# **LIACS Data Mining course**

#### an introduction



## **Course Textbook**

# Data Mining

# Practical Machine Learning Tools and Techniques

third edition, Morgan Kaufmann, ISBN 978-0-12-374856-0

by Ian Witten and Eibe Frank





(EUR 45,95 at Amazon.de)



## **Course Information**

Course website:

http://datamining.liacs.nl/DaMi/

(will be updated periodically)

- Practical exercises
- New style of exam
  - fewer definitions, more understanding and applying
  - old exams ( $\leq$  2009) should not be used
  - exam preparation important



# **Course Outline**

9-Sep 16-Sep	Knobbe Vanschoren	today
23-Sep	Knobbe	
30-Sep	Vanschoren	
07-Oct	Knobbe	+ practical exercise
14-Oct	Knobbe	
21-Oct	Knobbe	+ practical exercise
28-Oct	Knobbe	
04-Nov	Knobbe	
11-Nov exercise	Takes	guest lecture + practical
18-Nov	Knobbe	
25-Nov	Konijn, Achme	ea guest lecture
02-Dec	Knobbe	+ practical exercise
09-Dec	Knobbe	exam preparation!



# **Introduction Data Mining**

an overview and some examples



# **Data Mining definitions**

#### Data Mining:

the concept of extracting *previously unknown* and *potentially useful, interesting* knowledge from large sets of data.

secondary statistics: analyzing data that wasn't originally collected for analysis.



# Data Mining, the big idea

- Organizations collect large amounts of data
- Often for administrative purposes
- Large body of experience
- Learning from experience

#### Goals

- Prediction
- Optimization
- Forecasting
  - Diagnostics



#### 2 Streams



#### 2 Streams

Mining for insight

- Understanding a domain
- Finding regularities between variables
- Interpretable models
- Examples: Medicine, production, maintenance



## 2 Streams

Mining for insight

- Understanding a domain
- Finding regularities between variables
- Interpretable models
- Examples: Medicine, production, maintenance

#### Black-box' Mining

- Don't care how you do it, just do it well
- Optimization
- Examples: Marketing, forecasting (financial, weather)



Optimize the response to a mailing, by targeting only those that are likely to respond:

more response

fewer letters



- more response
- fewer letters





- more response
- fewer letters





- more response
- fewer letters





- more response
- fewer letters





- more response
- fewer letters





- more response
- fewer letters



# example: Bioinformatics

- Find genes involved in disease (Parkinson's, Celiac, Neuroblastoma)
- Measurements from patients (1) and controls (0)
- Gene expression: measurements of 20k genes
  dataset 20,001 x 100
- Challenges
  - many variables
  - few examples (patients), testing is expensive
  - interactions between genes



# **Data Mining paradigms**

#### Classification

- binary class variable
- predict class of future cases
- most popular paradigm
- Clustering
  - divide dataset into groups of similar cases
- Regression
  - numeric target variable
- Association
  - find dependencies between variables
  - basket analysis, ...

























Age	Gender	House	Price	
	Mortgag			
21	F	Rent	-	No
30	Μ	Rent	-	Yes
40	Μ	Rent	-	No
32	F	Buy	300K	No
30	F	Rent	-	Yes
55	Μ	Buy	260K	No
25	F	Buy	180K	Yes



Age	Gender	House	Price	
	Mortgag	e?		
21	F	Rent	-	No
30	Μ	Rent	-	Yes
40	Μ	Rent	-	No
32	F	Buy	300K	No
30	F	Rent	-	Yes
55	Μ	Buy	260K	No
25	F	Buy	180K	Yes
	Ť	-	T	



Age	Gender	House	Price		
	Mortgage?				
21	F	Rent	-	No	
30	Μ	Rent	-	Yes	
40	Μ	Rent	-	No	
32	F	Buy	300K	No	
30	F	Rent	-	Yes	
55	Μ	Buy	260K	No	
25	F	Buy	18 <b>0</b> K	Yes	
	<b>1</b>	-	<b>1</b>		
	-		-	📕 Rent	
		<			
				P Duy	
-1222 ·					
44				Other	
Universiteit Leider				Other	











# Applying a classifier (decision tree)

New customer: (House = Rent, Age = 32, ...)





# Applying a classifier (decision tree)

New customer: (House = Rent, Age = 32, ...)





# Applying a classifier (decision tree)

New customer: (House = Rent, Age = 32, ...) prediction = Yes





dataset with two variables + 1 class (+/-)
 graphical interpretation of decision tree





dataset with two variables + 1 class (+/-)
 graphical interpretation of decision tree



dataset with two variables + 1 class (+/-)
 graphical interpretation of decision tree



dataset with two variables + 1 class (+/-)
 other classifiers





dataset with two variables + 1 class (+/-)
 other classifiers





dataset with two variables + 1 class (+/-)
 other classifiers





# **Applications of DM**

- Marketing
  - outgoing
  - incoming
- Bioinformatics & Medicine
- Fraud detection
- Risk management
- Insurance
- Enterprise resource planning



# **Rhinoplastic surgery**



histogram over VAE improvement 30 pre E1c > 3 all patients 23 nr. patients 15 8 0 2 2.5 3.5 0 0.5 1.5 3 4 4.5 5 1 **VAE** improvement

#### 'beinvloedt deze bezorgdheid uw dagelijkse leven'

#### InfraWatch: monitoring of infrastructure





Continuous monitoring of a large bridge 'Hollandse Brug'

- 145 sensors
- time-dependent, at frequencies up to 100Hz
- multi-modal (sensor, video, differen freq.)
- managing large data quantities, >1 Gb per day



# InfraWatch: monitoring of infrastructure



- 34 geo-phones (vibration sensors)
- 44 embedded strain-gauges, 47 gauges outside
- 20 thermometers

weather station

video camera



niversiteit Leiden

#### sensor mining







### Real-world application: Maintenance planning at KLM

- Routine checks of aircrafts
- Maintenance requires up to 10k different parts
- Ordering parts incurs delay (costs)...
- ... but so does stocking
- In theory 10k individual predictions
- Input
  - maintenance history
  - flight history, Sahara/North Pole
- Only few parts predictable







# **Cashflow Online**

- Online personal finance overview
- All bank transactions are loaded into the application
- transactions are classified into different categories
- Data Mining predicts category



ABN-AMRO

Fortis |

Rabobank

ING



# **67 Categories**

Gas Water Licht Onderhoud huis en tuin Telefoon + Internet + TVContributie (sport-)verenigingen Levensverzekering / Lijfrente Rente ontvangen Boodschappen Hypotheekrente Naar spaarrekening Geldopname/chipknip Verzekeringen overig Loterijen Cadeau's Interne boeking Vakantie & Recreatie Uitgaan, hobby's en sport Creditcard Ziektekostenverzekering Brandstof Woonhuis / Opstalverzekering



#### **Fragmented results:**

#### Boodschappen (groceries)

25000 patterns found for td.RUBRIEK - 'Boodschappen', Novelty, all data						
Nr.	Depth	Tables	Coverage	Accuracy	Novelty	Condition list
1	2	td	70319	0,267893	0,054078	BIJ = '0' AND BETALING = '0'
2	2	td	70319	0,267893	0,054078	BETALING = '0' AND BIJ = '0'
3	2	td	31836	0,356986	0,052846	BANK = '0' AND RELATIEREKENING = "
4	2	td	31836	0,356986	0,052846	RELATIEREKENING = "AND BANK = '0'
5	2	td	32366	0,352623	0,052314	ABO = '0' AND RELATIEREKENING = "
6	2	td	32366	0,352623	0,052314	RELATIEREKENING = "AND ABO = '0'
7	2	td	71846	0,263773	0,052291	BIJ = '0' AND BETALINGSKENM = '0'
8	2	td	71846	0,263773	0,052291	BETALINGSKENM = '0' AND BIJ = '0'
9	2	td	32380	0,352471	0,052287	RABO_BANK = '0' AND RELATIEREKENING = "
10	2	td	32380	0,352471	0,052287	RELATIEREKENING = "AND RABO_BANK = '0'
11	2	td	32671	0,350647	0,052162	RENTE = '0' AND RELATIEREKENING = "

#### Contributie

25000 patterns found for td.RUBRIEK - 'Contributie (sport-)verenigingen', Novelty, all data						
Nr.	Depth	Tables	Coverage	Accuracy	Novelty	Condition list
1	2	td	1068	0,833333	0,008742	CONTRIBUTIE = '1' AND VERZ = '0'
2	2	td	1068	0,833333	0,008742	CONTRIBUTIE = '1' AND BETALINGSCENTRUM = '0'
3	2	td	1068	0,833333	0,008742	CONTRIBUTIE = '1' AND ASR = '0'
4	2	td	1073	0,82945	0,008742	CONTRIBUTIE = '1' AND ABONNEMENT = '0'
5	2	td	1073	0,82945	0,008742	CONTRIBUTIE = '1' AND INZAKE = '0'
6	2	td	1074	0,828678	0,008742	CONTRIBUTIE = '1' AND NUON = '0'
7	2	td	1075	0,827907	0,008741	CONTRIBUTIE = '1' AND VOF = '0'
8	2	td	1076	0,827138	0,008741	BETREFT = '0' AND CONTRIBUTIE = '1'
9	2	td	1076	0,827138	0,008741	BTW = '0' AND CONTRIBUTIE = '1'
10	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND DONATIE = '0'
11	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND BETREFT = '0'
12	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND BOER = '0'
13	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND KRUIS = '0'
14	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND BTW = '0'
15	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND HENGELO = '0'
16	2	td	1076	0,827138	0,008741	CONTRIBUTIE = '1' AND COOP = '0'
17	2	td	1077	0,82637	0,008741	VOORSCHOT = '0' AND CONTRIBUTIE = '1'
18	2	td	1077	0,82637	0,008741	KRUIDVAT = '0' AND CONTRIBUTIE = '1'
19	2	td	1077	0,82637	0,008741	SERVICES = 10' AND CONTRIBUTIE = "1"
20	2	td	1077	0,82637	0,008741	ALDI = '0' AND CONTRIBUTIE = '1'

#### **Decision Tree over all categories**





# **Data Mining at LIACS**

#### Applications

- bioinformatics (LUMC)
- rhinoplastic surgery (NKI)
- Hollandse Brug (Strukton, TU Delft, RWS)
- fraud detection at Achmea health insurance
- ChartEx, medieval documents (English, Latin)

#### Complex data

- graphical data (molecules)
- relational data (criminal careers)
- stream data (sensor-data, click-streams)

