Machine Learning & Data Mining

J. Savoy
University of Neuchatel

I. H. Witten, E. Frank, M.A. Hall: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

A. Rajaraman, J.D. Ullman: *Mining of Massive Datasets*. Cambridge University Press.

What I'm expecting

- Introduction new CS models
- Find useful *practical* solutions
- Basic ideas / evaluation
- From your part:
  - Precision (definition / equation)
  - Ask "why" (black box, transparent box)
  - 5 ECTS ≠ 4 h. / week
- Evaluation:
  - Written exam (2 h) (English, French *si on le demande avant*)
  - Practical exercises: 1 pt (no repetition!)
  - Final test: 4 pt

Overview of this Course

- Introduction to Machine Learning & Data Mining
- Simple Rules
- Bayes Learning
- Decision Trees
- Evaluation Methodology
- Rules Models
- Association Rules
- Nearest Neighbors (NN, *k*-NN, minHash)
- Advanced Models
- Combining Methods
- Practical Exercises (WEKA)
- *Another course on Statistical Learning Models (R)*

Overview of the Course « Statistical Learning Methods »

- Introduction to Statistical Learning
- Introduction to **R**
- Linear Regression
- Assessing Model Accuracy
- Classification Methods (logistic, LDA, k-NN)
- Advanced Linear Models (ridge, lasso)
- Support Vector Machines (SVM)
- Unsupervised Learning (PCA, clustering)

- Practical exercises (with R)

References

- I. H. Witten, E. Frank, M.A. Hall: *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

- A. Rajaraman, J.D. Ullman: *Mining of Massive Datasets*. Cambridge University Press, 2012

- C.M. Bishop: *Pattern Recognition and Machine Learning*. Springer, 2006.
- T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning*. Springer, New York, 2009.

The Market…

"…The US alone could face a shortage by 2018 of 140,000 to 190,000 people with the deep analytical skills and know-how needed to analyze data and discover new insights. "

McKinsey Global Institute report, 2011

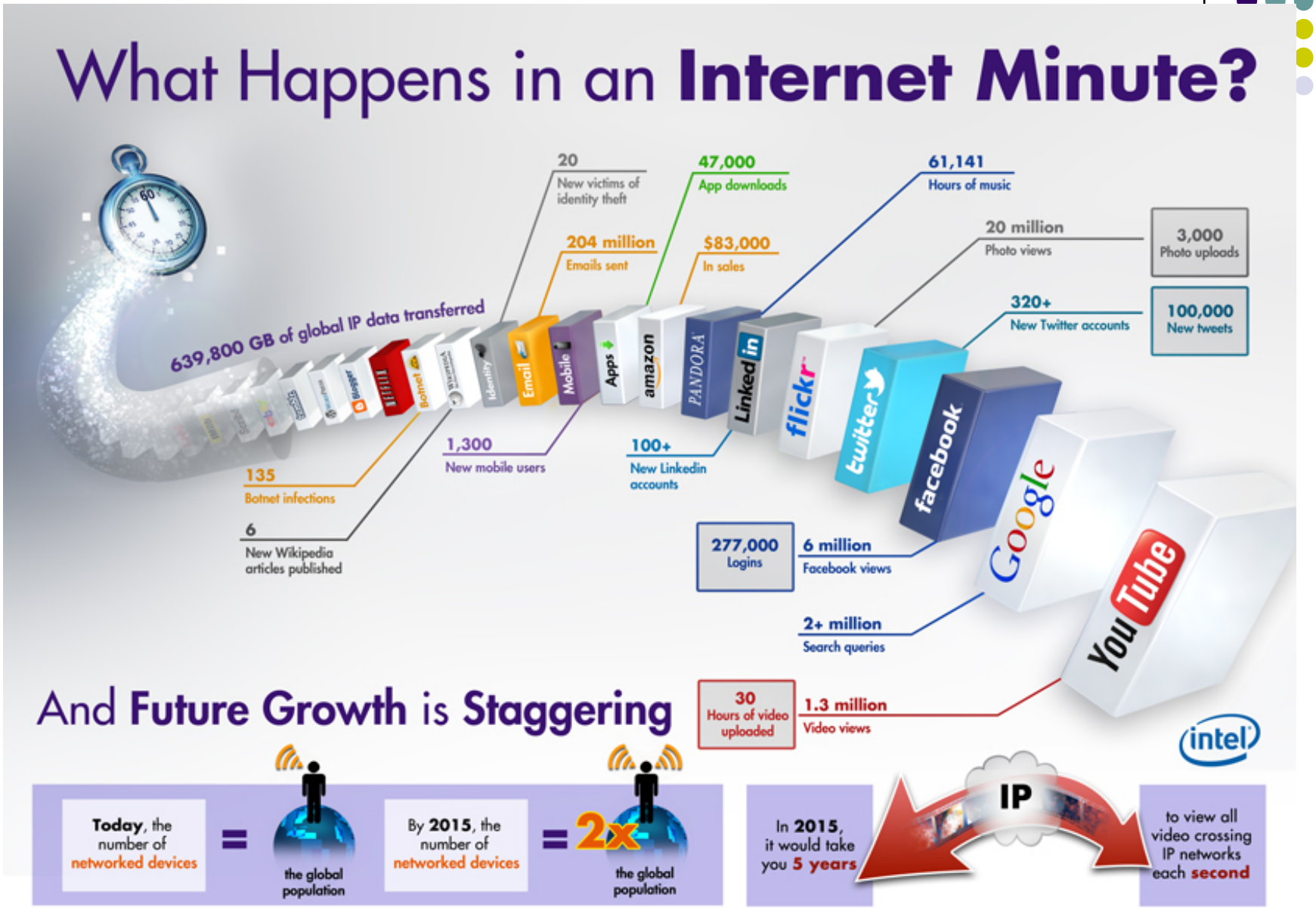
Overview

- **Introduction to ML & DM**
- Examples
- Terminology: Concept, Example, Attribute
- Search & Bonferroni's principle
- Curve fitting: a more complex example
- Table Look-up

Data?

- Data (encoded) ≠ information (meaning)
- Society produces huge amount of data
  - Sources: weather, business, science, medicine, economics, geography, environment, sports, …
  - Personal data: GPS, e-mails, tweets, …
- Really large?
- Larger than the Web?
  (visible web: 47 billions of web pages indexed by Google)

Data communication



[Picture from: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html]

© Copyright attached

Largest database (2013)

1. World Data Centre for Climate (6 PB)
2. National Energy Research Computing Center (3 PB)
3. AT&T (312 TB)
4. Google
5. Sprint
6. ChoicePoint (LexisNexis)
7. YouTube (45 TB)
8. Amazon (42 TB)
9. Central Intelligence Agency
10. Library of Congress (20 TB)

Data?

- Data (encoded) ≠ information (meaning)
- Society produces huge amount of data

- Potentially valuable resource?
- Overwhelmed by data (the amount of data stored in the world's databases doubles every 20 months)
  “More and more noise on the web”
- Raw data is useless: need techniques to automatically extract information from it
  - Data: recorded facts
  - Information: patterns underlying the data

What is Data Mining?

- Discovery of *useful*, possibly unexpected, *patterns* in data
- Extracting
  - implicit information to *explicit* information,
  - previously unknown,
  - potentially useful

  information from data
- More precisely
  Algorithms for acquiring structural descriptions from examples
- Needed: programs that detect patterns and regularities in the data

What is Data Mining?

- Structural descriptions represent patterns *explicitly*
  - Can be used to understand and explain
  - Can be used to predict outcome in new situation
  - Prediction is derived (*maybe even more important*)
- Patterns expressed as a *black box* or as a *transparent* box (revealing the structure of the pattern)?
- Example: Selection of the most appropriate predictors to detect a given form of cancer
- Methods originate from artificial intelligence, statistics, and research on databases

What is Data Mining?

- Subsidiary issues (not always simple to process):
  - Data cleaning: detection of bogus data (this is the rule, not the exception)
    - e.g., age = 150
    - Entity resolution
  - Visualization: something better than megabyte files of output (a picture is worth a 1,000 words)
  - Warehousing (organize) of data (for retrieval)
- and machine learning? or just learning?

Limits

- Text Mining (unstructured data)
  - Text categorization
    - Language identification
    - Topical text classification (metadata extraction)
    - Authorship attribution, profiling
  - Text summarization
- Web Mining (semi-structured data)
  - Social networks
    - Community finding
    - Reputation follow-up

Prediction (in real world)

- Standard & Poor's rating
  - give a AAA score for the CDO (collateralized debt obligation and the 2008 financial crisis)
  - estimated risk 0.12%, real 28%
  - large deviation, need to learn and improve the rating!
- US Election (2008)
  - Monica Crowley, one week before the election
    "McCain will win by half a point"
    Obama: 365 vs. McCain: 173
  - One week after, M. Crowley, said "McCain had run a terrific campaign… a change election year"
  - And we continue to trust Crowley's predictions!

Learning?

- Definitions of “learning” from dictionary:
  - To get knowledge of by study, experience, or being taught (difficult to measure)
  - To become aware by information or from observation
  - To commit to memory
  - To be informed of, ascertain; to receive instruction
    → relatively simple for computers
- Operational definition
  - Things learn when they change their behavior in a way that makes them perform better in the future.
- Does learning imply intention? (otherwise it is only trained)
- And relation to data mining?

Learning?

- Data Mining: practical topic, learning in a *practical*, not in a theoretical, sense
  - Methods to find and describe *structural patterns* (see below)
  - Opposed to theoretical learning, theorem proving, …
- Data (input)
  - Set of examples / observations.
- Output
  - Actual description of the data
  - Predictions about new examples
- Recall: garbage in, garbage out
  (nothing to learn (noisy data), we will learn nothing!)

Overview

- Introduction to ML & DM
- **Examples**
- Terminology: Concept, Example, Attribute
- Search & Bonferroni's principle
- Curve fitting, a more complex example
- Table Look-up

Structural Patterns

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

Structural Patterns

Example If-Then rules

- From each line, we can derive a rule
  3 x 2 x 2 x 2 = 24 (# possible values)
- Or using more general rule like

  If (`tear production rate` = `reduced`)
  then `recommendation` = `none`
  otherwise if (`age` = `young`) and (`astigmatic` = `no`)
  then `recommendation` = `soft`

  But this is a simple example
  - all combinations of possible values
  - values are specified for all the features
  - in practice, we have to cope with errors and noise in the data

Example: Weather problem

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Example

- Decision list (sequence)

  If (`outlook` = `sunny`) and (`humidity` = `high`)
  then `play` = `no`
  If (`outlook` = `rainy`) and (`windy` = `true`)
  then `play` = `no`
  If (`outlook` = `overcast`) then `play` = `yes`
  If (`outlook` = `normal`) then `play` = `yes`
  If none of the above then `play` = `yes`

- Interpret the rules in order (?)
- This is a *classification rule* (predict the class)

Example

Build rules that strongly associate different attribute values (*association rule*) and predict the value of arbitrary attribute

If (`temperature` = `cold`) then `humidity` = `normal`

If (`humidity` = `normal`) and (`windy` = `false`)
then `play` = `yes`

If (`outlook` = `sunny`) and (`play` = `no`)
then `humidity` = `high`

If (`windy` = `false`) and (`play` = `no`)
then `outlook` = `sunny` and `humidity` = `high`

…

We have more association rules than classification rules.
And with numerical values?

Example: Weather problem

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Example

If (`outlook` = `sunny`) and (`humidity` > `83`)
then `play` = `no` ...

- The construction process is more complex due to the presence of numerical values.

- Another example: the contact lenses problem
- Prescription of *hard*, *soft* or *none* based four attributes
- An example that classify all the examples (next slide)
- The rules are complete and deterministic (a unique prescription for every conceivable example).

Contact Lenses Example

If (tear production rate = reduced) then recommendation = none

If (age = young) and (astigmatic = no) and (tear production rate = normal) then recommendation = soft

If (age = pre-presbyopic) and (astigmatic = no) and (tear production rate = normal) then recommendation = soft

If (age = presbyopic) and (spectacle prescription = myope) and (astigmatic = no) then recommendation = none

If (spectacle prescription = hypermetrope) and (astigmatic = no) and (tear production rate = normal) then recommendation = soft

If (spectacle prescription = myope) and (astigmatic = yes) and (tear production rate = normal) then recommendation = hard

If (age = young) and (astigmatic = yes) and (tear production rate = normal) then recommendation = hard

If (age = pre-presbyopic) and (spectacle prescription = hypermetrope) and (astigmatic = yes) then recommendation = none

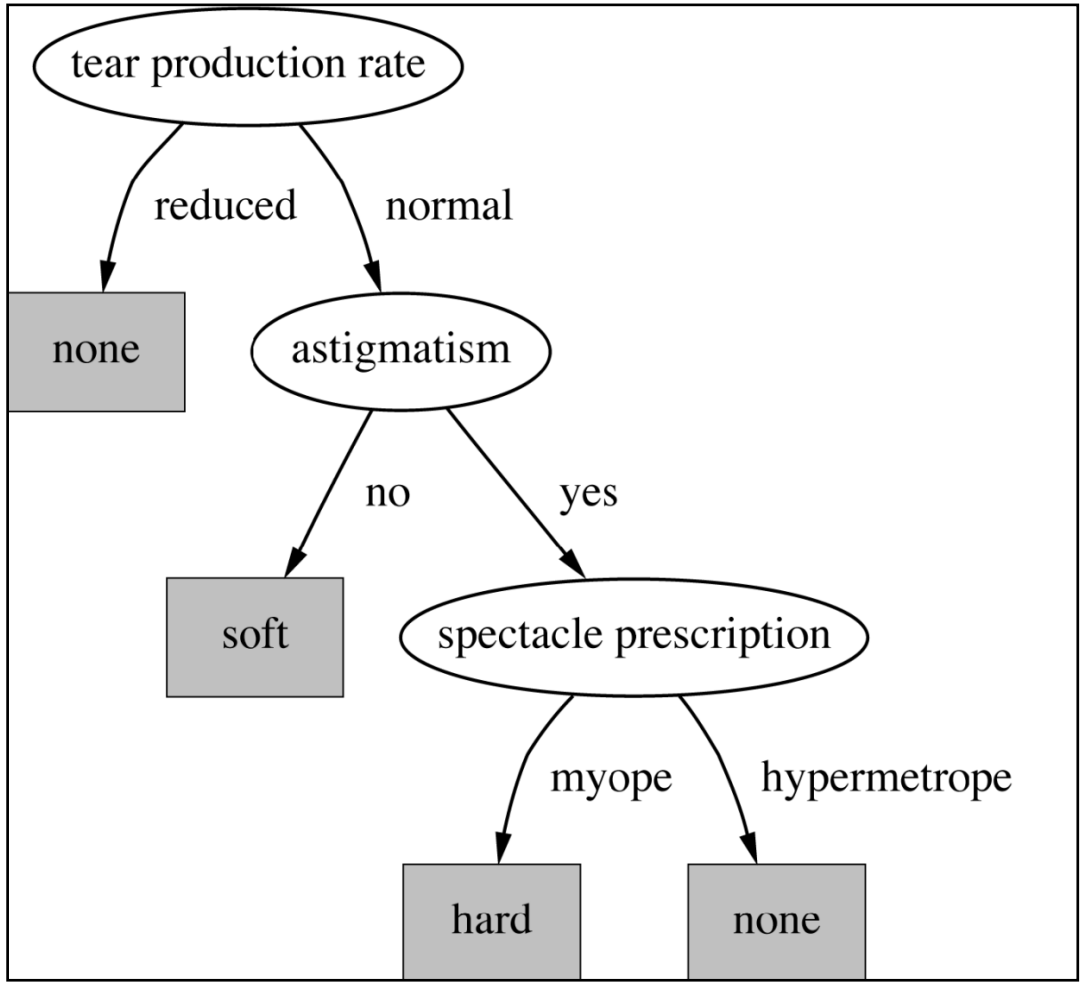
If (age = presbyopic) and (spectacle prescription = hypermetrope) and (astigmatic = yes) then recommendation = none

Example

- Generally, we do not have examples with all attributes values
  - Sometimes no rule applies
  - More than one rule may apply (give weights / probabilities to each rule)
- Good question
  "Can we find a smaller rule set performing as well?"
- A more concise representation (decision tree)
  (Small is beautiful!, Smaller is excellent!)
  Simpler to maintain

Example

- Decision tree (which classifies two examples incorrectly)



Example

- The Iris dataset (R.A. Fisher, mid-1930)
- Prescription for three possible iris types:
  setosa, *versicolor* or *virginica*
- 150 examples (50 from each type)
- Four attributes (all measurements in cm).

Example: The Iris

	sepal length	sepal width	petal length	petal width	type
1	5.1	3.5	1.4	1.4	setosa
2	4.9	3.0	1.4	1.4	setosa
3	4.7	3.2	1.3	1.3	setosa
4	4.6	3.1	1.5	1.5	setosa
…					
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
…					
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica

Example: The Iris

If (petal length < 2.45) then Iris setosa

If (sepal width < 2.10) then Iris versicolor

If (sepal width < 2.45) and (petal length < 4.55)
then Iris versicolor

If (sepal width < 2.95) and (petal length < 1.35)
then Iris versicolor

If (petal width < 1.75) and (sepal width < 3.05)
then Iris virginica

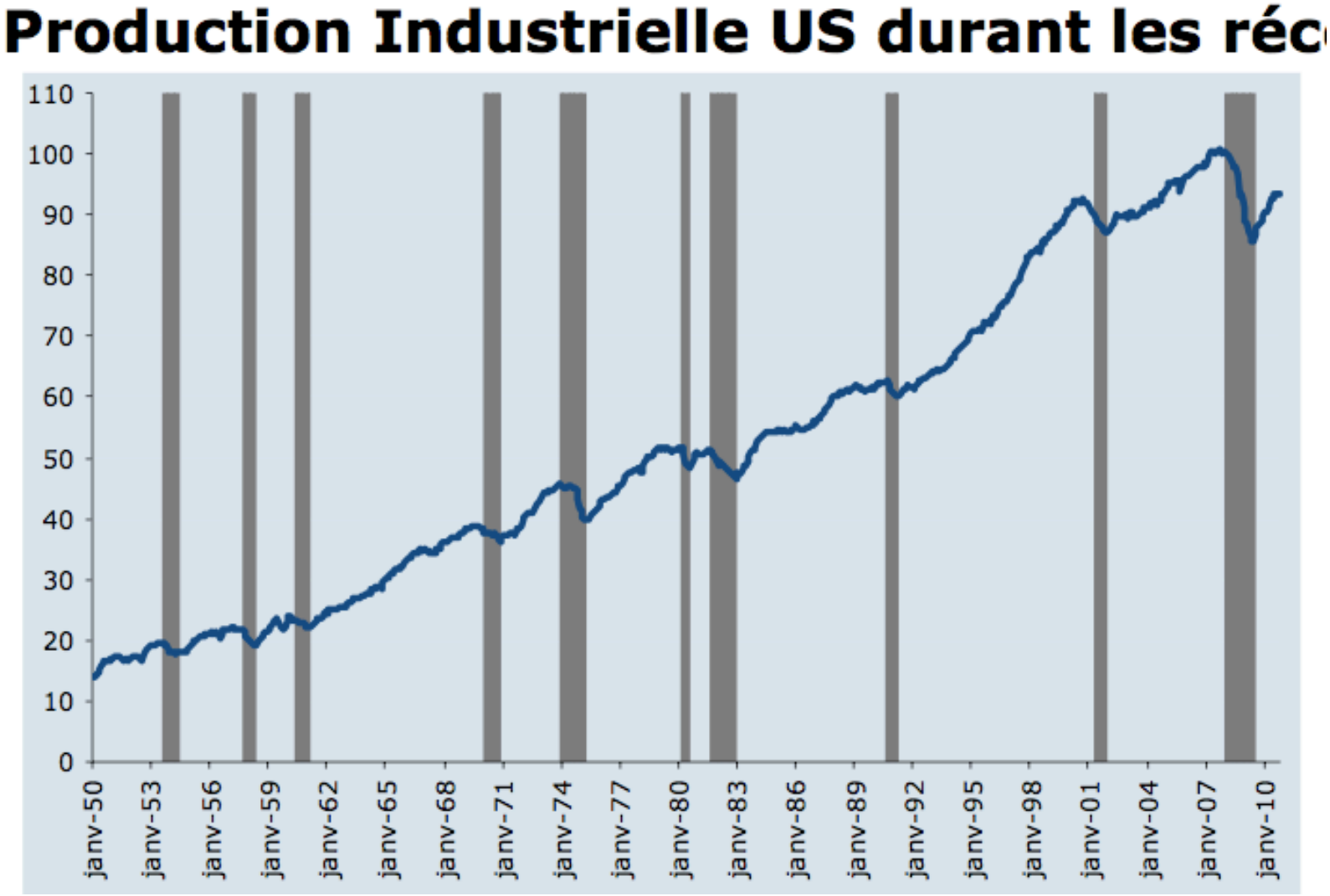
If (petal length > 4.95) and (petal width < 1.55)
then Iris virginica

Example: Econometrics

- Defining today if an economy *is in recession* or not (all sectors are subject to a decrease of their revenues) Not an easy task!
- Define *a posteriori* (two years after), it is easier
- Defining currently, with objective measures, more problematic (subject to discussion, political concern)
- Based on
  - indprod: Industrial production
  - payroll: total number of paid US workers (non-farm, gov)
  - unempl: unemployment rate

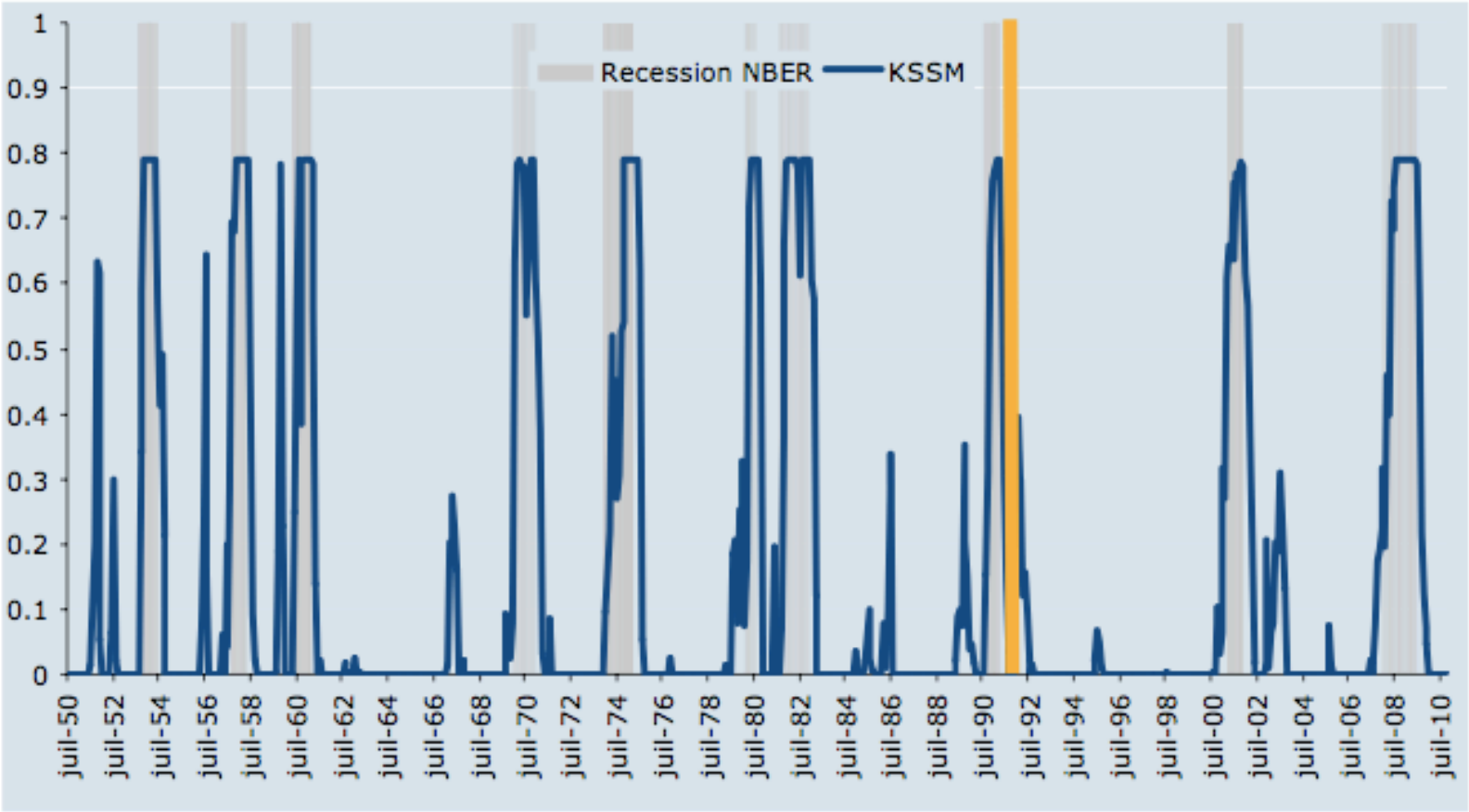
We can define when the US economy is in recession (defined by the NBER, National Bureau of Economic Research) with a latency of around 3 months

Example: Econometrics



Example: Econometrics

On travaille ici aussi avec les indicateurs économiques suivants: INDPROD, PAYROLL, UNEMPL (variations sur 6 mois)



Example

- Example: 209 different computer configurations

Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
125	256	6000	256	16	128	198
29	8000	32000	32	8	32	269
480	512	8000	32	0	0	67
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

With numeric class

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	5
sunny	80	90	true	0
overcast	83	86	false	55
rainy	70	96	false	45
rainy	68	80	false	65
rainy	65	70	true	45
overcast	64	65	true	60
sunny	72	95	false	0
sunny	69	70	false	70
rainy	75	80	false	45
sunny	75	70	true	50
overcast	72	90	true	55
overcast	81	75	false	75
rainy	71	91	true	10

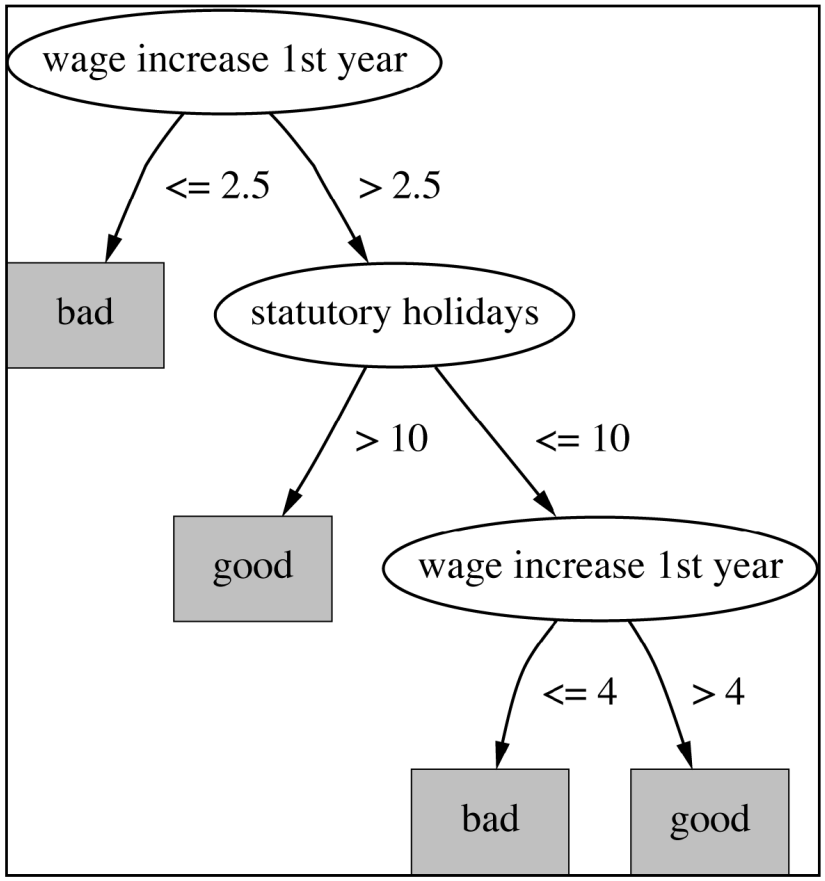
Example: Labor negotiations

- Contract negotiations (in 1987-88 in Canada)
  Prescription in *acceptable*, or *unacceptable*
  40 examples (17 attributes)

Attribute	Type	1	2	3
Duration	Number of years	1	2	3
Wage increase first year	Percentage	2%	3%	4.2%
Wage increase 2nd year	Percentage	?	4%	4.5%
Wage increase 3rd year	Percentage	?	?	5.5%
Cost of living adjustment	{none, tcf, tc}	none	tcf	?
Working hours / week	number of hours	28	35	38
…				

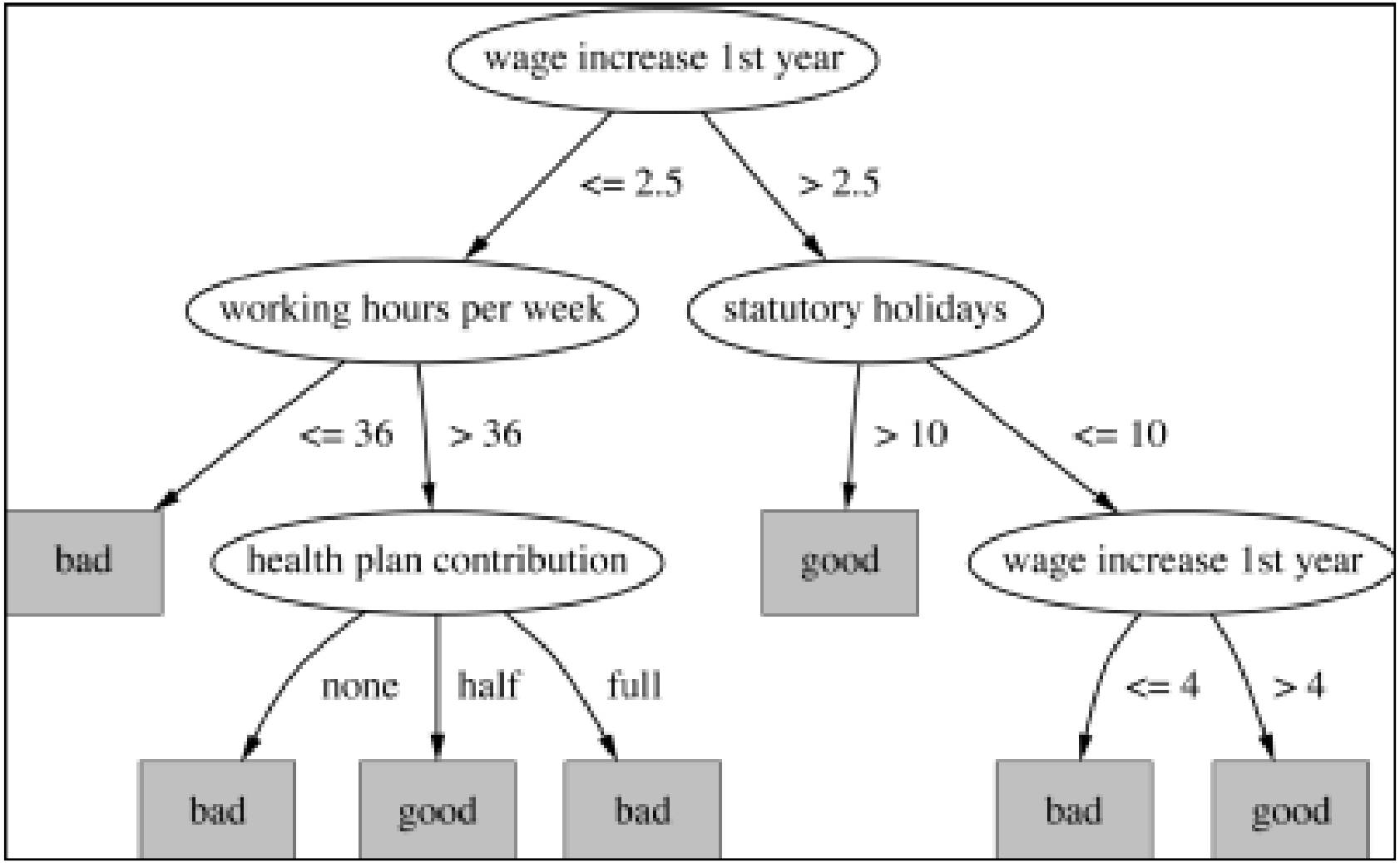
Example: Labor negotiations

The resulting decision tree



Example: Labor negotiations

Another decision tree



Example: Soybean classification

- Identification of rules for diagnosing soybean diseases
- 680 examples
- 35 attributes

Attribute	Number of values	Sample value
Time of occurrence	7	July
Precipitation	3	Above normal
Condition	2	Normal
Mold growth	2	Absent
Condition of fruit pods	4	Normal
Fruit spots	5	?
Condition	2	Abnormal
Leaf spot size	3	?
Condition	2	Abnormal
Stem lodging	2	Yes
Condition	3	Normal
	19	Diaporthe stem canker

Soybean classification

Examples of resulting rules

If (leaf condition is normal) and
(stem condition is abnormal) and
(stem cankers is below soil line) and
(canker lesion color is brown)
then diagnosis is “rhizoctonia root rot”

If (leaf malformation is absent) and
(stem condition is abnormal) and
(stem cankers is below soil line) and
(canker lesion color is brown)
then diagnosis is “rhizoctonia root rot”

But domain knowledge is also useful
(leaf condition is normal) = (leaf malformation is absent)

Applications (Among Many)

- The result of learning —or the learning method itself— is deployed in practical applications
- Processing loan applications
- Screening images for oil slicks
- Electricity supply forecasting
- Diagnosis of machine faults
- Marketing and sales
- Separating crude oil and natural gas
- Reducing banding in rotogravure printing
- Finding appropriate technicians for telephone faults
- Scientific applications: biology, astronomy, chemistry

Useful for the Bank

- Given: questionnaire with financial and personal information
- Question: should money be lent?
  (binary decision "accept" or "reject")
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers (human processing)
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?
  No trouble, no problem, but ...
  - Borderline cases are most active customers
    you can win a lot of money with them!

Useful for the Bank

- 1,000 training examples of borderline cases
- 20 attributes:
  - age
  - years with current employer
  - years at current address
  - years with the bank
  - other credit cards possessed,…
- Learned rules: correct on 70% of cases
  - human experts only 50%
- Rules could be used to explain decisions to customers

Applications

- Companies precisely record massive amounts of marketing and sales data
- Applications:
  - Customer loyalty:
    identifying customers that are likely to defect by detecting changes in their behavior (e.g. banks/phone companies)
  - Special offers:
    identifying profitable customers (e.g., reliable owners of credit cards that need extra money during the holiday season)
    (need a special offer to recruit old customer, e.g., iPhone & Swisscom)
  - Market basket analysis
    Association techniques find groups of items that tend to occur together in a transaction (used to analyze checkout data)

Market Basket Analysis

Where should detergents be placed in the store to maximize their sales?

Are window cleaning products purchased when detergents and orange juice are bought together?

Is soda typically purchased with bananas?
Does the brand of soda make a difference?

How are the demographics of the neighborhood affecting what customers are buying?

Market Basket Analysis

- There has been a considerable amount of research in the area of Market Basket Analysis. Its appeal comes from the clarity and utility of its results, which are expressed in the form *association rules*.
  Given
  - A database of transactions
  - Each transaction contains a set of items
  - Objective: Find all rules X→Y that correlate the presence of one set of items X with another set of items Y
    Example: when a customer buys bread and butter, they buy milk 85% of the time.

Market Basket Analysis

- While association rules are easy to understand, they are not always useful.
- **Useful:** On Fridays convenience store customers often purchase diapers and beer together.
- **Trivial:** Customers who purchase maintenance agreements are very likely to purchase large appliances.
- **Inexplicable:** When a new Super Store opens, one of the most commonly sold item is light bulbs.

Market Basket Analysis

- Examine what people buy together to discover patterns.
  1. What pairs of items are unusually often found together at the supermarket?

     Answer: diapers and beer.
     Run a sale on diapers; raise the price of beer.
  2. What books are likely to be bought by the same Amazon customer?
     Previous sales with a given book, customer profile, …
  3. → Recommendation systems, online advertising

Overview

- Introduction to ML & DM
- Examples
- **Terminology**: Concept, Example, Attribute
- Search & Bonferroni's principle
- Curve fitting, a more complex example
- Table Look-up

Terminology

- CS: concepts/attributes → model → prediction
- Statistician: independent var. → model → dependant var.
- ML: features → model → responses
- ML: predictors → function approximation → responses
- Nature:
  - nominal (discrete, categorical)
  - ordinal (ordered categorical)
  - quantitative
- Response:
  - classification (qualitative)
    binary (numeric code or target)
  - regression (quantitative)

What is …

- What's a concept?
  - Classification, association, clustering, numeric prediction
- What's in an example?
  - Relations, flat files, recursion
- What's in an attribute?
  - Nominal, ordinal, interval, ratio

What's a Concept?

- Styles of learning:
  - Classification learning:
    predicting a discrete class
  - Association learning:
    detecting associations between features
  - Clustering:
    grouping similar instances into clusters
  - Numeric prediction:
    predicting a numeric quantity
- Concept: thing to be learned
- Concept description:
  output of learning scheme

What's a Concept?

- Classification learning
- Example problems: weather data, contact lenses, irises, labor negotiations
- Classification learning is *supervised*
  - Scheme is provided with actual outcome
- Outcome is called the *class* of the example
- Measure success on fresh data for which class labels are known (*test data*)
- In practice success is often measured subjectively

What's a Concept?

- Association learning
- Can be applied if no class is specified and any kind of structure is considered “interesting”
- Difference to classification learning:
  - Can predict any attribute’s value, not just the class, and more than one attribute’s value at a time
  - Hence: far more association rules than classification rules
  - Thus: constraints are necessary
    Minimum coverage and minimum accuracy

What's a Concept?

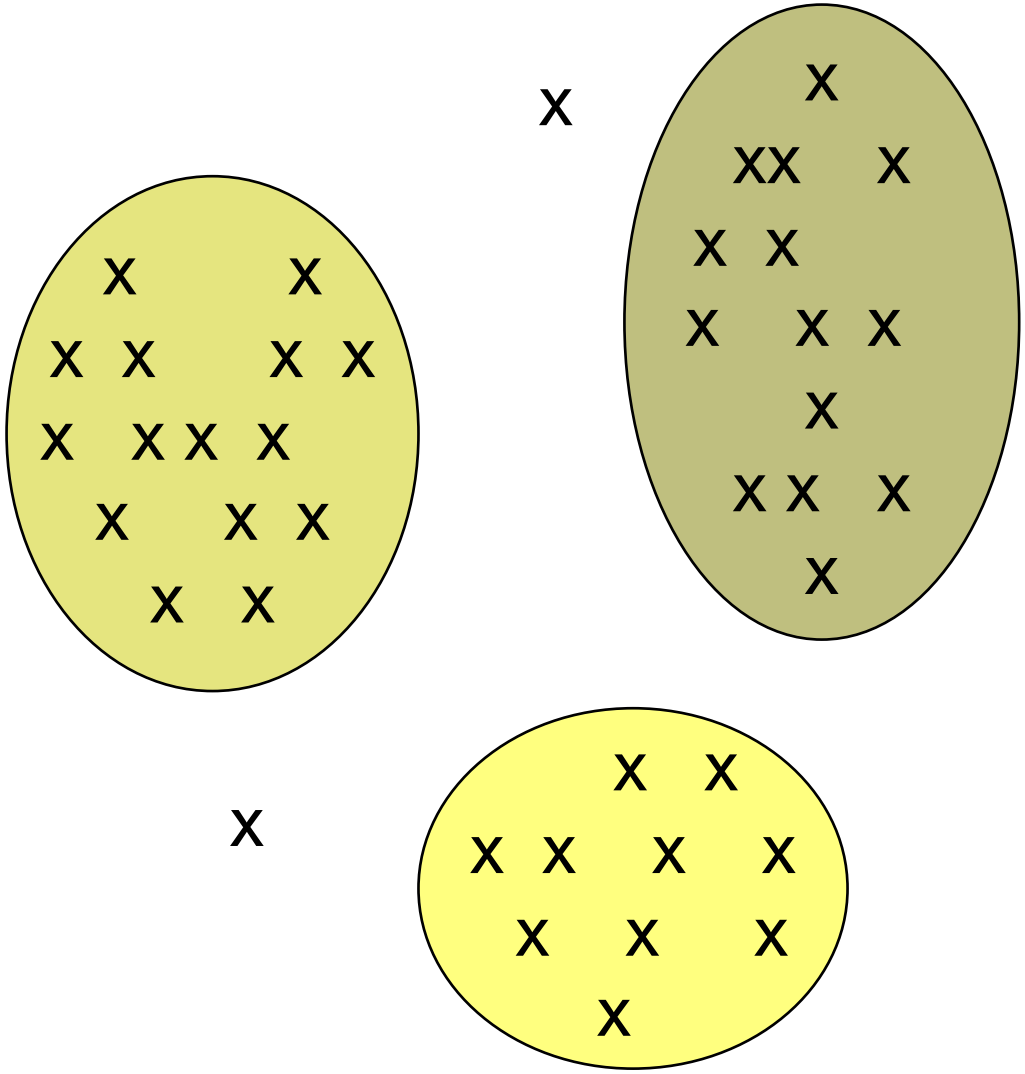
- Clustering
- Finding groups of items that are similar
- Clustering is *unsupervised*
  The class of an example is not known
- Success often measured subjectively

Sepal length	Sepal width	Petal length	Petal width	Type
5. 1	3. 5	1. 4	0. 2	Iris setosa
4. 9	3. 0	1. 4	0. 2	Iris setosa
7. 0	3. 2	4. 7	1. 4	Iris versicolor
6. 4	3. 2	4. 5	1. 5	Iris versicolor
6. 3	3. 3	6. 0	2. 5	Iris virginica
5. 8	2. 7	5. 1	1. 9	Iris virginica

Clustering

Classify these pictures

Clusters



What's a Concept?

- Numerical prediction
- Variant of classification learning where “class” is numeric (called “*regression*”)
- Learning is supervised
  - Scheme is being provided with target value
- Measure success on test data

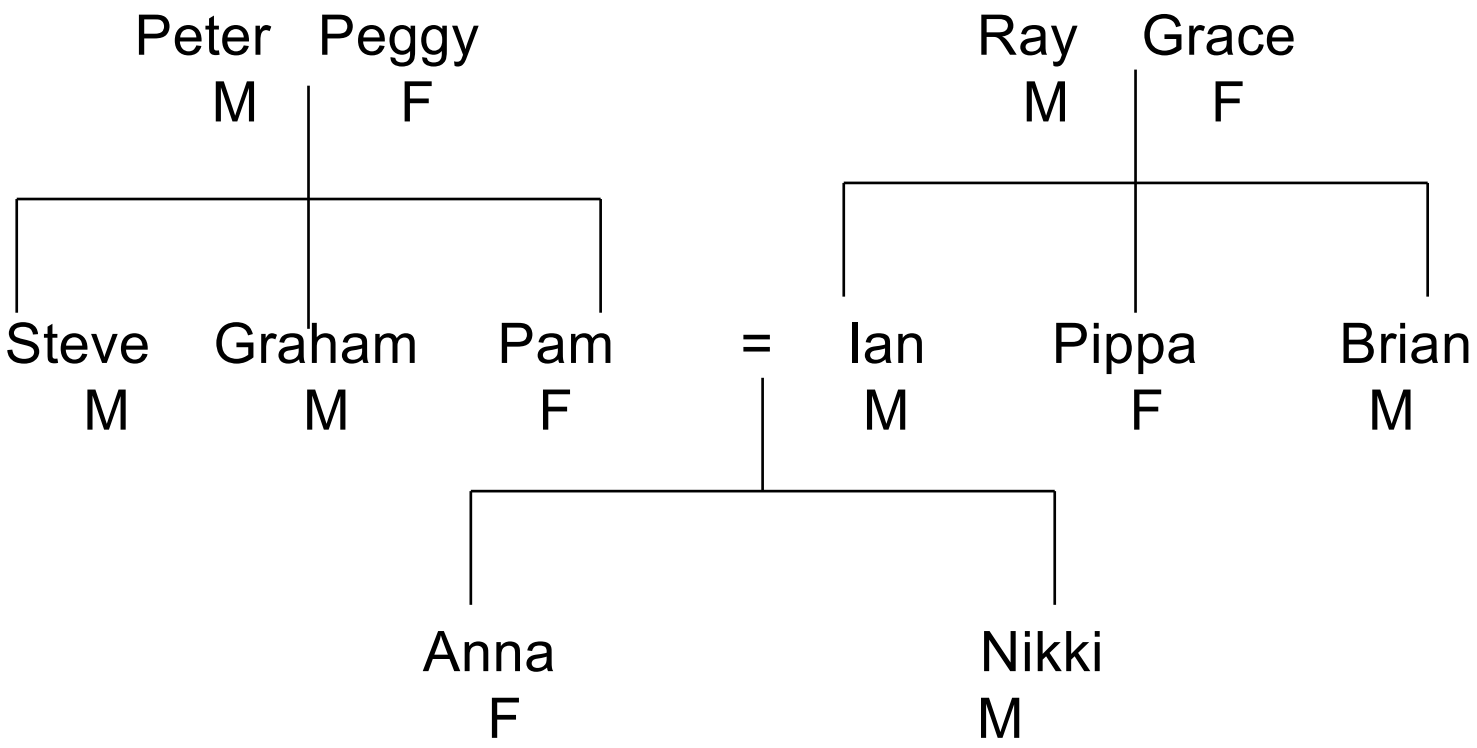
Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	5
sunny	hot	high	true	0
overcast	hot	high	false	55
rainy	mild	high	false	40
…				

What's in an Example?

- Instance: specific type of example
  - Things to be classified, associated, or clustered
  - Individual, independent example of target concept
  - Characterized by a predetermined set of attributes
- Input to learning scheme: set of instances/dataset
  - Represented as a single relation/flat file
- Rather restricted form of input
  - No relationships between objects
- Most common form in practical data mining

What's in an Example?

Easy…A Family Tree



How to represent it?

What's in an Example?

- Representation as a table

Name	Gender	Parent1	Parent2
Peter	male		
Peggy	female		
Steven	male	Peter	Peggy
Graham	male	Peter	Peggy
Pam	female	Peter	Peggy
Ian	male	Grace	Ray
Pippa	female	Grace	Ray
Brian	male	Grace	Ray
Anna	female	Pam	Ian
Nikki	male	Pam	Ian

What's in an Example?

The "*sister-of*" relation

Person1	Person2	Sister-of
Peter	Peggy	no
Peter	Steven	no
…	…	…
Steven	Peter	no
Steven	Graham	no
Steven	Pam	yes
…	…	…
Ian	Pippa	yes
…	…	…
Anna	Nikki	yes
…	…	…
Nikki	Anna	yes

Person1	Person2	Sister-of
Steven	Pam	yes
Graham	Pam	yes
Ian	Pippa	yes
Brian	Pippa	yes
Anna	Nikki	yes
Nikki	Anna	yes
all the rest		no

Closed-world assumption

What's in an Example?

A full representation in one table

First Person				Second Person				Sister
Name	Ge	Parent1	Parent2	Name	Ge	Parent1	Parent2	of?
Steven	m	Peter	Peggy	Pam	f	Peter	Peggy	yes
Graham	m	Peter	Peggy	Pam	f	Peter	Peggy	yes
Ian	m	Grace	Ray	Pippa	f	Grace	Ray	yes
Brian	m	Grace	Ray	Pippa	f	Grace	Ray	yes
Anna	f	Pam	Ian	Nikki	f	Pam	Ian	yes
Nikki	f	Pam	Ian	Nikki	f	Pam	Ian	yes
all the				rest				no

if (`second person's gender` = *`female`*) and
(`first person's parent` = `second person's parent`)
then `sister-of` = *`yes`*

What's in an Example?

- Process of flattening called “denormalization”
  - Several relations are joined together to make one
- Possible with any finite set of finite relations
- Problematic: relationships without pre-specified number of objects
  - Example: concept of *nuclear-family* (parents + children)
- Denormalization may produce spurious regularities that reflect structure of database
  - Example: “supplier” predicts “supplier address”
  - and large (100 suppliers x 1,000 products x 10,000 customers)

What's in an Example?

The "*ancestor-of*" relation: large number of possible answers

First Person				Second Person				Ancest of?
Name	Ge	Parent1	Parent 2	Name	Ge	Parent1	Parent 2	
Peter	m			Steven	m	Peter	Peggy	yes
Peter	m			Pam	f	Peter	Peggy	yes
Peter	m			Anna	f	Pam	Ian	yes
Peter	m			Nikki	f	Pam	Ian	yes
Pam	f	Peter	Peggy	Nikki	f	Pam	Ian	yes
Grace	f			Ian	m	Grace	Ray	yes
		other	positive	example				yes
		all	the	rest				no

What's in an Example?

- Infinite relations require recursion
  - **If (*`person1`* is a `parent of` *`person2`*) then *`person1`* is an `ancestor of person2`**
  - **If (*`person1`* is a `parent of` *`person2`*) and (*`person2`* is `an ancestor of` *`person3`*) then *`person1`* is `an ancestor of` *`person3`***
- Appropriate techniques are known as "inductive logic programming"
  - (e.g. Quinlan's FOIL)
  - Problems: (a) noise and (b) computational complexity

What's in an Attribute?

- Each instance is described by a fixed predefined set of features, its “attributes”
- But: number of attributes may vary in practice
  - Possible solution: “irrelevant value” flag (e.g., the number of wheels for car, trucks, and ships)
- Related problem: existence of an attribute may depend of value of another one
- Possible attribute types (“levels of measurement”):
  - *Nominal, ordinal, interval* and *ratio*

What's in an Attribute?

- **Nominal** quantities (categorical, enumerated, “discrete”)
- Values are distinct symbols
  - Values themselves serve only as labels or names
  - *Nominal* comes from the Latin word for name
- Example: attribute “outlook” from weather data
  - Values: “sunny”, ”overcast”, and “rainy”
  - or values in “blue”, “green”, “blue”, “red”
- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

What's in an Attribute?

- **Ordinal** quantities
- Impose order on values
- But: no distance between values defined
- Example: attribute “temperature” in weather data
  - Values: “hot” > “mild” > “cool”
- Note: addition and subtraction don’t make sense
- Example rule:
  `temperature` < *hot* → `play` = *yes*
- Distinction between nominal and ordinal not always clear (e.g. attribute “outlook”)

What's in an Attribute?

- **Interval** quantities
- Interval quantities are not only ordered but measured in fixed and equal units
- Example 1: attribute “temperature” expressed in degrees Fahrenheit
- Example 2: attribute “year”
- Difference of two values makes sense
  but the sum of two dates does not make sense
- Sum or product doesn’t make sense
  - Zero point is not well defined (arbitrary)

What's in an Attribute?

- **Ratio** quantities
- Ratio quantities are ones for which the measurement scheme defines a zero point
- Example: attribute “distance”
  - Distance between an object and itself is zero
- Ratio quantities are treated as real numbers
  - All mathematical operations are allowed

What's in an Attribute?

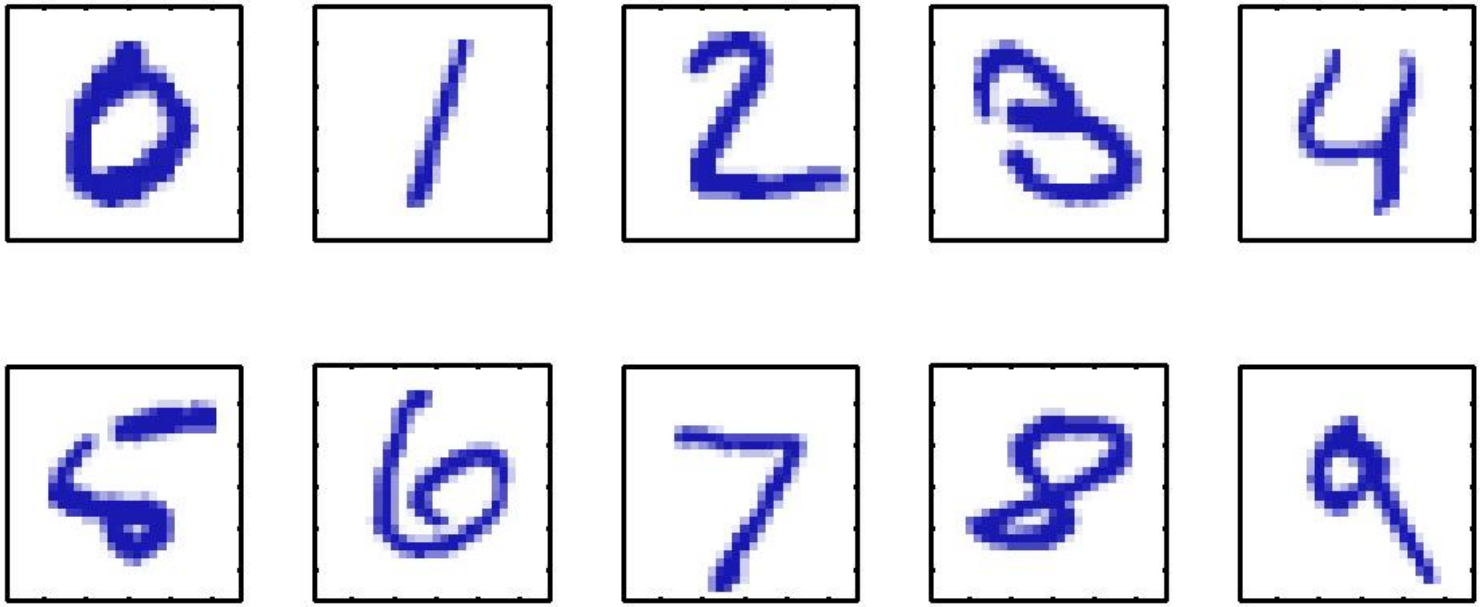
- Attributes types used in practice
- Most schemes accommodate just two levels of measurement: nominal and ordinal
- Nominal attributes are also called “categorical”, ”enumerated”, or “discrete”
  - But: “enumerated” and “discrete” imply order
- Special case: dichotomy (Boolean attribute)
- Ordinal attributes are called “numeric”, or “continuous” (Yes this is not correct!)
  - But: “continuous” implies mathematical continuity

What's in an Attribute?

- Metadata
- Information about the data that encodes background knowledge
- Can be used to restrict search space
- Examples:
  - Dimensional considerations
    (i.e. expressions must be dimensionally correct)
  - Circular orderings
    (e.g., degrees in compass)
  - Partial orderings
    (e.g., generalization/specialization relations)

Example

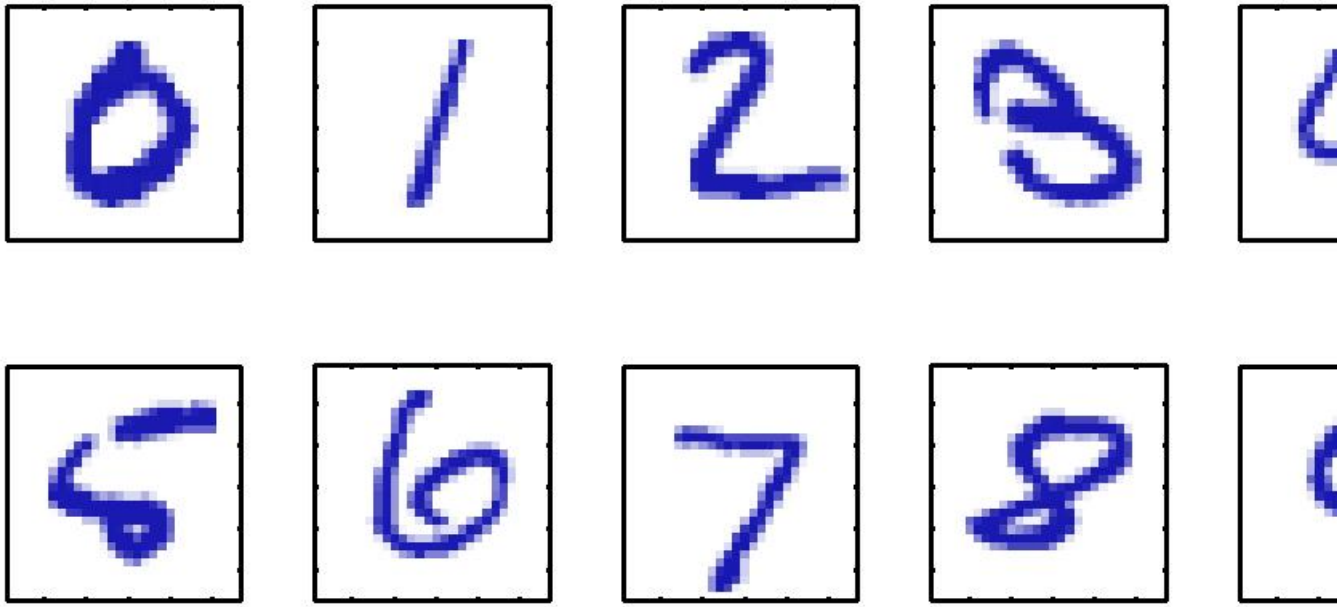
Supervised learning: hand-writing digits (zip codes)

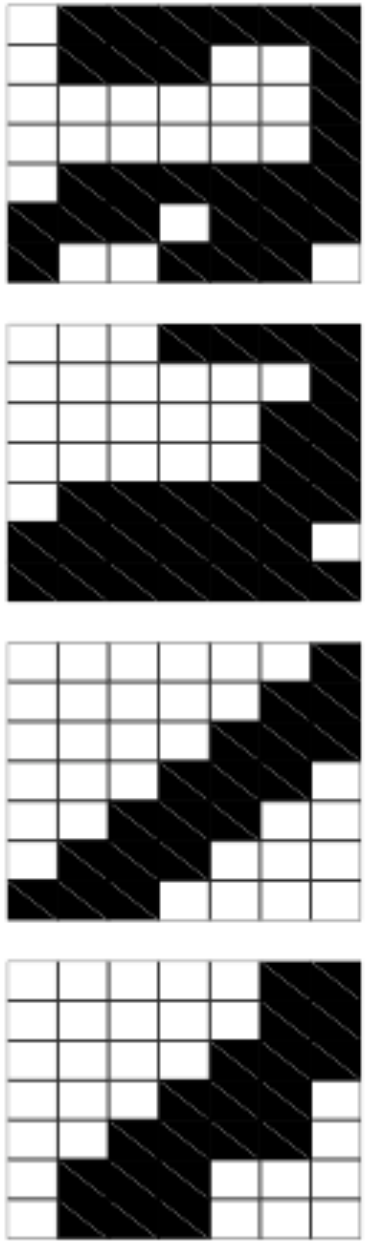


- How can you build examples?

Learning

Supervised learning: hand-writing digits (zip coc





- Categorization or regression?
- Which features?
- Can you use some prior knowledge?

Cultures

- Databases: concentrate on large-scale (non-main-memory) data.
- AI (machine-learning): concentrate on complex methods, small data (toy-size?).
- Statistics: concentrate on models.

Overview

- Introduction to ML & DM
- Examples
- Terminology: Concept, Example, Attribute
- **Search & Bonferroni's principle**
- Curve fitting, a more complex example
- Table Look-up

Generalization as Search

- Inductive learning: find a concept description that fits the data
- Example: rule sets as description language
  - Enormous, but finite, search space
- Simple solution:
  - enumerate the concept space
  - eliminate descriptions that do not fit examples
  - surviving descriptions contain target concept

Generalization as Search

- Search space for weather problem
  4 x 4 x 3 x 3 x 2 = 288 possible combinations
  With 14 rules $\rightarrow 288^{14} = 2.7 \times 10^{34}$ possible rule sets
- Other practical problems:
  - More than one description may survive (more than one max value)
  - No description may survive
  - Language is unable to describe target concept
    or data contains noise
- Another view of generalization as search:
- Hill-climbing in description space according to pre-specified matching criterion
- Most practical algorithms use heuristic search that cannot guarantee to find the optimum solution

Meaningfulness of Answers

- A big risk when data mining is that you will "discover" patterns that are meaningless.
- Statisticians call it Bonferroni's principle (1892-1960): (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.

  Due to pure chance, having a significance level of 5%, our statistical test will accept H_0 1 over 20.

  If you apply *n* tests on the same data, you will have to accept H_0 once (if $n = 20$, in mean once).

Story Behind the Story

- The “hotels” example got picked up by a newspaper reporter who spun it as
  - Stanford professor proves tracking terrorists is impossible
- This example illustrates a problem with intelligence-gathering.
- Suppose we believe that certain groups of terrorists are meeting occasionally in hotels to plot doing evil.
- We want to find people who at least twice have stayed at the same hotel on the same day.

Bonferroni's Principle

- 10^9 people being tracked
- 1,000 days
- Each person stays in a hotel 1% of the time (10 days out of 1,000)
  On a given day, we have 10^7 persons in the hotels
- Each hotel hold 100 people (so we have 10^5 hotels)
- If everyone behaves randomly (i.e., no evil-doers) will the data mining detect anything suspicious?

Bonferroni's Principle

q at some hotel

the same hotel

- Probability that persons *p* and *q* will be at the *same* hotel on day *d* :
  - $1/100 * \boxed{1/100} * \boxed{10^{-5}} = 10^{-9}$
- Probability that *p* and *q* will be at the same hotel on two given days:
  - $10^{-9} * 10^{-9} = 10^{-18}$
- Pairs of days:
  - $1/2 *[10^3*10^3] = 5*10^5$

Bonferroni's Principle

- Probability that *p* and *q* will be at the same hotel on some two days:
  - $5*10^{5} * 10^{-18} = 5*10^{-13}$
- Pairs of people:
  - $1/2 *[10^{9}*10^{9}] = 5*10^{17}$
- Expected number of suspicious pairs of people:
  - $5*10^{17} * 5*10^{-13} = 250,000$

Bonferroni's Principle

- Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice.
- Analysts have to sift through 250,010 candidates to find the 10 real cases.
  - Detection impossible!
  - But how can we improve the scheme?
- Moral:
  When looking for a property (e.g., "two people stayed at the same hotel twice"), make sure that there are not so many possibilities that *random* data will not produce facts "of interest."

Example

- We think that some people may have Extra-Sensory Perception (ESP).
- We have an experiment to verify this.
  - Predict *red* or *black* for a sequence of 10 trials (e.g., playing cards)
  - The experiments were done with a little bit than 1,000 persons.
  - We discover than 1 over 1,000 persons has a ESP (perfect prediction for 10 cards)!
- Moral?

Example

- We told this person s/he has ESP.
  - We then repeat the experiment.
  - The person fails.
- Moral?
  - As soon as you inform a person s/he has ESP, this person loses his/her ESP!
- Do you have another explanation?

Overview

- Introduction to ML & DM
- Examples
- Terminology: Concept, Example, Attribute
- Search & Bonferroni's principle
- **Curve fitting**, a more complex example
- Table Look-up

Example: Curve fitting

- Simplest form: learn a function from examples

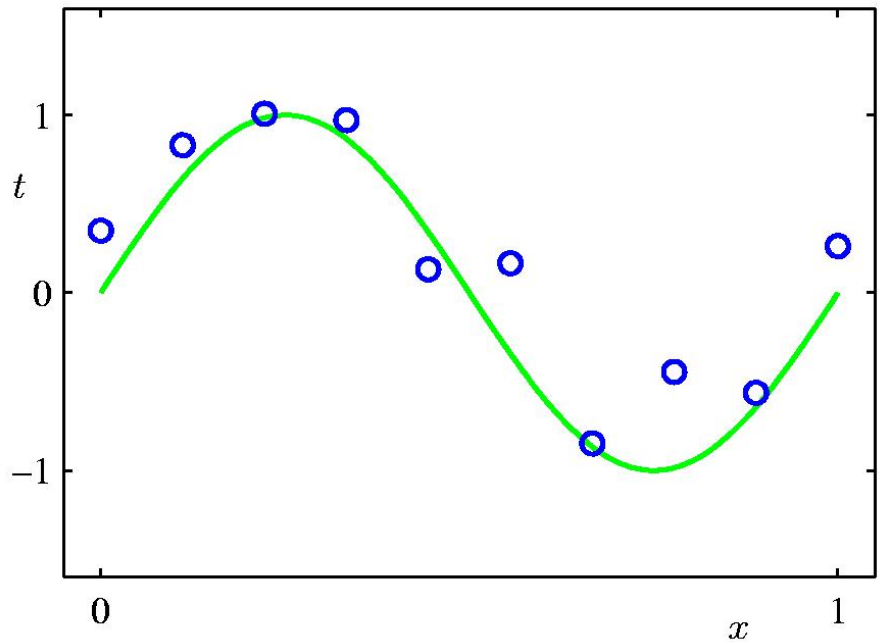
  f is the target function

  An example is a pair $(x, f(x))$ (or $(x, c(x))$)
- Problem: find a hypothesis h (a function)
  - such that $h \approx f$
  - given a training set of examples
- This is a highly simplified model of real learning:
  - Ignores prior knowledge
  - Assumes examples are given
  - Is h a good approximation? How good?

Example

- Construct/adjust *h* to agree with *f* on training set (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting: f with N = 10 blue points

$$y(x, \mathbf{w}) = w_0 + w_1 \cdot x + w_2 \cdot x^2 + \cdots + w_M \cdot x^M = \sum_{j=0}^{M} w_j \cdot x^j$$

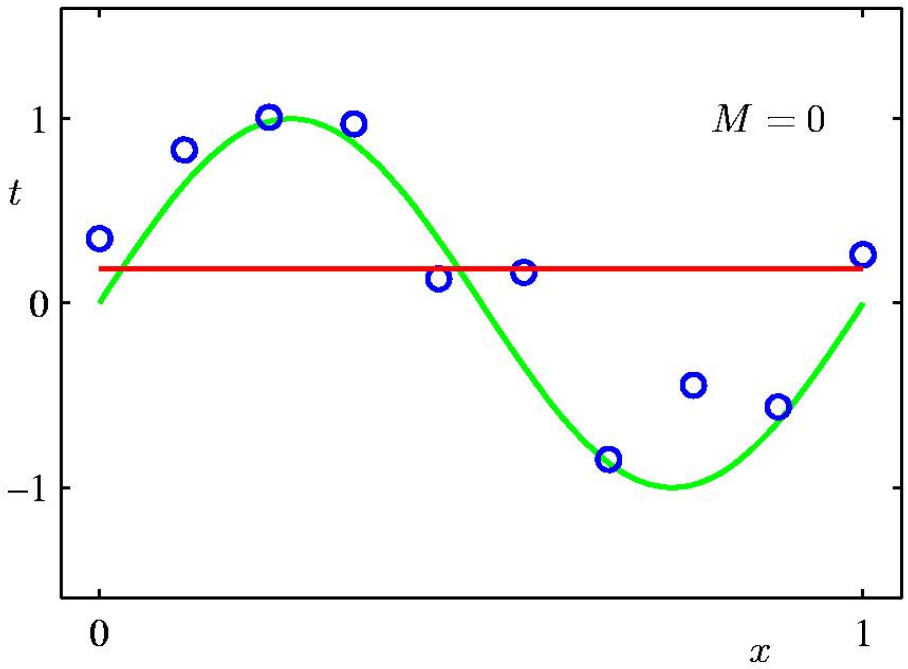


Behind, the real process in green is f = sin(2πx)

regularity + noise = observation

Example

- A very simple solution for *h* (order 0) with its graphical representation is red

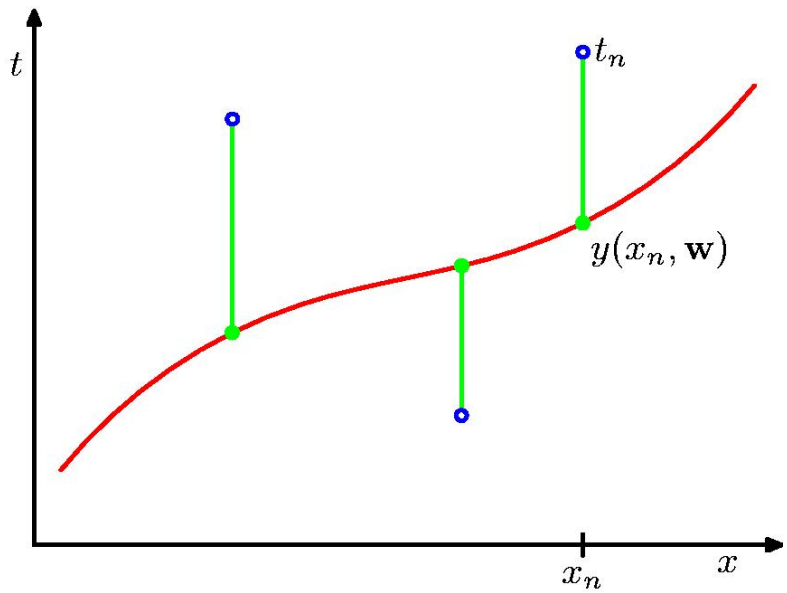


- How do you compute the quality of the model?

Example

- We can define an overall error function measuring the misfit between y(*x*,**w**) and the training set

$$Err(\mathbf{w}) = \frac{1}{2} \cdot \sum_{n=1}^{N} (y(x_n, \mathbf{w}) - t_n)^2$$

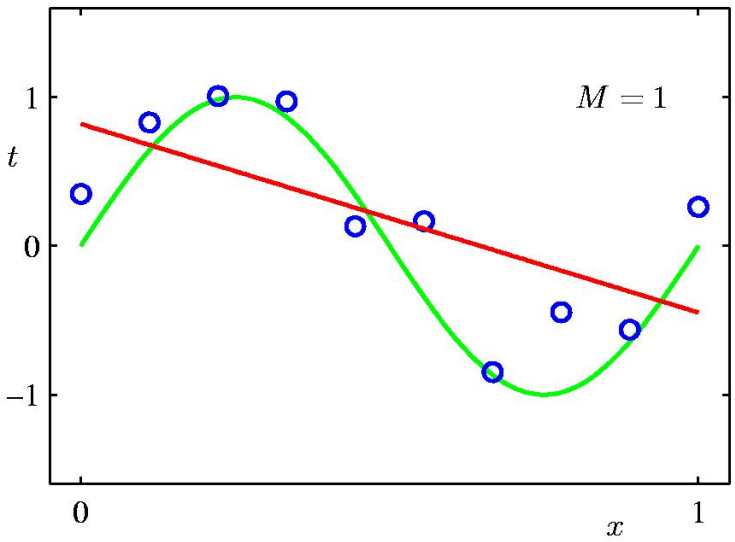


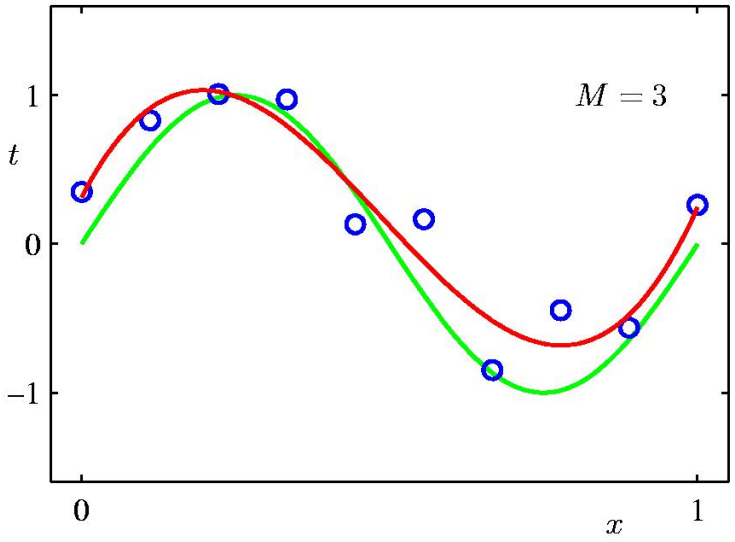
- To compare performance on different training set (or sample) with different size (N), use the root-mean-square

$$Err_{RMS} = \sqrt{2 \cdot Err(\mathbf{w})/N}$$

Example

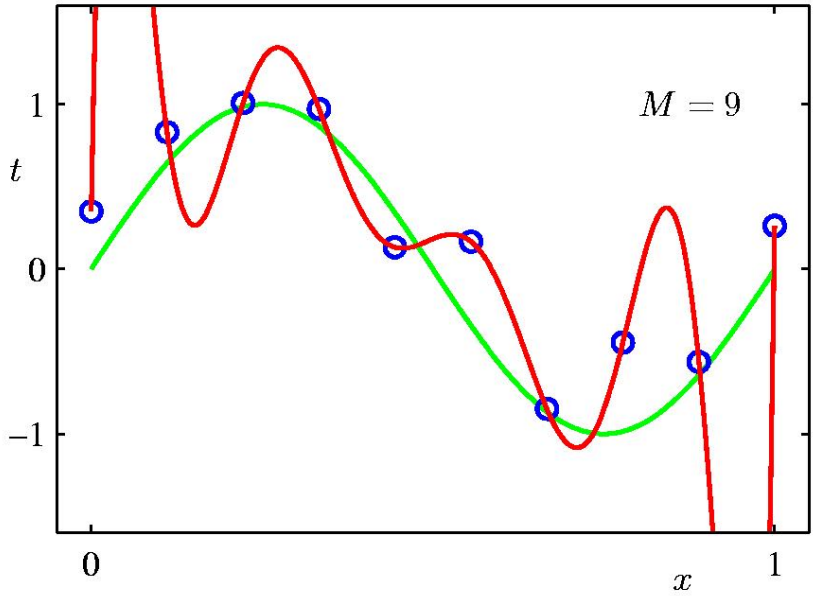
- Other solutions for *h* (a line, degree 1) or a polynomial of order 3





Example

- But we can consider more complex formula for *h* (e.g., a polynomial of order 9)
- The function *h* passes by all blue points. Great?



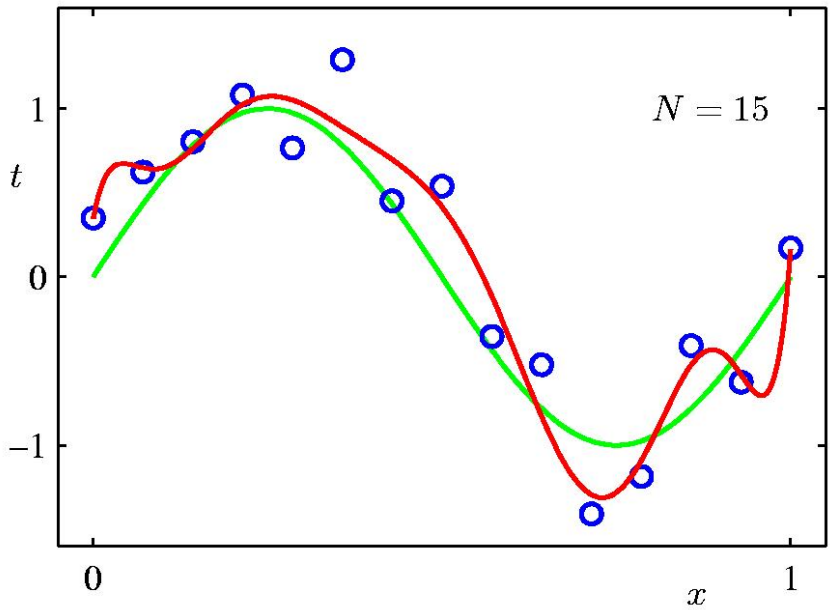
Example

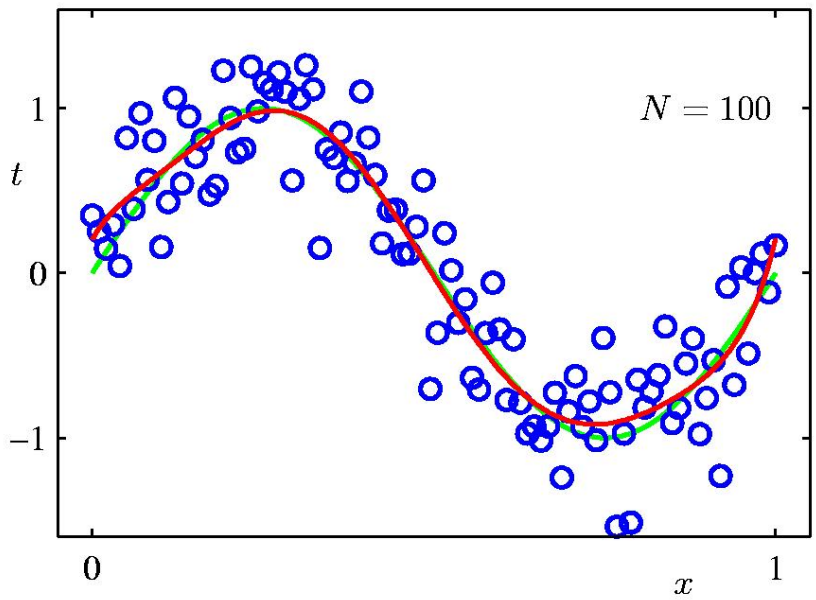
- Look at the parameter values

	M = 0	M = 1	M = 3	M = 9
w_0	0.19	0.82	0.31	0.35
w_1		-1.27	7.99	232.37
w_2			-25.43	-5,321.83
w_3			17.37	48,568.31
w_4				-231,639.3
w_5				640,042.26
w_6				-1,061,800.52
w_7				1,042,400.18
w_8				-557,682.99
w_9				125,201.43

Example

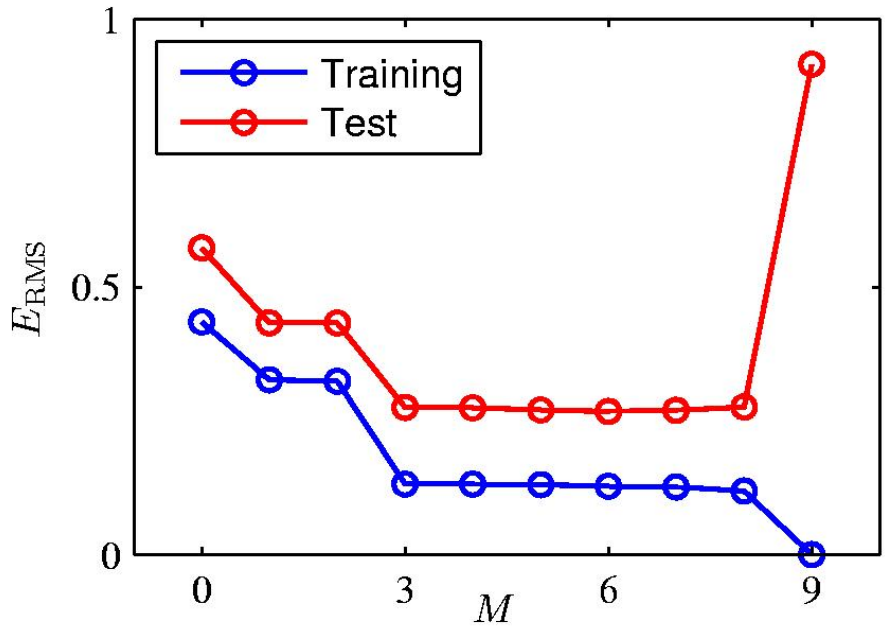
- If we increase the training set (parameter *N*), the overfitting tend to be reduced
  (the function *h* was of order 9 with *N* = 10)
- Rule of thumb: $5*h \leq N$ or $10*h \leq N$

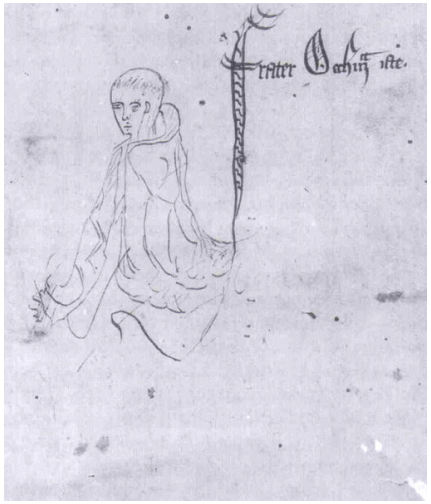




Example

- Ockham's razor: prefer the simplest hypothesis consistent with data.
  This principle proposed by William of Ockham in the fourteenth century:
  "Pluralitas non est ponenda sine neccesitate", which translates as "entities should not be multiplied unnecessarily"
  Ockham uses the razor to eliminate unnecessary hypotheses.





Example

- Maybe we can penalize using large values for the underlying coefficients
- Change the error function to

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \cdot \sum_{n=1}^{N} (y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2} \cdot \| \mathbf{w} \|^2$$

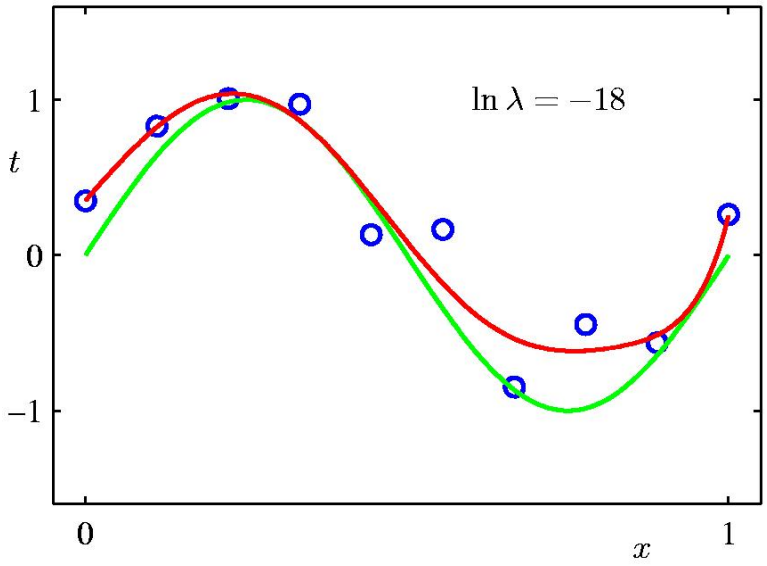
with

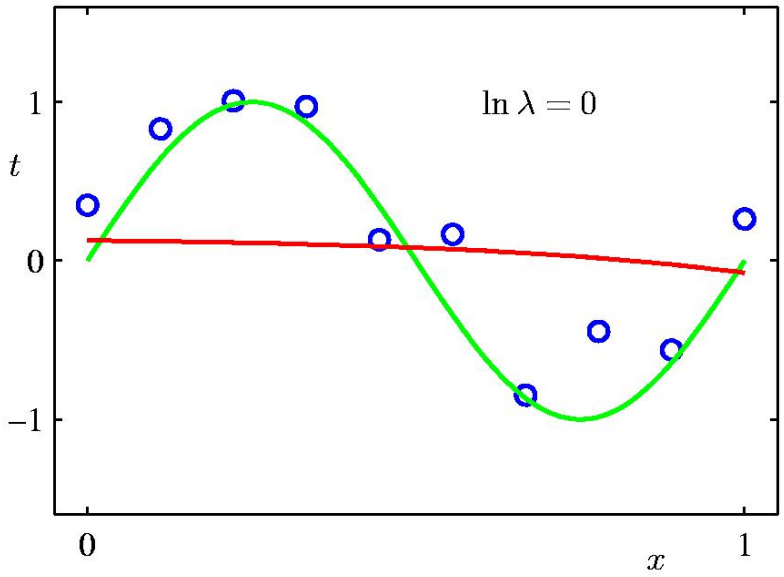
$$\| \mathbf{w} \|^2 = \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$

we can ignore w_0 in some cases

Example

- With ln(λ) = -18 and M = 9, with really penalize the large coefficient values and we obtain a better curve estimate (that is close to the real function sin(2πx).
- With ln(λ) = 0, the large coefficients are too largely penalize (only small values are possible → no effect)





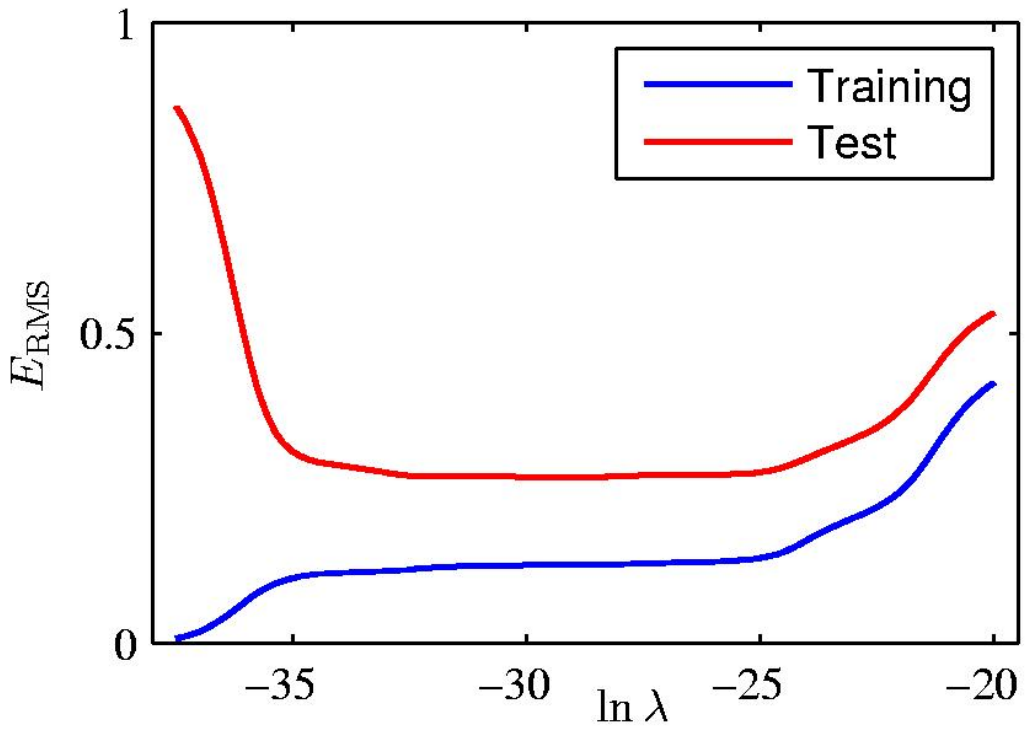
Example

- The underlying coefficients are

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

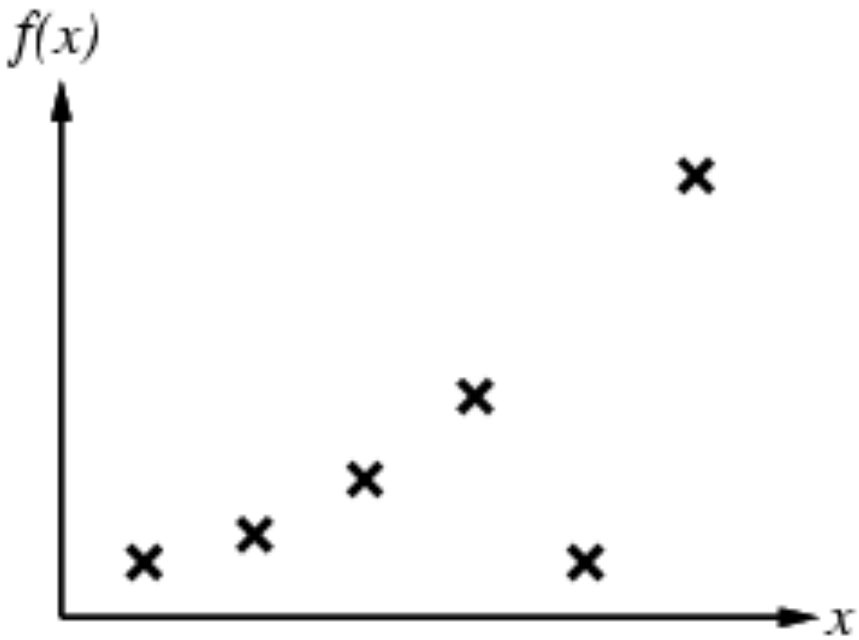
Example

- If we look at the training / test error (Err_{RMS})
- The error rate (on the test data) is decreasing for a while, reaches a minimum and then increases



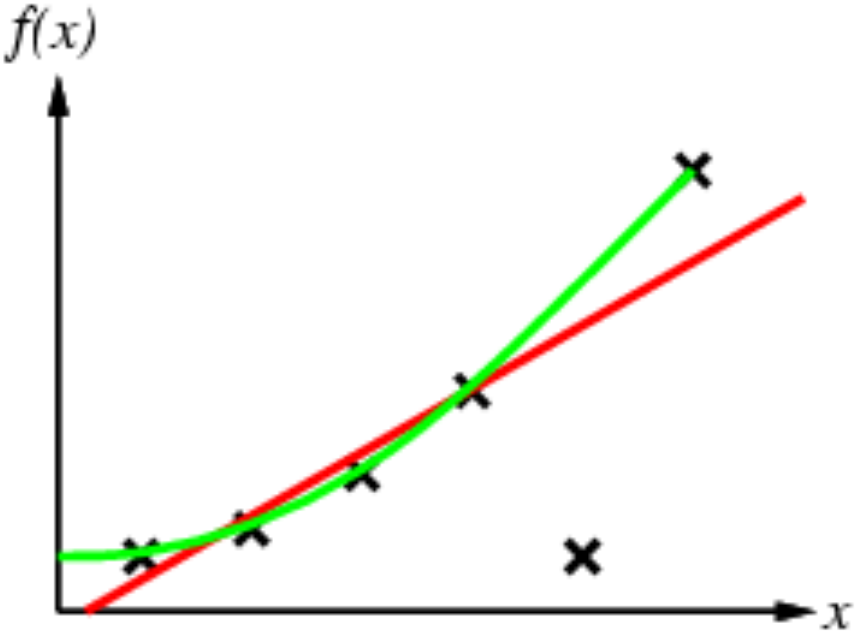
Example 2

- Another example. We have the following training sample (set of points)



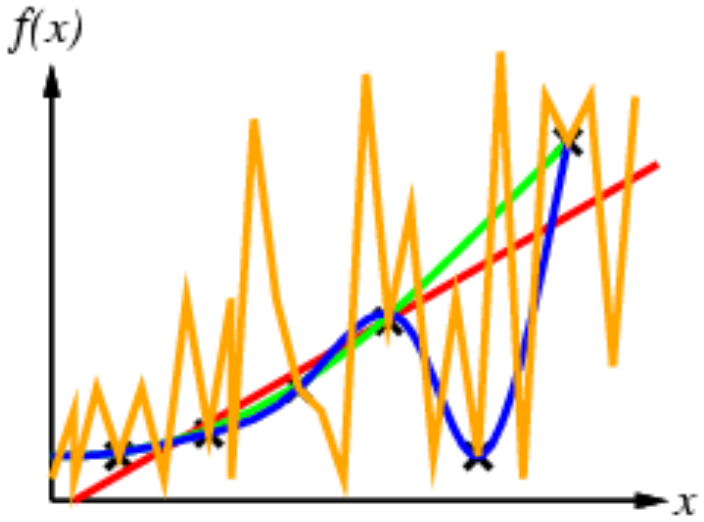
Example 2

- Which *h* function is the best?
  1. A line?
  2. A polygon of degree 2?
  3. More complex formulation?



Example 2

- The various possibilities …



- The fourth point could be an outlier!
- Tradeoff: complexity of *h* vs. degree of fit to the data

Overview

- Introduction to ML & DM
- Examples
- Terminology: Concept, Example, Attribute
- Search & Bonferroni's principle
- Curve fitting, a more complex example
- **Table Look-up**

Table Look-up

- “Learning” by storing every thing, trivial memorization of experiences

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	true	no
overcast	mild	normal	true	yes
sunny	mild	high	false	no
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Table Look-up

- Each training example *x* is simply stored.
- When a query example x_q is given
  retrieve the stored example with the same attributes values (x_i) and simply return $c(x_i)$

Outlook	Temperature	Humidity	Windy	Play
rainy	cool	normal	true	?

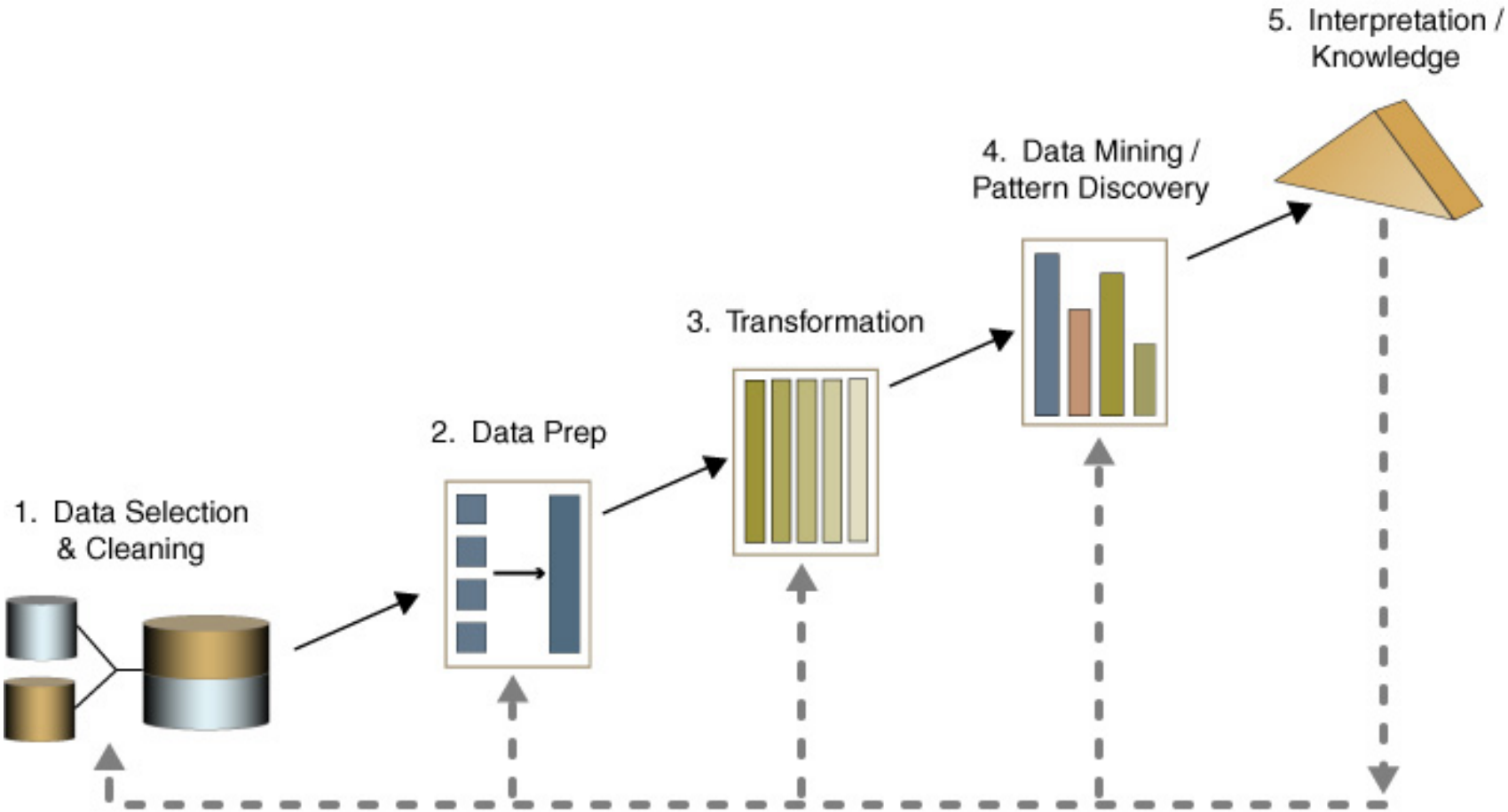
- In this case, return "no" (5th example)

Table Look-up

- Such scheme does not really “learn”
  (no inference or induction)
- Can be very effective (training sample = test sample)
- Does not respect our definition of learning (“improve over time”).
- Use as baseline / dummy approach

Conclusion: Real Data

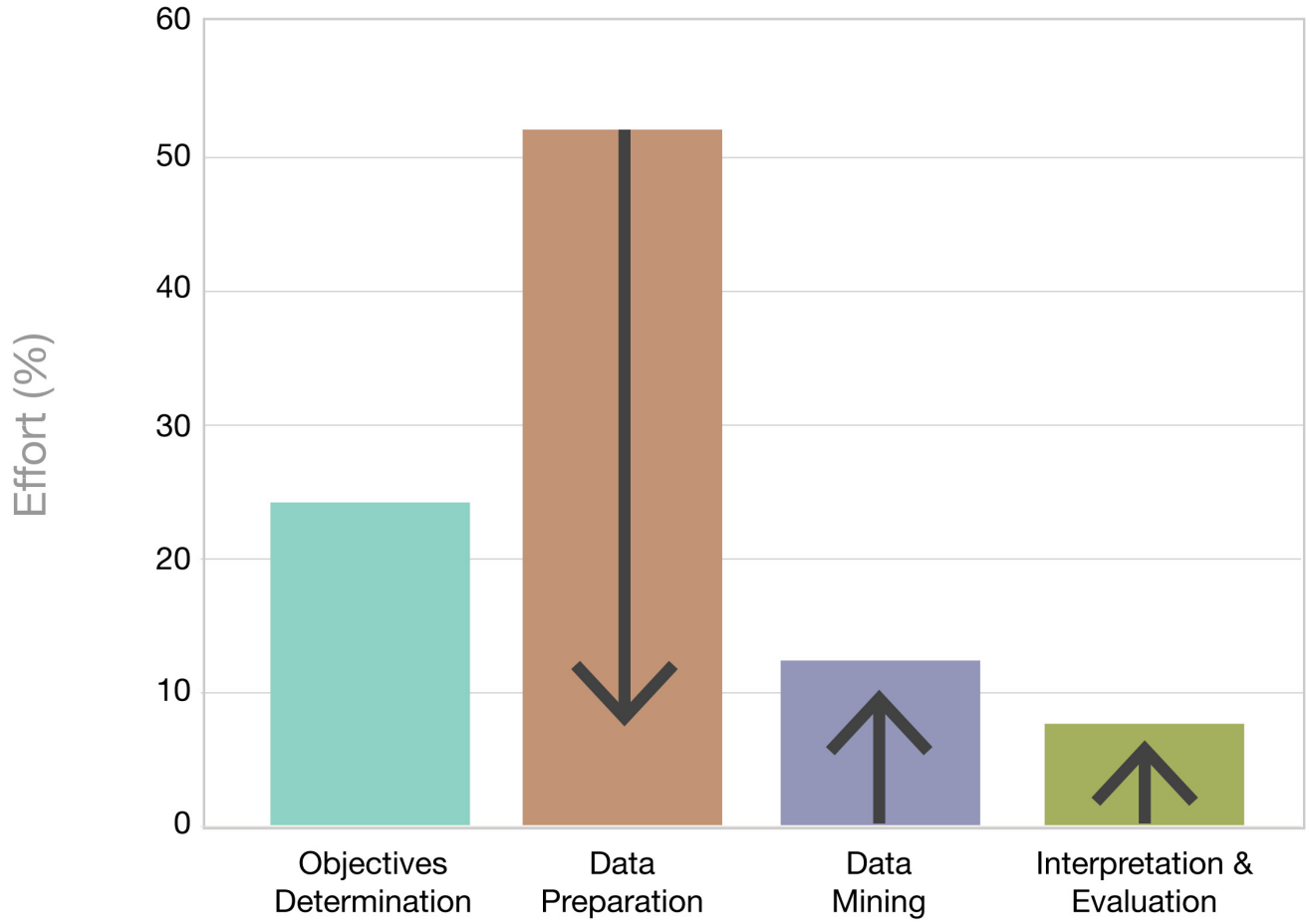
A whole process is often hidden



An Overview of the Steps That Compose the KDD Process

Conclusion: Real Data

Arrows indicate the direction we hope the effort should go



Conclusion: Preparing the Data

- Denormalization is not the only issue
- Problem: different data sources (e.g. sales department, customer billing department, …)
  - Differences: styles of record keeping, conventions, time periods, data aggregation, primary keys, errors
  - Data must be assembled, integrated, cleaned up
  - "Data warehouse": consistent point of access
- External data may be required ("overlay data")
- Critical: type and level of data aggregation
- Telecom companies and banks are really concerned with their data… les other companies

Conclusion

- What is Data Mining, Machine Learning
- The Input: concept, example, attribute
- The Output: explain, predict
- DB, Statistics, ML
- Always consider the complexity of your learning structures and algorithms. Simple could be beautiful!
- Over-fitting
- Do not consider that the training sample is error-free (error in the data)
- Don't forget to compare with the simple "Table look-up" strategy (error rate = 0%)

Conclusion

- Data Mining, Machine Learning & Big Data
- A solution ot extract information from data
- Big data means
  - Velocity of the date
  - Variaty of the data
  - Volume of the data
  - Veracity of the data (quality)
- Detecting fraud
- Sharing data

Data Mining & Ethics

- Ethical issues arise in practical applications
- Data mining often used to discriminate
  e.g. loan applications: using some information (e.g. sex, religion, race) is unethical
- Ethical situation depends on application
  e.g., same information ok in medical application
- Attributes may contain problematic information
  e.g., area code may correlate with race
- Important questions:
  Who is permitted access to the data?
  For what purpose was the data collected?
  What kind of conclusions can be legitimately draw from it?

References

- Ian H. Witten, Eibe Frank: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.
- Tom Mitchell: Machine Learning. McGraw Hill, 1997.
- Christopher M. Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
- T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning. Springer, 2nd Ed., 2009

- Max Bramer: Principles of Data Mining. Springer, 2007.
- S. Marsland: Machine Learning. An Algorithmic Prespective. CRC Press, 2009.
- Soumen Chakrabarti et al.: Data Mining, Know it all. Elsevier, Amsterdam, 2009.

Defining the Coefficients

- How can we find the values for the coefficient w_i?
- Use the error function

$$Err(\mathbf{w}) = \frac{1}{2} \cdot \sum_{n=1}^{N} (y(x_n, \mathbf{w}) - t_n)^2$$

- Derive it in function of w_i and set that derivation = 0.

$$\frac{\partial Err(\mathbf{w})}{\partial w_i} = 0$$

$$Err(\mathbf{w})' = \frac{2}{2} \cdot \sum_{n=1}^{N} (y(x_n, \mathbf{w}) - t_n) \cdot y(x_n, \mathbf{w})'$$

Defining the Coefficients

- If we use the simplest model (M = 0, an horizontal line), the general model is simplified to

$$\begin{aligned} y(x, \mathbf{w}) &= w_0 + w_1 \cdot x + w_2 \cdot x^2 + \cdots + w_M \cdot x^M \\ y(x, \mathbf{w}) &= w_0 \end{aligned}$$

- and the unique coefficient w_0 is

$$\frac{\partial Err(\mathbf{w})}{\partial w_0} = 0 = \frac{2}{2} \cdot \sum_{n=1}^{N} (w_0 - t_n) \cdot 1$$

$$N \cdot w_0 = \sum_{n=1}^{N} t_n \rightarrow w_0 = \bar{t}_n$$

Defining the Coefficients

- If we use the simplest model (M=1, a line), the model is

$$y(x, \mathbf{w}) = w_0 + w_1 \cdot x$$

- and the coefficient w_0 and w_1 are

$$\frac{\partial Err(\mathbf{w})}{\partial w_0} = 0 = \sum_{n=1}^{N} ((w_0 + w_1 \cdot x_n) - t_n)$$

$$\frac{\partial Err(\mathbf{w})}{\partial w_1} = 0 = \sum_{n=1}^{N} ((w_0 + w_1 \cdot x_n) - t_n) \cdot w_1$$

Defining the Coefficients

- The coefficient w_0 and w_1 are

$$
\begin{aligned}
w_0 &= \bar{t} - w_1 \cdot \bar{x} \\
w_1 &= \frac{N \cdot \sum_{n=1}^{N} (x_n \cdot t_n) - \sum_{n=1}^{N} x_n \cdot \sum_{n=1}^{N} t_n}{N \cdot \sum_{n=1}^{N} x_n^2 - \left(\sum_{n=1}^{N} x_n\right)^2}
\end{aligned}
$$

Defining the Coefficients

- In general, we have the following set of equations

$$\sum_{j=0}^{M} A_{ij} \cdot w_j = T_i$$

$$\text{where} \quad A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j} \quad \text{and} \quad T_i = \sum_{n=1}^{N} (x_n)^i \cdot t_n$$