Data Mining and Analysis Analiza i eksploracja danych



Lecturer: JERZY STEFANOWSKI Institute of Computing Science Poznań University of Technology

Software Engineering – Master Course Computer Science, PUT, revised 2015 / 2016

Background literature [Polish translations]

Translations:

- Larose D., Odkrywanie wiedzy z danych. Wprowadzanie do eskploracji danych, PWN, 2006.
- Larose D., Metody i modele eksploracji danych, PWN 2008.
- Hand D., Mannila H., Smyth P. Eksploracja danych, WNT, 2005 (Principles of Data Mining, MIT Press, 2001).
- Polskie książki Polish language books
 - Tadeusz Morzy, Eksploracja danych. Metody i algorytmy. PWN 2013!!!!!
 - Koronacki J., Ćwik J., Statystyczne systemy uczące się, WNT 2005 (kolejne wydanie w drodze).
 - Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. PP, 2003.



Background literature (English)

- Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001 (1 ed.), there is 2d
 - Hand D., Mannila H., Smyth P. Principles of Data Mining, MIT Press, 2001.
 - Kononenko I., Kukar M., Machine Learning and Data Mining: Introduction to Priniciples and Algorithms. Horwood Pub, 2007.
 - Maimon O., Rokach L., The data mining and knowledge discovery Handbook, Springer 2005.
 - Witten I., Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.



Lecture 1 a.

Data Mining: Introduction

Motivations - data explosion problem

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.
- More data is generated:
 - Bank, telecom, other business transactions ...
 - Scientific data: astronomy, biology, etc
 - Web, text, and e-commerce
- Very little data will ever be looked at by a human!
- We are drowning in data, but starving for knowledge!

Data Flood and Answers

- Data mining?
 - Extraction of useful information patterns from data
 - More than typical data analysis, machine learning or classical decision support!
- Knowledge Discovery is NEEDED to make sense and use of data.

Data mining: what is it?



• Data mining is

- Extraction of useful patterns from data sources, e.g., databases, texts, web, images.
- Patterns (knowledge representation) must be:
 - Valid, novel, potentially useful, understandable to the users.



Data Mining: On What Kind of Data?

- Attribute-value tables (standard form / data table)
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, ...)
- Sequence data bases
- Other more complex data repositories
 - Object-oriented and object-relational databases
 - Spatial databases
 - Time-series data and temporal data
 - Data streams
 - Text databases and multimedia databases
 - WWW resources





Flat files

- Actually the most common data source for data mining, especially at the research level.
- Simple data files in text or binary format with a structure known by the data mining algorithm to be applied.
- The data in these files can be transactions, time-series data, scientific measurements, etc.
- Big data efficiency of access and management.

Instance	f 1	 f k	Y
x1	V 1,1	 V 1,k	V1,k+1
x i	V i,1	 V i,k	Vi,k+1
xn	V n,1	 V n,k	Vn,k+1

Types of attributes

- The most common distinction comes from measurement scale and statistics:
 - Nominal (also binary)
 - Ordinal
 - Interval-scaled
 - Ratio-scaled.
- Other names:
 - Categorical vs. numeric/continuous ones.
- Other types:
 - Criteria (preference-ordered), hierarchical, …

Typical approach to the classification task.

Age	Car Type	Risk
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Numeric prediction – regression function

• Example: 209 different computer configurations

	Cycle time (ns)	Main r ۴)	nemory (b)	Cache (Kb)	Cha	nnels	Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Linear regression function

PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX

Transforming text documents into a standard form

Transformation into Vector Representation



Data Preparation for Knowledge Discovery



A crucial issue: The majority of time / effort is put there.

Data Understanding: Quantity

- Number of instances (records)
 - Rule of thumb: 5,000 or more desired
 - if less, results are less reliable; use special methods (boostrap sampling, ...)
- Number of attributes (fields)
 - Rule of thumb: for each field (attribute) find 10 or more instances
 - If more fields, use feature reduction and selection
- Number of targets
 - Rule of thumb: >100 for each class
 - if very unbalanced, use stratified sampling or specific preprocessing (SMOTE, NCR, etc.)

Why Data Preprocessing?

- Data in the real world is "dirty" ...
 - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., occupation=""
 - noisy: containing errors or outliers
 - e.g., Salary="-10"
 - inconsistent: containing discrepancies (disagreements) in codes or names
 - e.g., Age="42" Birthday="03/07/1997"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
 - e.g., discrepancy between duplicate records

Basic forms of data preprocessing



From J.Han's book

Basis problems in "Data Cleaning"

- Data "acquisition" / integration and metadata
- Unified formats and other transformations
- Erroneous values
- Missing values
- Data validation and statistics

Erroneous / Incorrect values

• What suspicious can you see in this table?

Ì	Dane: TPI	Odatacleani	ng.STA 7v	🗂 Statystyki opisowe				
TEK VAI	1	2	3	4	5	6	7	AGE
	ID_CUST	CODEPOST	SEX	INCOME	AGE	MARTIALS	TRANS_SU	
1	1001	10048	М	75 000	С	М	5000,00	Szczegółowe statystyki opisowe
2	1002	74002	F	40 000	40	V	4000,00	Opcje
3	1003	90210		50 000	54	S	5400,00	🗖 Usuwanie BD, przypadkami
4	1004	J2S7K7	F	-40 500	34	S	4500,00	Wyświetł długie nazwy zmiennych
5	1005	6269	М	54 000	37	М	6500,00	
6	1006	45210	F	?	23	D	4500,00	Duczenia zmiększonej precyzli
7	1007	60210	М	99 450	0	М	3000,00	Rozkład
8	1008	65430	M	10000000	56	S	1000,00	🕅 Tabele liczebności 🖓 Histogramy
9	1009	60211	М	3000	43	S	2400,00	
10	1009	60211	М	3000	43	S	2400,00	Normalne częstości oczekiwane
		· · · · · · · · · · · · · · · · · · ·						🗌 📃 Testy normalności K-S i Lillieforsa
								🗌 🗖 Test W Shapiro-Wilka

Incorrect values

- Reason: data has not been collected for mining it
- Result: errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes ⇒ values need to be checked for consistency
- Typographical and measurement errors in numeric attributes ⇒ outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)
- Other problems: duplicates, …

Tools?

Outliers – graphical identification

• Use simple statistics and araph tools - Statistica



Redundant Data

- Redundant data occur often when integration of multiple databases
 - The same attribute may have different names in different databases
 - One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlation analysis
- Large number of redundant data may slow-down or confuse knowledge discovery process.

Looking for correlated columns

🖾 STATISTICA -	Workbook2* - [(Correlations	(EnginePerforman	ce.sta)]										7 🗙
<u>File E</u> dit <u>V</u> iew	<u>File Edit V</u> iew Insert F <u>o</u> rmat <u>S</u> tatistics <u>G</u> raphs <u>T</u> ools <u>D</u> ata Workbook <u>W</u> indow <u>H</u> elp													
0 🛩 🖬 🔁	🔿 🖪 👗 🖻	🛍 🝼 🗠	🖙 🎒 Add to Wor	kbook 👻	Add to R	eport 👻 🐧	🛛 🧼 💦 🗸							
Arial	▼ 10	• B I	⊻ ≣ ≣ ⊒ 🗗	<u>A</u> - <	> - 💽 -	• 🇞 📗	◆.0 .00 ⊉	- • @ @	🔒 👌 ×=?	%1=? %2=? Vars ▼ Ca	ises 🕶 🔯 🖸	- 🖾 -	-	
Deter Franke	Destaurant	. (70	20-1			, , , , , , , , , , , , , , , , , , , ,				-				
🏢 Data: Engine	Performance.st	а (790 ру 1.	28C)											
	1	2	2	4		E	c	7	0	0	10		<u></u>	
	Sorial Number	Z Efficiency	J Eugl Economy(%)	4 Dowor(%) In	5 mut01	b Innut02	/ InnutΩ	0 3 Innutí	9 M Input05	UI Inputi	าด เก	nut	
1	#05/157	102 384		rower(aa :	70) III 317 100	1 186545	16 6255147	11 9297	997 15 <u>450</u>	1075 16 71993	19 17 / 75	4064 20	753	
2	#25458	81 405	89,798	110	392 98	4136317	16 3445083	13,5326	772 14 0011	3087 15 63472	214 17.05	4004 20, N197 20'	303	
3	#25459	94 070	92 072	87 9	317 98	7403916	16 5964348	12,0007	502 15 507	7475 15 78571	13 18 617	5749 20.	527	
4	#25460	108,855	89,369	90,9	345 99	5529412	16,7615965	12,0610	633 14,2580	0726 13,86958	301 17,885	1961 19	0.81	
5	#25461	107,903	89,453	95	312 98	8236109	16 6525248	12 2789	147 14 650	1313 20.6343	384 17 121	8605 21	11	
6	#25462	86,475	94,063	30	Workb	ook2* - C	orrelations	(Fngine	Performance	.sta)				
7	#25463	105,583	94,868		Workboo			(5		,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,				
8	#25464	109,303	95,652		Gran	n∼ hs of Block			Correlations	(EnginePerfor	mance.sta)			
9	#25465	103,633	91,181		Grop	Histogram ((Marked corr	elations are si	gnificant at	p < ,0500	0	
10	#25466	95,300	93,490		🔄 Basic	: Statistics/			N=128 (Cas	ewise deletion	of missing	data)		
11	#25467	102,334	90,320		ē-🔂 (Correlations			Efficiency	Fuel	Power(%)	Input01	Input02	InputC
12	#25468	94,456	118,944			Correlat	Variable		1.00	Economy(%)	0.40	0.40	0.40	
13	#25469	109,349	107,956	1			Efficiency	(0()	1,00	-0,09	0,12	0,12	0,19	U,
14	#25470	105,943	89,392				Fuel Econo	omy(%)	-0,09	1,00	0,53	0,67	0,50	U,
15	#25471	101,390	102,309				Power(%)		0,12	0,53	1,00	0,26	0,14	U,
16	#25472	105,911	107,008	1			Input01		0,12	0,07	0,26	1,00	1.00	-0,
17	#25473	78,027	91,527				Input02		0,19	0,50	0,14	0,83	1,00	-0,
18	#25474	107,266	89,611				Input03		0,06	0,10	0,12	-0,01	-0,05	i
19	#25475	99,571	101,998	1			Input04		-0,07	-0,08	0,00	-0,20	-0,23	-0,
20	#25476	107,466	102,613	1			Linput06		-0,00	-0,00	0,05	-0,10	-0,04	U, 0
21	#25477	109,327	95,364	1			mpuloo	_	0,15	0,11	0,17	0,14	0,16	U,
22	#25478	104,091	91,369											
23	#25479	95,655	90,542											
24	#25480	107,033	96,745											
25	#25481 405400	108,802	107,768											
26	#25482 #05482	98,975	117,309											
27	#25483 405404	104,152	100,064											
28	1#25464	67,792	116,900											
			1											

Scatterplot matrix

M STATIS	STICA - Workboo)k3* - [Correlat	ions (Irisdat.	sta 5v*150c))]		_ 6
<u> </u>	t <u>V</u> iew Insert F	ormat Statistics	<u>G</u> raphs <u>T</u> ools	Work <u>b</u> ook <u>W</u>	(indow <u>H</u> elp		
0 🖻 🛛	- 1	አ 🖻 🛍 🚿	004	Add to Workbo	ook 👻 Add to Report 👻	र्देख 🛷 💦 🗸	
S Norm	al Graph [modi 💌	3.8.2	y G 🖂 🛱	‡ ⊕ Q @	1 🙉 F 🛍 🖆 🖒	3 ° × ° × ° a a a == ₩ • • • • • • • •	
Data:	EngineDerform	ance eta (70v l	w 120a)				
III Data;	🗰 Data: Irisdat	t.sta (5v bv 15	0c)				
		Fisher (1936) ir	is data: length	. & width of se	enals and netals 3 ty	vnes of Iris	
		1 ISHEF (1000) II	2	3	4 5		1
		SEPALLEN SE	PALWID PET	TALLEN PETA	ALWID IRISTYPE		5
	125	5,1	3,5	1,4	0,3 SETOSA		3
	126	7,2	3,6	6,1	2,5 VIRGINIC		7
	127	6,5	3,2	5,1	2,C 🏢 Workbook	k3* - Correlations (Irisdat.sta 5v*150c)	
	128	6,1	2,9	4,7	1.2 🕞 Workbook3*	*	
	129	5,6	2,9	3,6	1.3 🔄 🔄 Basic Sta	ati: Correlations (Irisdat.sta 5v*150c)	
	130	6,9	3,1	4,9	1. 🗧 🗍 🔂 Corre	reli serallen :	3
	131	6,4	2,7	5,3			
	132	6,8	3,0	5,5	2,1		
	133	5,5	2,5	4,0			
	134	4,0		1,0	0.4		
<u> </u>	135	4,0		1,4	0.1		•••
	130	4,3	2,5	- ,J 5 0	20	I HARRING I BON I HARRING I HARRING	10
	138	57	2,5	17	03		
	139	51	3.8	15	0.2		
	140	55	2.3	4 0	1.3	PETALEN	
	141	6.6	3.0	4.4	1.4		
	142	6.8	2.8	4.8	1.4		
	143	5,4	3,4	1,7	0,2		
	144	5,1	3,7	1,5	0,4	• 8.8.** PETAUNI	,
	145	5,2	3,5	1,5	0,2		
	146	5,8	2,8	5,1	2,4		Ser
	147	6,7	3,0	5,0	1,7		
	148	6,3	3,3	6,0	2,5		

Data Cleaning: Missing Values

- Missing data can appear in several forms:
 - <empty field> ? "0" "." "999" "NA" …
- Missing data may be due to
 - equipment malfunction
 - inconsistent with other recorded data and thus deleted
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
 - not register history or changes of the data
- Missing data may need to be inferred inputation!

Missing and other absent values of attributes

- Value may be missing because it is unrecorded or because it is inapplicable
- In medical data, value for Pregnant? attribute for Jane or Anna is missing, while for Joe should be considered Not applicable
- Don't care values

Hospital Check-in Database

Name	Age	Sex	Pregnant	
Mary	25	F	N	
Jane	27	F	?	
Joe	30	М	-	
Anna	2	F	?	

Handle Missing Values

- Ignore / delete the instance: (not effective when the percentage of missing values per attribute varies considerably).
- Fill in the missing value manually: expert based + infeasible?
- Fill in a more advanced way :
 - a global constant : e.g., "unknown", a new class? don' t use it!
 - the attribute mean or the most common value.
 - the attribute mean for all examples belonging to the same class.
 - the most probable value: inference-based such as Bayesian formula or decision tree // prediction - regression model
 - result of global closest fit (distance base approaches)
 - Use a prediction technique

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a smaller, specified range
 - min-max normalization
 - z-score normalization
- Attribute/feature construction
 - New attributes constructed from the given ones

Data Transformation: Normalization

min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

z-score normalization

$$v' = \frac{v - mean_A}{stand \ dev_A}$$

normalization by decimal scaling

 $v' = \frac{v}{10^{j}}$ Where *j* is the smallest integer such that Max(|v'|)<1

Discretization

- Some methods require discrete values, e.g. most versions of Naïve Bayes, CHAID, Associations,
- Discretization → transformation of numerical values into codes / values of ordered subintervals defined over the domain of an attribute.
- Discretization is very useful for generating a summary of data
- Many approaches have been proposed:
 - Supervised vs. unsupervised,
 - Global vs. local (attribute point of view),
 - Dynamic vs. static choice of parameters

Discretization: Equal-Width (Length)



[64,67) [67,70) [70,73) [73,76) [76,79) [79,82) [82,85]

Equal Width, bins Low <= value < High

Discretization: Equal-Width may produce clumping



Salary in a corporation

Discretization: Equal-Frequency

Temperature values: 64 65 68 69 70 71 72 72 75 75 80 81 83 85



Equal Height = 4, except for the last bin

Supervised (class) discretization

Use information about attribute value distribution + class assignment.



values of the attribute

• Minimal entropy based approaches; Chi-Merge, others

Evaluate purity of information about learning examples with Entropy (similar to decision tree split)

$$e_i = -\sum_{j=1}^k p_{ij} \log_2 p_{ij}$$

However, you also need a conditional entropy (with and attribute splitting)

Entropy-Based Discretization

• For learning examples S; If S discretized into two subintervals S1 i S2 using (cut point) T, conditional entropy is defined as: $|S_1| = |S_2| = |S_2|$

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- Scan all possible cut points
- Choose the one mininizing the entropy.
- Continue until a stopping conditions such as $Ent(S) - E(T,S) > \delta$
- MDL principle could be also exploited

A Toy example

- Starting entropy $Ent(S) = -\frac{3}{6} \cdot \lg \frac{3}{6} \frac{3}{6} \cdot \lg \frac{3}{6} = 1$
- Attribute /Q and a cut point T=107 (inter. Left < T)

105	107	107	109	113	115
yes	no	no	no	yes	yes

$$Ent(S \mid T) = \frac{1}{6}(-1 \cdot \lg 1) + \frac{5}{6}(-\frac{3}{5} \cdot \lg \frac{3}{5} - \frac{2}{5}\lg \frac{2}{5}) = 0.811$$

- Yet another cut point T=113 Ent(S|T) = 0.541 the better choice.
- Fayyad and Irani theoretical advice limit tested cut points

Outliers and Errors

- Outliers are values thought to be out of range.
- Approaches:
 - do nothing
 - enforce upper and lower bounds
 - let binning handle the problem

Examine Data Statistics

**************** Field 9: MILES_ACCUMULATED

Total entries = 865636 (23809 different values). Contains non-numeric values. Missing data indicated by "" (and possibly others).

Numeric items = 165161, high = 418187.000, low = -95050.000 mean = 4194.557, std = 10505.109, skew = 7.000

Most frequent entries:

Value	Total					
:		700	474	(80.9	00
0:		327	48	(3.8%)
1:		416	(0.	0응)	
2:		337	(0.	0응)	
10:		321	(0.	0응)	
8:		284	(0.	0%)	
5:		269	(0.	0응)	
6:		267	(0.	0응)	
12:		262	(0.	0응)	
7:		246	(0.	0%)	
4:		237	(0.	0%)	

Outliers – quite far from typical distributions



 Box plot of data from the Michelson–Morley experiment displaying four outliers in the middle column, as well as one outlier in the first column.



Experiment No.

Multi-dimensional case



Cluster analysis⁻

Another perspective - regression



- Numeric prediction regression model
- Linear model $y = a_1 x_1 + a_2 x_2 + ... + a_m x_m$

Data Preprocessing: Attribute Selection

First: Remove fields with no or little variability

- Examine the number of distinct field values
 - Rule of thumb: remove a field where almost all values are the same (e.g. null), except possibly in minp % or less of all records.
 - *minp* could be 0.5% or more generally less than 5% of the number of targets of the smallest class
- More sophisticated (statistical or ML) techniques specific for data mining tasks
 - In WEKA see attribute selection

Too many attributes!

"Curse of dimensionality" [Bellman 1961]

- For a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve!
- "the number of samples required per variable increases exponentially with the number of variables"



Toy classification example [D.Mladenic 2005]



- Data set
 - Five Boolean features
 - $C = F_{1 V} F_2$
 - $F_3 = \neg F_2, F_5 = \neg F_4$
 - Optimal subset:

 $\{F_1,\,F_2\}$ or $\{F_1,\,F_3\}$

 optimization in space of all feature subsets 2^F (possibilities)

(tutorial on genomics [Yu 2004])

Real working K-NN with many attributes



Different attribute selection methods

Filtering (single attributes)

- Correlation-based measure.
- Contextual-merit.
- Info-Gain.
 - Gain ratio
- Chi-squared statistic
- Liu Consistency measure
- Subsets and more advanced search
 - Relief method
 - Wrapper model

WEKA – attribute selection tools



Ranking with ...? WEKA

👙 Weka Explorer	
Preprocess Classify Cluster Associate	Select attributes Visualize
Attribute Evaluator	
Choose ChiSquaredAttributeEval	
Search Method	
Choose Ranker -T -1.7976931348623	11 57E308 -N -1
Attribute Selection Mode	Attribute selection output
 Use full training set 	DI:
O Cross-validation Folds 10	Evaluation mode: evaluate on all training data
Seed 1	
(Nom) D1:	
(Noin) D1.	=== Attribute Selection on all input data ===
Start Stop	
-Reput list (vight click for antiona)	Search Method:
	Attribute ranking.
21:37:48 - Ranker + ChiSquaredAttributeEval	Attribute Fueluator (sumervised Class (nominal), 9 Dl.).
	Chi-squared Ranking Filter
	Ranked attributes:
	71.9035 2 A3:
	68.5634 1 A2:
	67.8595 4 A5:
	67.629 8 A9:
	64.2122 7 A8:
	64.0766 3 A4:
	14.0986 6 A7.
	Selected attributes: 2,1,4,8,7,3,5,6 : 8

How could you exploit rankng

000	Weka 3.5.4 - Explorer			
Program Applications Tools Visualiza	ation Windows Help			Find thorshold r
	Explorer			
Preprocess C	lassify Cluster Associate Select at	tributes Visualize		
Attribute Evaluator				Madian an anath alag?
Choose InfoGainAttributeEval				Median or smth else?
Search Method				
Choose Ranker -T - 1.7976931348	623157E308 - N - 1			
Attribute Selection Mode	Attribute selection output			
Use full training set	Information Gain Ranking Fil	lter	ř.	
Cross-validation Folds 10 Seed 1	Ranked attributes: 1.6896 11 rewred-mean 1.6043 13 rewgreen-mean 1.6027 10 intensity-mean 1.6041 19 hue-mean			
(Nom) class	1.5677 17 value-mean 1.5492 12 ravblue-mean 1.3426 2 region-centroid-row			
Start Stop	1.3317 16 exgreen-mean 1.3039 18 saturation-mean		0.111	
Result list (right-click for options)	1.0016 14 exred-mean		1	Segment I-Gain
17:04:47 - Ranker + ReliefFAttributeEval	0.4939 6 vodgo-mean			
17:06:49 - Ranker + InfoGainAttributeEval	0.3786 7 vegde-sd 0.1064 1 region_centroid_col	1.8		
	0.028 5 short-line-density-2 0 4 short-line-density-5			
	0 3 region-pixel-count	1.6 +		
	Selected attributes: 11,13,10,19,17,	1.4		
		1.2		
Status				
OK				
		0.6		
		0.4		
		0.2		
		0.2		
		0+		
		rawed-near-near-near-	learnearnear	row rear near near near near near est est color and the service rear near near near near near near nea
			(eg) e	res and and res

Search in Subset Space



How to move in the space

• An example of search space (John & Kohavi 1997)



Heuristic Subset Search Techniques

- There are 2^d possible sub-features of *d* features
- Several heuristic feature selection methods:
 - Best step-wise feature selection (forward):
 - The best single-feature is picked first
 - Then next best feature condition to the first, ...
 - Step-wise feature elimination (backward):
 - Repeatedly eliminate the worst feature
 - Best combined feature selection and elimination
 - Partly non-deterministic seach (genetic and other techqniques)

Wrapper approach

• Filter vs. Wrapper approach (Kohavi et al. 94, and ...)



- The classifier is used by the evaluation function
- Search algorithms:
 - Forward selection
 - Backward elimination

. . .

Constructing new attribute

- Following A.Berge find new attributes
- In general two approaches for dimensionality reduction
 - Feature selection: choose a subset of the features



 Feature extraction: create a subset of new features by combining existing features

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} \right)$$

Principal Component Analysis (PCA)

- Given N data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) that can be best used to represent data
- Steps
 - Normalize input data: Each attribute falls within the same range
 - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
 - Each input data (vector) is a linear combination of the k principal component vectors
 - The principal components are sorted in order of decreasing "significance" or strength
 - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

Principal Component Analysis



Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
 - Data cleaning and data integration
 - Data reduction and feature selection
 - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research

Knowledge Discovery Process, in practice



Any questions, remarks?

