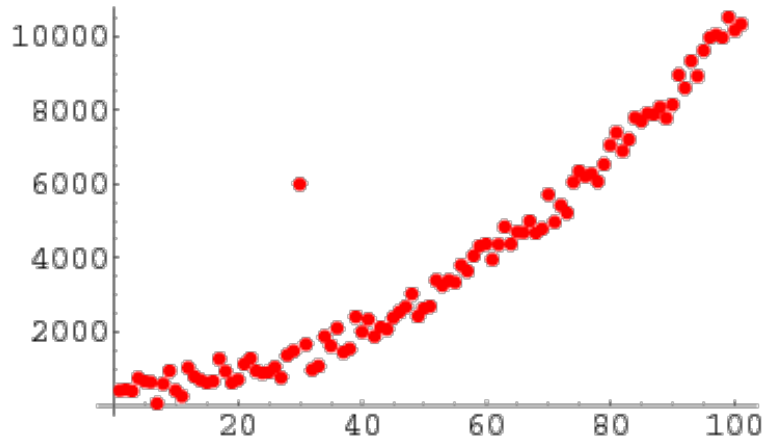
- Select variables (and apply dimension reduction techniques)
  - More variables = greater risk of overfitting



- How many variables and how much data?
- 6 $\times$ nr_outcome_classes $\times$ nr_variables

Preprocessing and cleaning the data

- Detect, inspect, and possibly remove **outliers**
  - Outliers can result from an input error or be part of the data
  - Manual 1-dim outlier detection through sorting in e.g. spreadsheet software
  - Manual 2-dim outlier detection through scatterplots
  - Automatic outlier detection through clustering

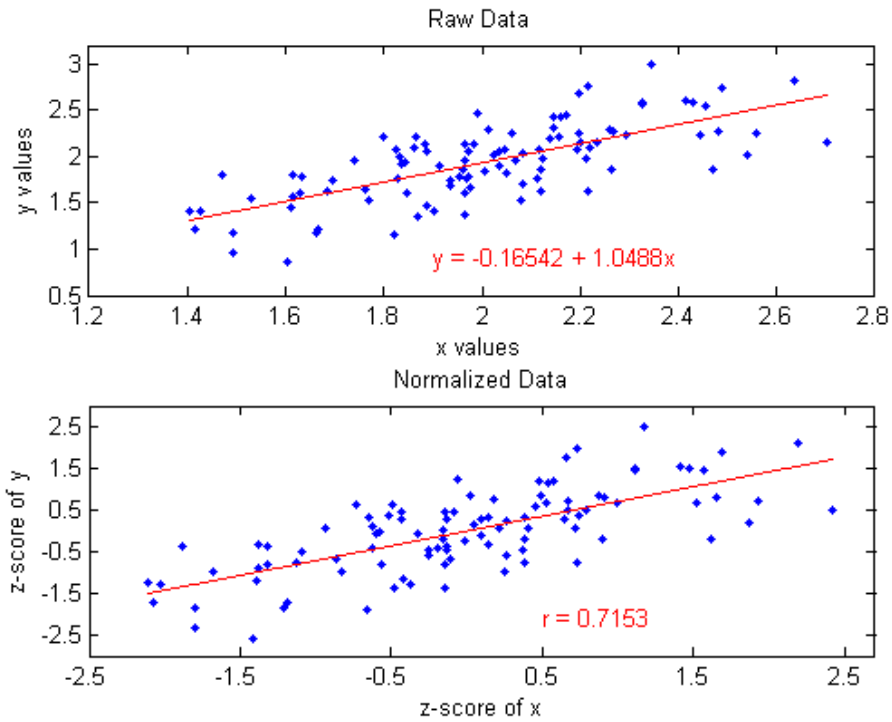


Preprocessing and cleaning the data

- Missing values
  - If the number of missing values is small, the records can be omitted
  - With a large number of variables even small amount of missing values causes a large amount of records to be omitted (e.g. 30 variables, 5% values missing $\rightarrow$ amount of data retained $= 0.95^{30} = 21.5\%$).
  - Input a value, e.g. mean (loses variance which is not a problem as we use a separate test dataset)

Preprocessing and cleaning the data

- Normalize data
  - Some algorithms require normalized data
  - Subtract mean and divide by the standard deviation $\rightarrow$ z-score, "number of standard deviations away from the mean"



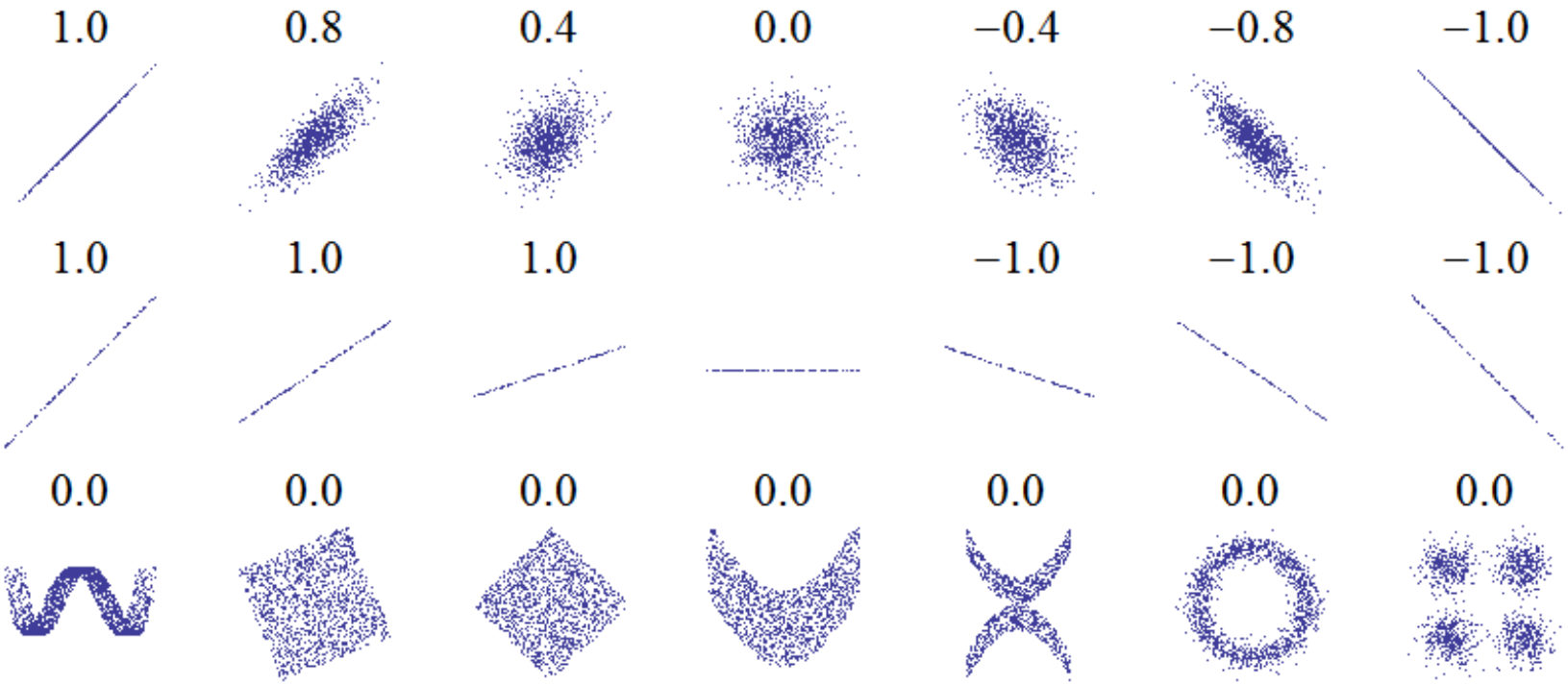
Dimensionality reduction

- Prerequisite for dimensionality reduction is understanding the data, using e.g. data summaries (min, max, avg, mean, median, stdev) and visualization
- Domain knowledge should always be applied first to remove predictors known to be unapplicable (e.g. height for predicting client income)
- Correlation analysis, principal component analysis, and binning

Correlation analysis

- With many variables there is usually overlap in the covered information.
- A simple technique for finding redundancies is to look at the **correlation coeffients** in a **correlation matrix**.
- Pairs that have a very strong positive or negative correlation contain a lot of overlap and are subject to removal

Correlation coefficients



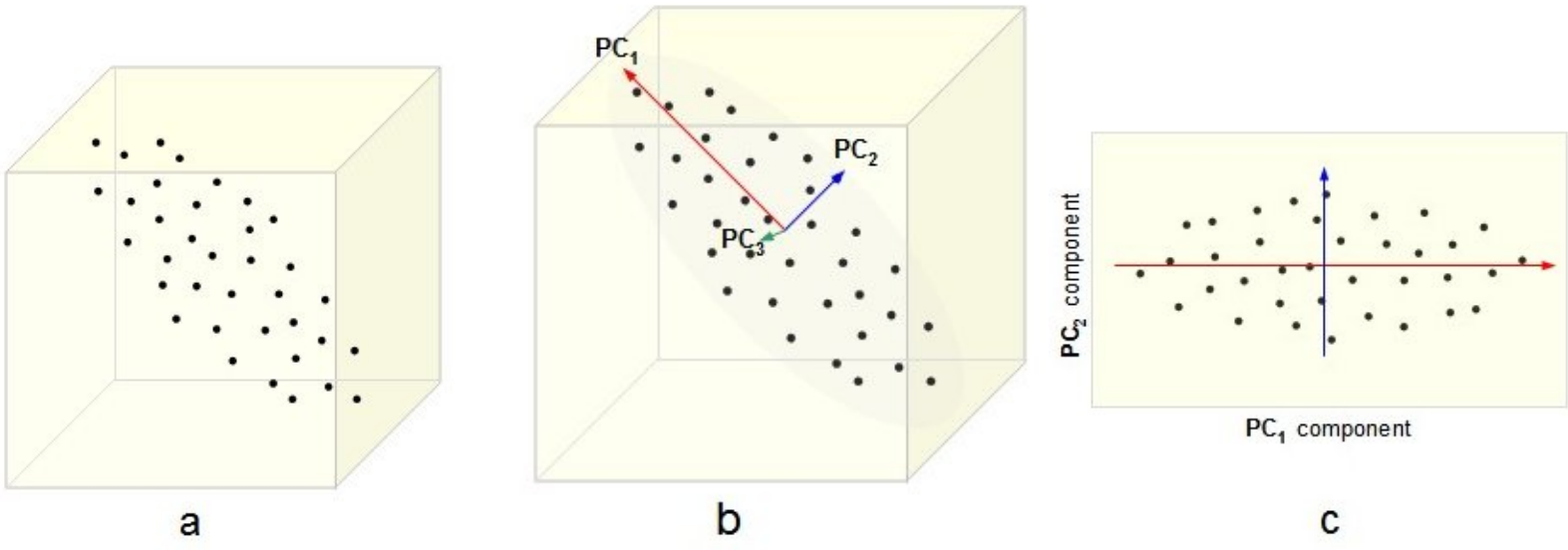
Correlation matrix

Correlations	TIMELR	MEDTOR	AVGDON	LSTDON	ANNDON
TIMELR	1.00				
MEDTOR	-0.11	1.00			
AVGDON	-0.36	0.03	1.00		
LSTDON	-0.04	0.09	0.69	1.00	
ANNDON	-0.28	0.01	0.87	0.63	1.00

Principal Component Analysis (PCA)

- Allows to reduces the number of predictors by finding the weighted linear combinations of predictors that retain most of the variance in the data set
- These are called **principal components**
- PCA works only with continuous variables

PCA example



PCs = weighted averages of original variables after subtracting their means

Example principal components

	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
TIMELR	0.99	0.09	0.06	-0.01	0.00
MEDTOR	-0.19	0.98	-0.01	-0.01	0.00
AVGDON	-0.37	-0.03	0.84	-0.13	0.36
LSTDON	-0.11	0.08	0.79	0.60	-0.02
ANNDON	-0.37	-0.05	0.89	-0.23	-0.07
Percent of Trace:	0.60	0.27	0.11	0.02	0.00

Example principal components

	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
TIMELR	0.99	0.09	0.06	-0.01	0.00
MEDTOR	-0.19	0.98	-0.01	-0.01	0.00
AVGDON	-0.37	-0.03	0.84	-0.13	0.36
LSTDON	-0.11	0.08	0.79	0.60	-0.02
ANNDON	-0.37	-0.05	0.89	-0.23	-0.07
Percent of Trace:	0.60	0.27	0.11	0.02	0.00

Fail, because data not normalized

PCA: normalization

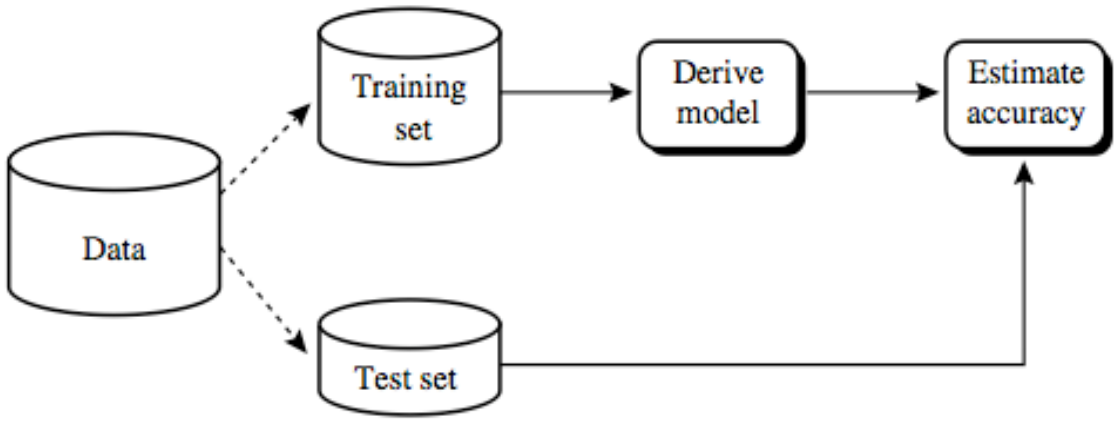
- If variables have different scales [min, max], these get reflected in principal components (e.g. MEDTOR [0, 209] and ANNDON [0.19, 759.80])
- If the scales don't reflect importance of the indicator by being commensurable, e.q. sales of jet fuel, sales of heating oil, you should normalize before applying PCA
- Normalization to unit deviation is achieved by dividing each variable by its standard deviation (z-score)

Use of principal components

- Reduce number of variables, use the PCs as predictors in the model. For test set, apply weights from training set to variables to obtain validation "PCs"
- Produce uncorrelated variables (correlation coefficient $= 0$)
- Describe data

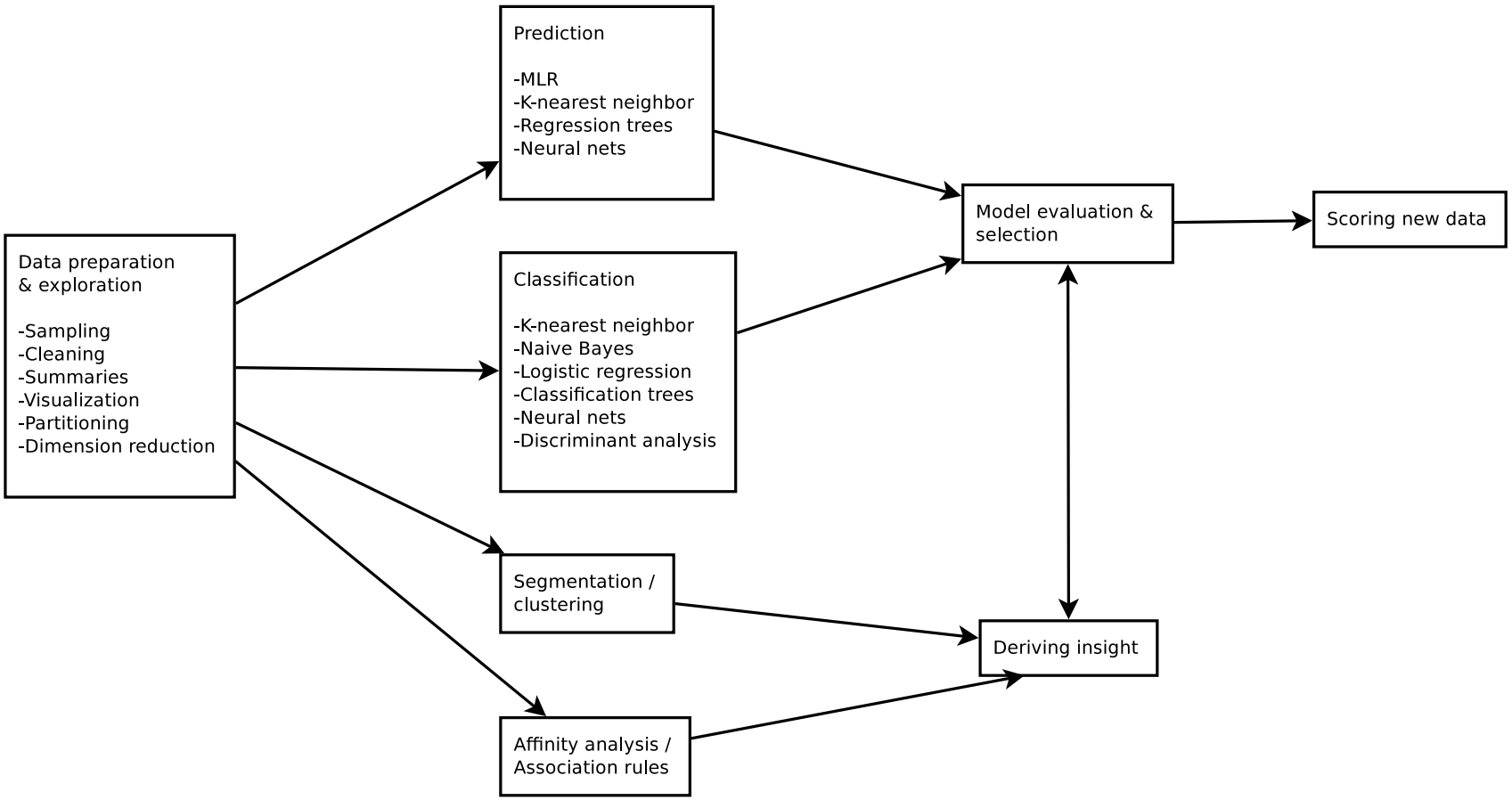
Partition data (if applying supervised learning)

- The derived model can contain bias due to training data matching the model by chance
- The model should always be evaluated/tuned with a separate test set



- Sometimes also a third partition, validation set, is used

Data mining process



Next week

- Load training set data into rapid miner
- Perform correlation analysis, decide on which variables to keep
- Make sure you can train your model in rapidminer
- Start reading about your method