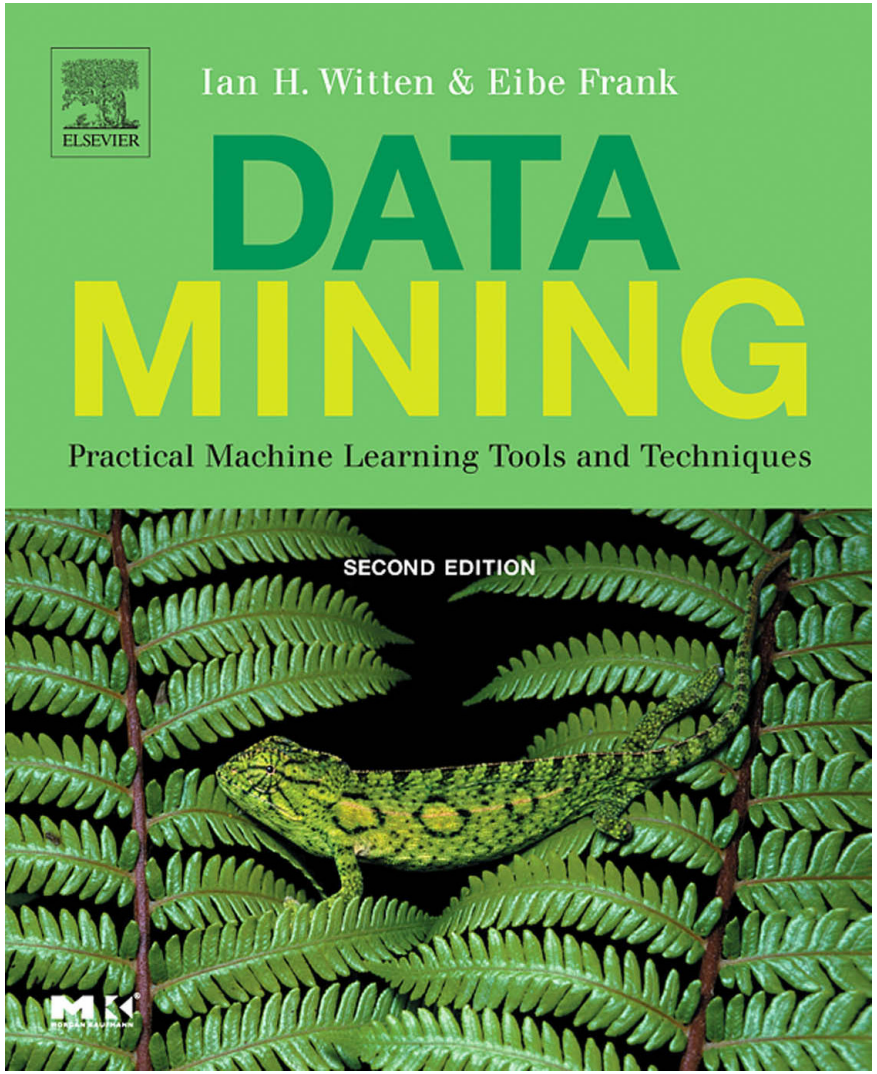
Data Mining



Textbook:

DATA MINING:
Practical Machine Learning Tools and Techniques, 2nd Edition,
by Ian H. Witten and Eibe Frank, Morgan Kaufmann Publishers, 2005.

Chapter 1: What's it all about?

1.1 Data mining and machine learning

Data vs. Information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
- Potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it.
  - Data: recorded facts
  - Information: patterns underlying the data

Examples

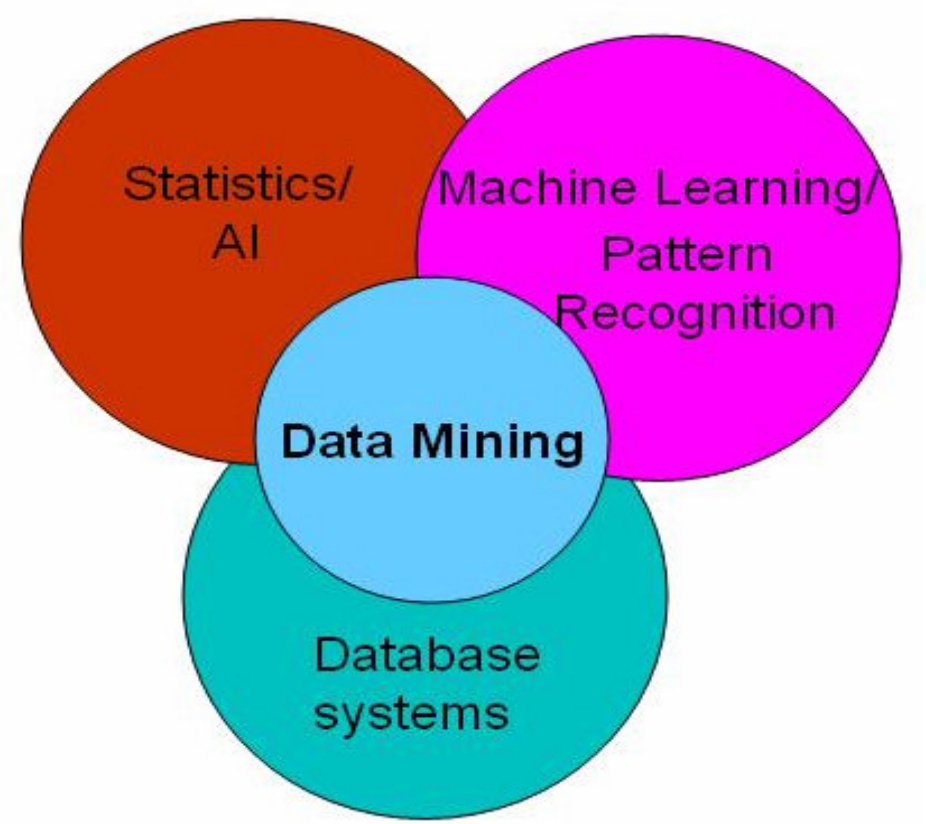
- Cow culling
  - Given: cows described by many features
  - Problem: selection of cows that should be culled
  - Data: historical records and farmers' decisions
- Hunters seek patterns in animal migration behavior
- Farmers seek patterns in crop growth
- Politicians seek patterns in voter opinion

Data Mining

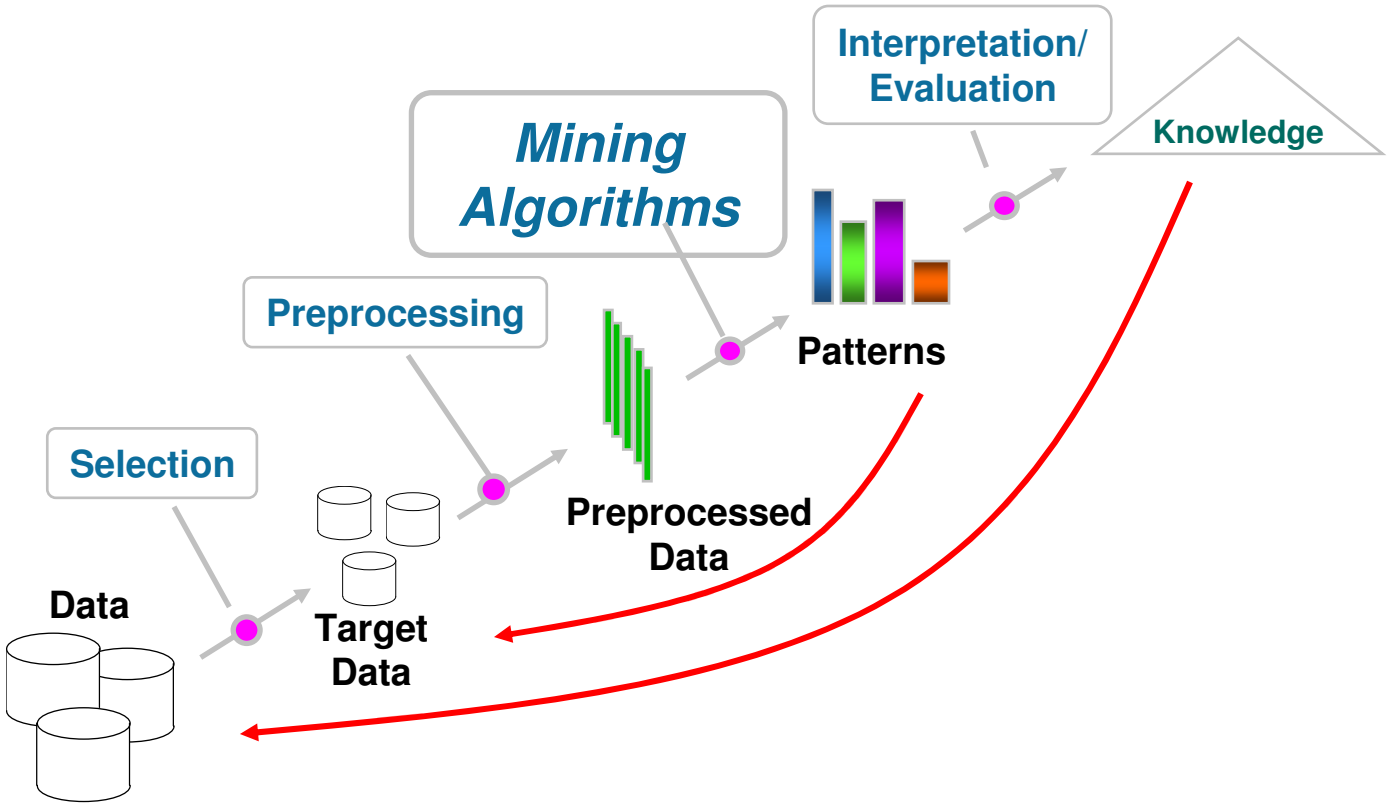
- Extracting
  - implicit,
  - previously unknown,
  - potentially useful

  information from data
- Needed: programs that detect patterns and regularities in the data
- ***Data Mining*** is the process of discovering useful patterns, automatically or semiautomatically, in large quantities of data.

Origins of data mining



Data Mining: An Engineering Process



Data mining: interactive and iterative process.

Machine learning techniques

- *Algorithms for acquiring structural descriptions from examples*
- Structural descriptions represent patterns
- Explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived
- Methods originate from statistics and artificial intelligence

Structural descriptions: The contact lens data

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

Example: if then rules

```
If tear production rate = reduced then recommendation = none
Otherwise, if age = young and astigmatic = no
                then recommendation = soft
```

Can machines really learn?

- Definitions of “learning” from dictionary:
  - To get knowledge of by study, experience, or being taught
  - To become aware by information or from observation
  - To commit to memory
  - To be informed of
  - To receive instruction
- Operational definition:
  Things learn when they change their behavior in a way that makes them perform better in the future.

1.2 Simple examples

The weather problem

- Instances in a dataset are characterized by the values of features, or *attributes.*
- In this case there are four attributes: *outlook, temperature, humidity,* and *windy.*
- The four attributes have values that are symbolic categories rather than numbers.
  - *Outlook* can be *sunny, overcast,* or *rainy*
  - *Temperature* can be *hot, mild,* or *cool*
  - Humidity can be *high* or *normal*
  - Windy can be *true* or *false*

Dataset: The weather data

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

Conditions for playing

- Rules:

```
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 humidity = normal                   then play = yes
If none of the above                   then play = yes
```

- A set of rules that are intended to be interpreted in sequence is called a ***decision list***.
- Some of the rules are incorrect if they are taken individually. For example, the rule if humidity = normal then play = yes

Weather data with some numeric attributes

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

Classification vs. association rules

- ***Classification rule:*** predicts value of a given attribute (the classification of an example)

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

- ***Association rule:*** predicts value of arbitrary attribute (or combination)

```
If temperature = cool                    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.
```

Mixed-attribute problem

- The problem with numeric attributes is called a ***numeric-attribute problem***
- This case is a ***mixed-attribute problem*** because not all attributes are numeric.
- The first rule given might take the following form:

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

Contact lenses data

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

A complete and correct rule set

```
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
```

Figure 1.1 Rules for the contact lens data.

A decision tree for this problem

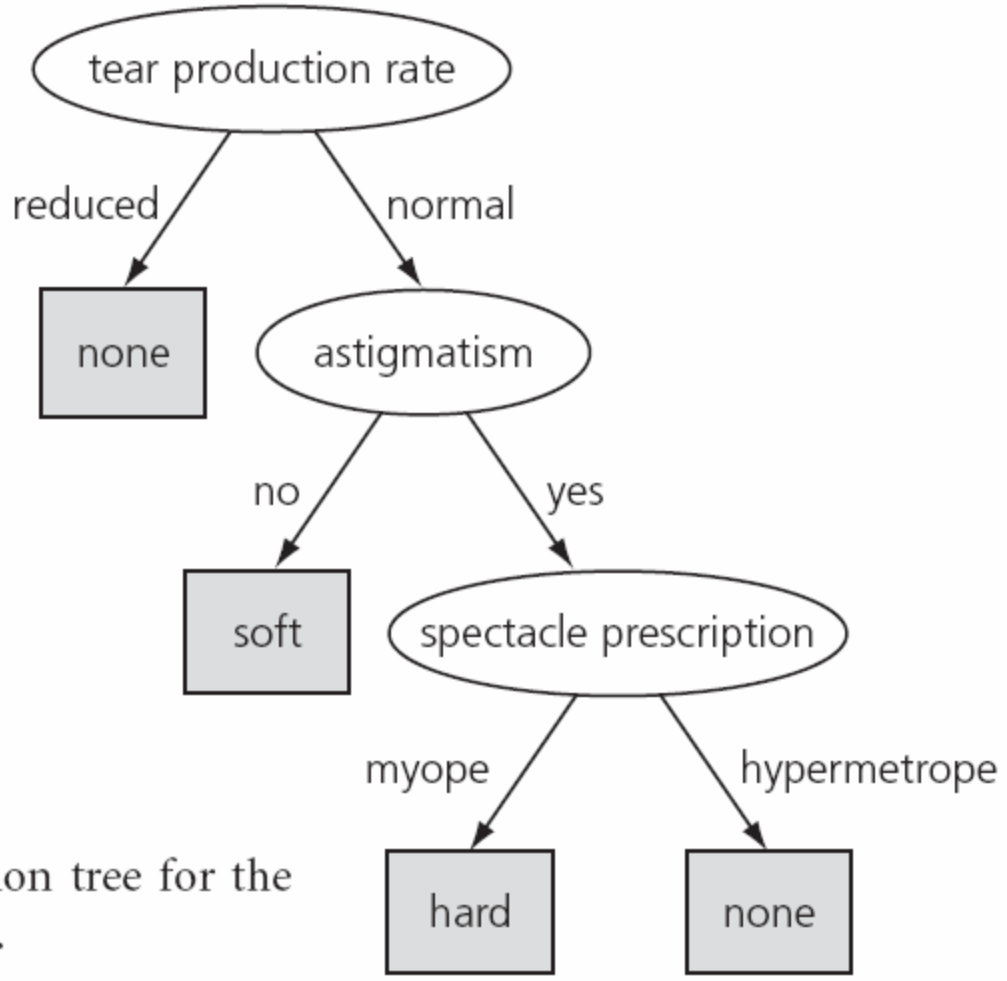


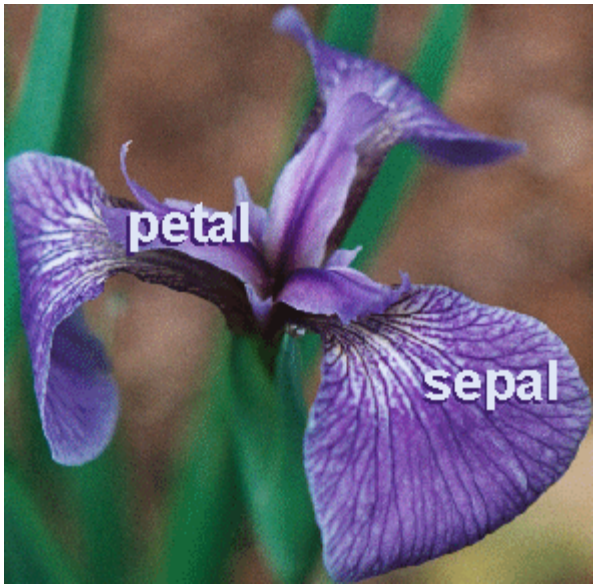
Figure 1.2 Decision tree for the contact lens data.

Iris flowers Classifying problem

- This dataset back to work by R.A. Fisher in the mid-1930s
- The iris dataset is the most famous dataset used in data mining
- It contains 50 examples each of three types of plant:
  Iris setosa, Iris versicolor, and *Iris virginica.*

Attributes

- There are four attributes: *sepal length*, *sepal width*, *petal length*, and *petal width* (all measured in centimeters)



Iris Dataset

	Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
. . .					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
. . .					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
. . .					

Classifying iris flowers

```
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 width < 1.35 then Iris versicolor
If petal length ≥ 2.45 and petal length < 4.45 then Iris versicolor
If sepal width < 2.55 and petal length < 4.95 and
   petal width < 1.55 then Iris versicolor
If petal length ≥ 2.45 and petal length < 4.95 and
   petal width < 1.55 then Iris versicolor
If sepal length ≥ 6.55 and petal length < 5.05 then Iris versicolor
If sepal width < 2.75 and petal width < 1.65 and
   sepal length < 6.05 then Iris versicolor
If sepal length ≥ 5.85 and sepal length < 5.95 and
   petal length < 4.85 then Iris versicolor
If petal length ≥ 5.15 then Iris virginica
If petal width ≥ 1.85 then Iris virginica
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
```

CPU performance Problem

- The iris dataset involves numeric attributes, the outcome—the type of iris—is a category
- This problem shows some data for which the outcome and the attributes are numeric.
- It concerns the relative performance of computer processing power on the basis of a number of relevant attributes

Linear regression function

- Attribute Information:
  - MYCT: machine cycle time in nanoseconds (integer)
  - MMIN: minimum main memory in kilobytes (integer)
  - MMAX: maximum main memory in kilobytes (integer)
  - CACH: cache memory in kilobytes (integer)
  - CHMIN: minimum channels in units (integer)
  - CHMAX: maximum channels in units (integer)
  - PRP: published relative performance

The CPU performance data

	Cycle time (ns) MYCT	Main memory (KB) Min. MMIN	Main memory (KB) Max. MMAX	Cache (KB) CACH	Channels Min. CHMIN	Channels Max. CHMAX	Performance PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
3	29	8000	32000	32	8	32	220
4	29	8000	32000	32	8	32	172
5	29	8000	16000	32	8	16	132
. . .							
207	125	2000	8000	0	2	14	52
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Linear regression function

- *regression equation*:

$$\begin{aligned}\text{PRP} = &-55.9 + 0.0489\ \text{MYCT} + 0.0153\ \text{MMIN} + 0.0056\ \text{MMAX} \\ &+ 0.6410\ \text{CACH} - 0.2700\ \text{CHMIN} + 1.480\ \text{CHMAX}.\end{aligned}$$

- The process of determining the weights is called *regression*
- We will examine different representations that can be used for predicting numeric quantities.
- Practical situations frequently present a mixture of numeric and nonnumeric attributes.

Labor negotiations

- The labor negotiations dataset is summarized the outcome of Canadian contract negotiations in 1987 and 1988.
- It includes agreements reached for organizations with at least 500 members (teachers, nurses, university staff, police, etc.).
- Each case concerns one contract, and the outcome is whether the contract is deemed ***acceptable*** or ***unacceptable***.

Data from labor negotiations

Attribute	Type	1	2	3	...	40
duration	years	1	2	3		2
wage increase 1st year	percentage	2%	4%	4.3%		4.5
wage increase 2nd year	percentage	?	5%	4.4%		4.0
wage increase 3rd year	percentage	?	?	?		?
cost of living adjustment	{none, tcf, tc}	none	tcf	?		none
working hours per week	hours	28	35	38		40
pension	{none, ret-allw, empl-cntr}	none	?	?		?
standby pay	percentage	?	13%	?		?
shift-work supplement	percentage	?	5%	4%		4
education allowance	{yes, no}	yes	?	?		?
statutory holidays	days	11	15	12		12
vacation	{below-avg, avg, gen}	avg	gen	gen		avg
long-term disability assistance	{yes, no}	no	?	?		yes
dental plan contribution	{none, half, full}	none	?	full		full
bereavement assistance	{yes, no}	no	?	?		yes
health plan contribution	{none, half, full}	none	?	full		half
acceptability of contract	{good, bad}	bad	good	good		good

Decision tree for the labor data (a)

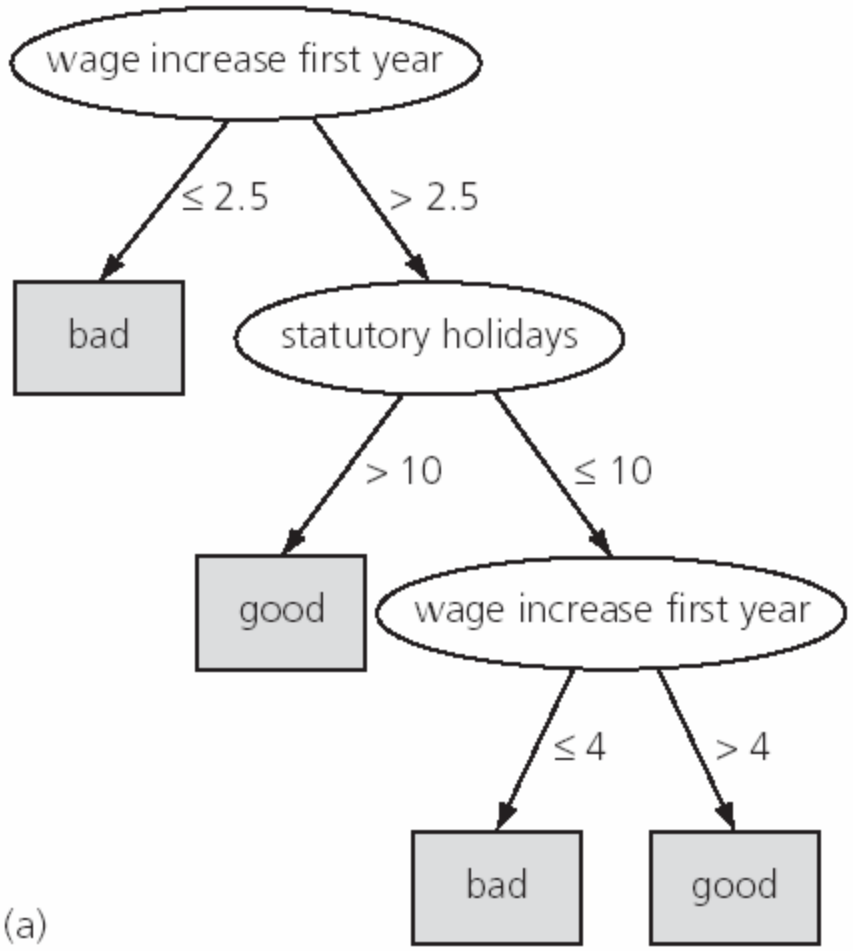
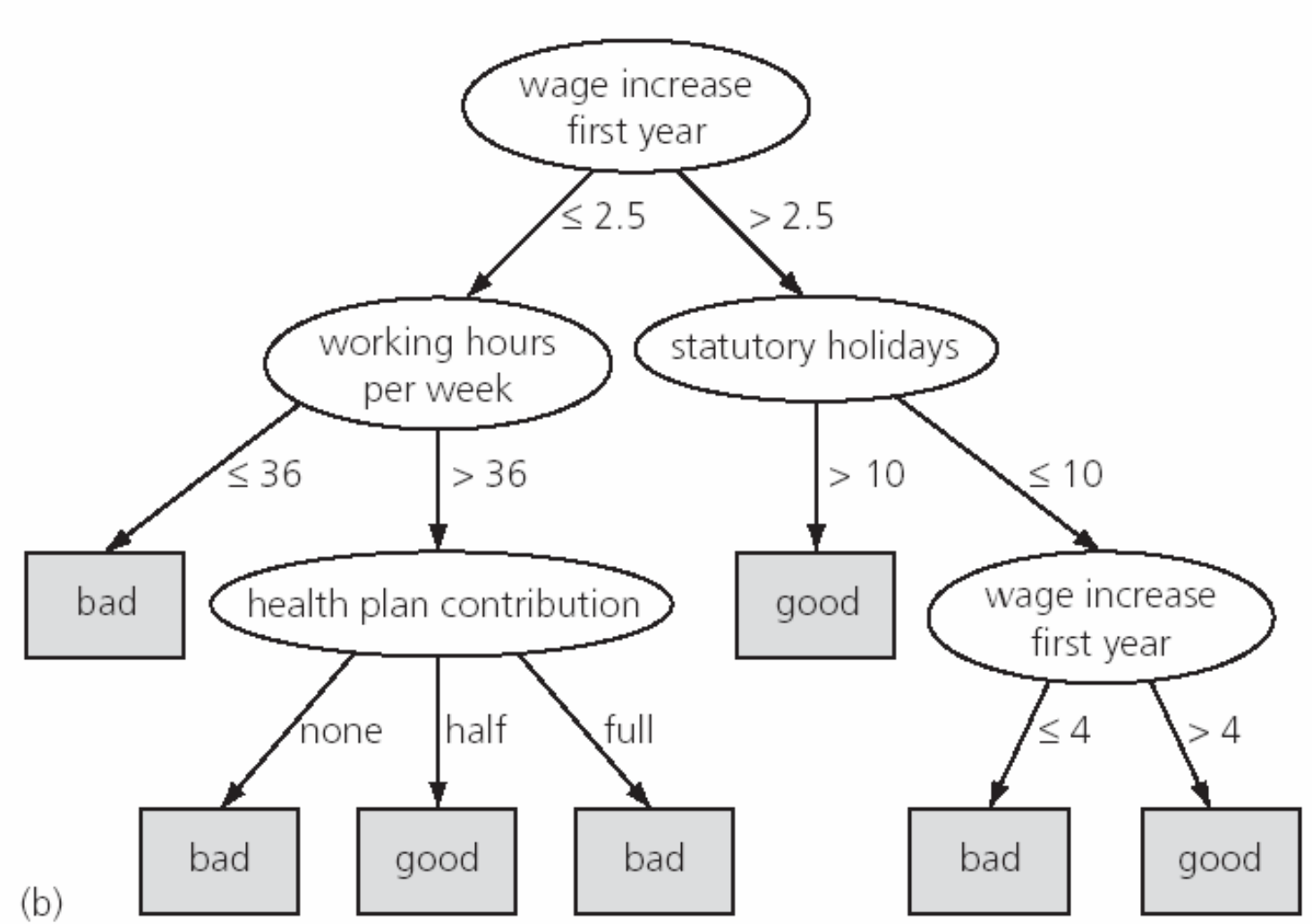


Figure 1.3 Decision trees for the labor negotiations data.

Decision tree for the labor data (b)



Soybean classification

- Identification of rules for diagnosing soybean diseases
- The data is taken from questionnaires describing plant diseases
- There are about 680 examples
- Plants were measured on 35 attributes
- There are 19 disease categories altogether

The soybean data

	Attribute	Number of values	Sample value
Environment	time of occurrence	7	July
	precipitation	3	above normal
....			
Seed	condition	2	normal
	mold growth	2	absent
....			
Fruit	condition of fruit pods	3	normal
	fruit spots	5	—
Leaf	condition	2	abnormal
	leaf spot size	3	—
....			
Stem	condition	2	abnormal
	stem lodging	2	yes
....			
Root	condition	3	normal
Diagnosis		19	diaporthe stem canker

The role of domain knowledge

```
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
```

1.3 Fielded applications

Processing loan applications

- Given: questionnaire with financial and personal information
- Question: should money be lent?
- Statistical methods are used to determine clear “accept” and “reject” cases
- Statistical method covers 90% of cases
- Borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?

Enter machine learning

- 1000 training examples of borderline cases for which a loan had been made
- 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
- Rules could be used to explain decisions to customers

Screening images

- Given: radar satellite images of coastal waters
- Problem: detect oil slicks in those images
- Oil slicks appear as dark regions
- Not easy: look-alike dark regions can be caused by weather conditions (e.g. high wind)
- Expensive process requiring highly trained personnel

Enter machine learning

- Input Attributes:
  - size of region
  - shape
  - area
  - intensity
  - sharpness and jaggedness of boundaries
  - proximity of other regions
- Output:
  - Extract dark regions
- Some constraints:
  - Few training examples—oil slicks are rare!
  - Unbalanced data: most dark regions aren't slicks

Load forecasting

- Electricity supply companies need forecast of future demand for power
- Forecasts of min/max load for each hour
- Given: constructed load model using over the previous 15 years
- Static model consist of:
  - base load for the year
  - load periodicity over the year
  - effect of holidays
- It assumes “normal” climatic conditions
- Problem: adjust for weather conditions

Enter machine learning

- Prediction corrected using “most similar” days
- Attributes:
  - temperature
  - humidity
  - wind speed
  - cloud cover readings
  - plus difference between actual load and predicted load
- Average difference among eight “most similar” days added to static model
- Linear regression coefficients form attribute weights in similarity function

Marketing and sales

- 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)
  - Special offers: identifying profitable customers and detecting their patterns of behavior that could benefit from new services (e.g. phone companies)

Marketing and sales (II)

- Market basket analysis
  - Association techniques find groups of items that tend to occur together in a transaction
  - e.g. used to analyze supermarket checkout data may uncover the fact that on Thursdays, customers who buy diapers also buy chips
- Identifying prospective customers
  - Focusing promotional mail outs (targeted campaigns are cheaper than mass-marketed ones)

1.4 Machine Learning and Statistics

Machine learning and statistics

- Historical difference (grossly oversimplified):
  - Statistics: testing hypotheses
  - Machine learning: finding the right hypothesis
- But: huge overlap
  - Decision trees (Breiman et al. 1984 as a statisticians & J. Ross Quinlan as a ML researcher)
- Today: perspectives have converged
  - Most ML algorithms employ statistical techniques

1.5 Data mining and ethics

Data mining and 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

Data mining and ethics (II)

- Important questions:
  - Who is permitted access to the data?
  - For what purpose was the data collected?
  - What kind of conclusions can be legitimately drawn from it?
- Are resources put to good use?

The end of

Chapter 1: What's it all about?