CHAPTER 1

What’s It All About?

Outline

- Data mining and machine learning
- Simple examples
- fielded applications
- Data mining and ethics

Data mining and machine learning (1/3)

- Data mining
  - The process of discovering patterns, automatically or semiautomatically, in large quantities of data—and the patterns must be useful
  - People frequently use data mining to gain knowledge, not just predictions
- Machine learning
  - Most of techniques for finding and describing structural patterns in data

Data mining and machine learning (2/3)

- Describing structural patterns
  - Rules
  - Decision trees
  - Association rules
  - Regression function
  - Networks
  - ……

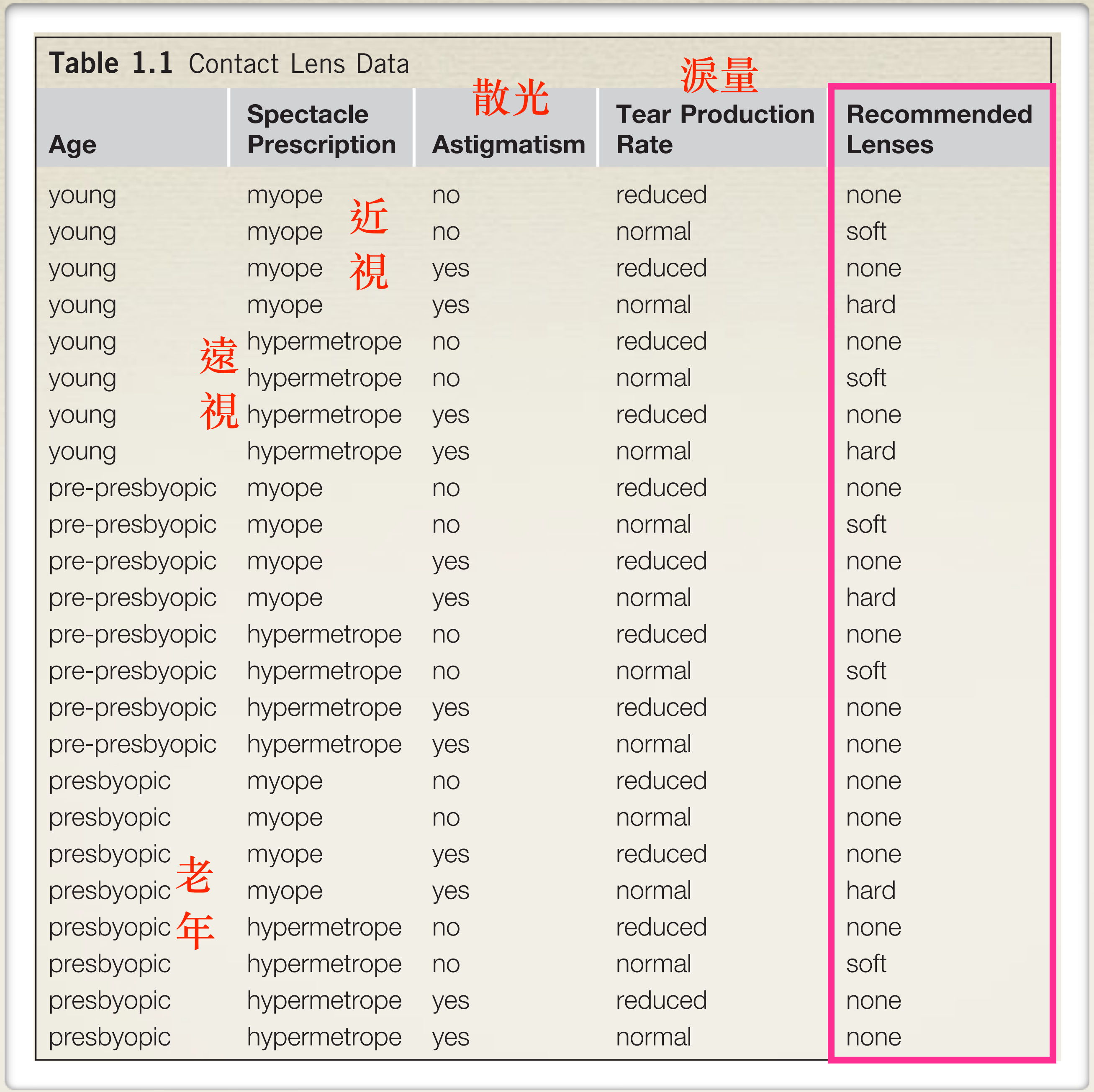


Table 1.1 Contact Lens Data

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

散光

淚量

近視

遠視

中年

老年

nominal or categorical

All combinations of possible values (not always)

IF tear-production-rate=reduced THEN recommended-lenses=none (12/12)
ELSEIF age=young and astigmatism=no THEN recommended-lenses=soft (2/2)

Simple examples: weather (1/7)

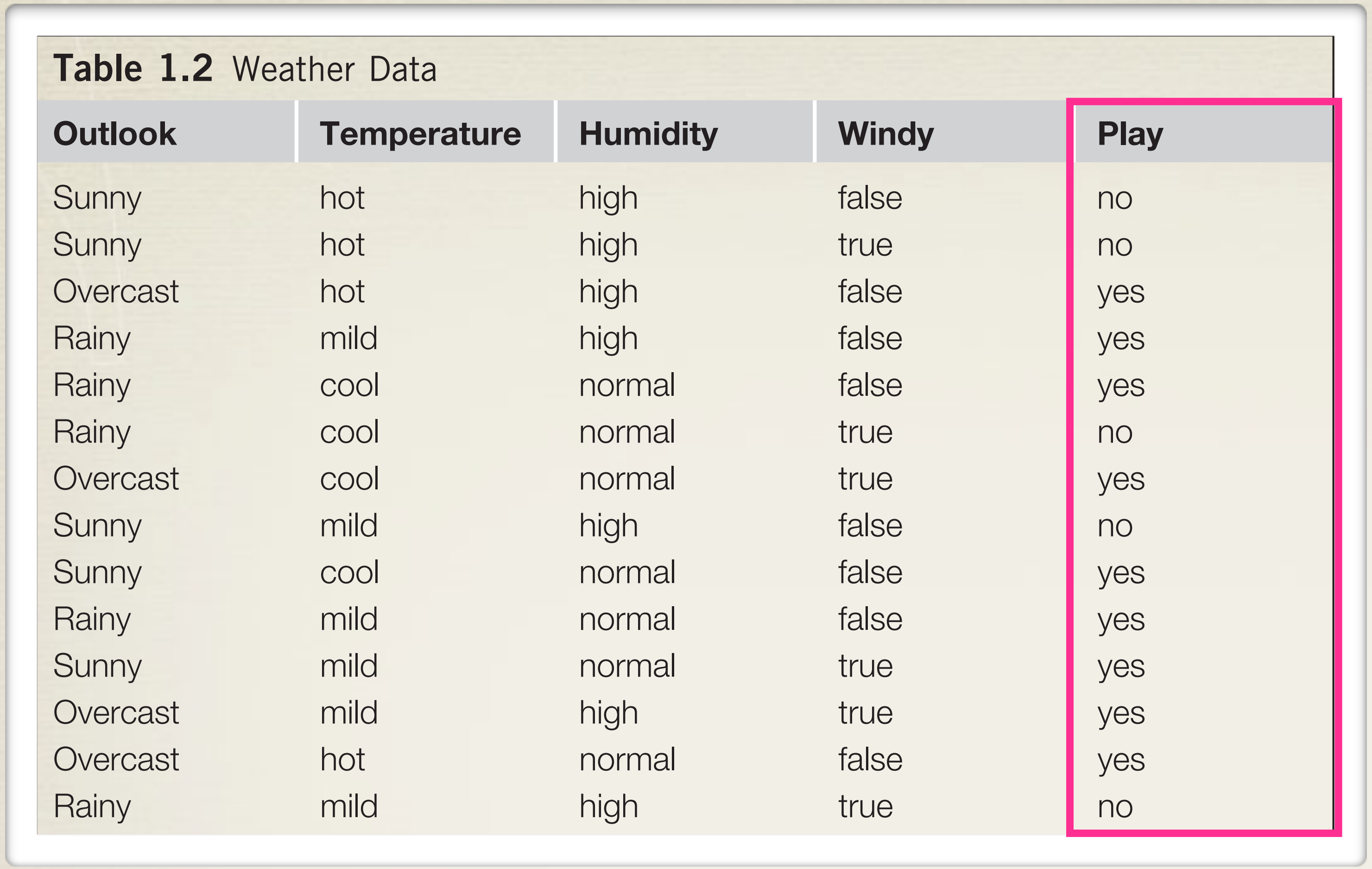


Table 1.2 Weather Data

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

nominal or categorical

Classification Rule

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

decision list
interpreted in sequence

Association Rule

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

Simple examples: weather (2/7)

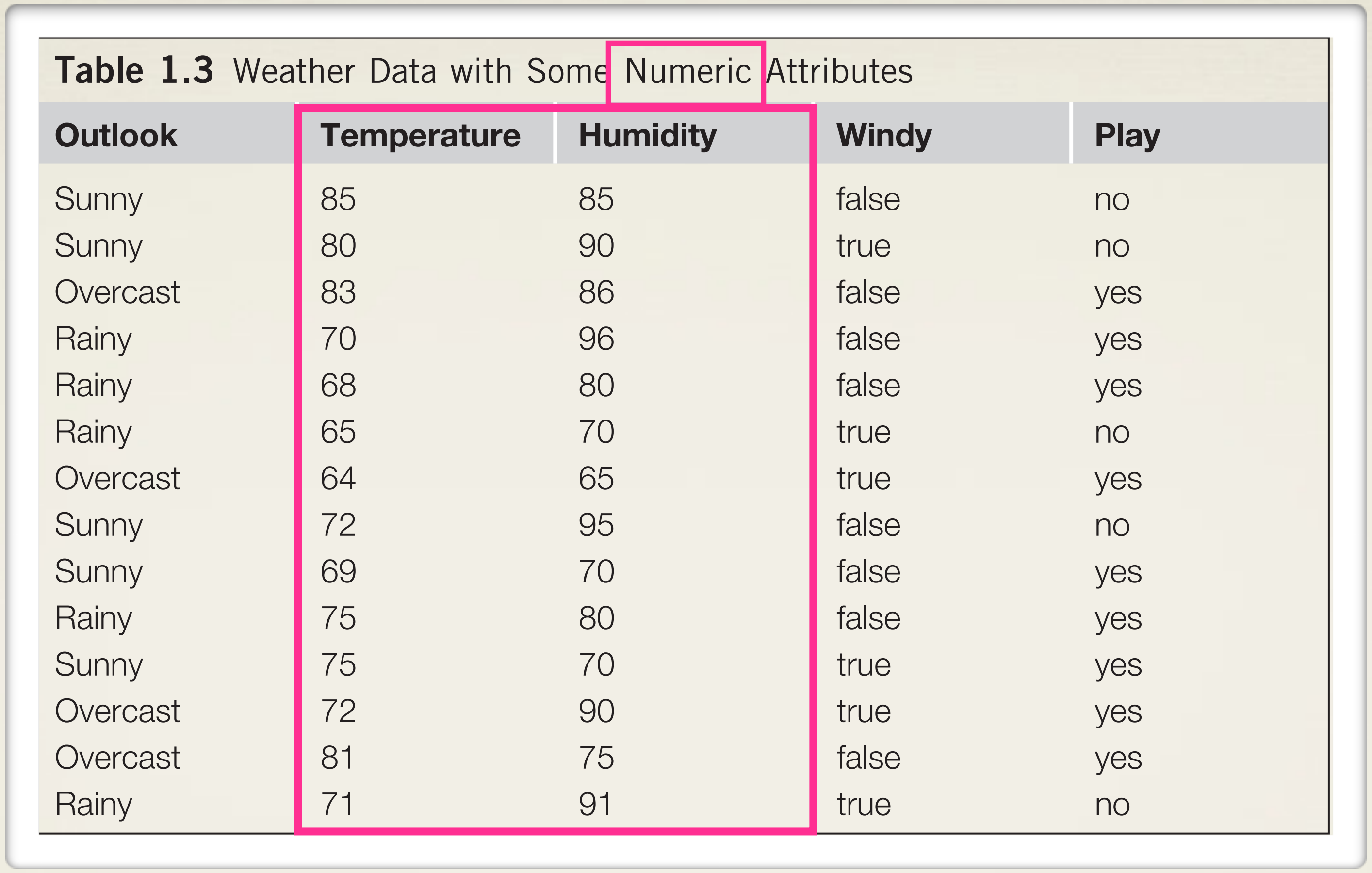


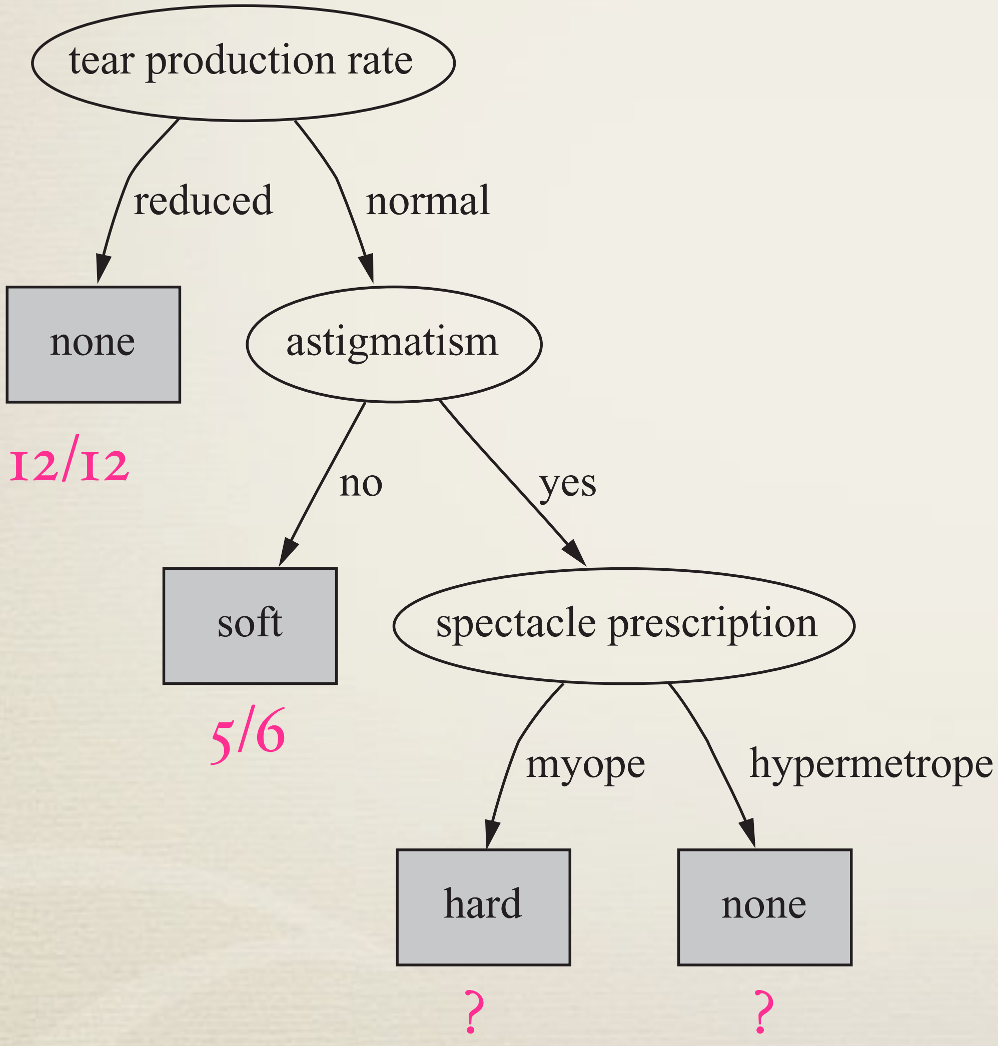
Table 1.3 Weather Data with Some Numeric Attributes

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

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

Simple examples: contact lens (3/7)

Decision tree



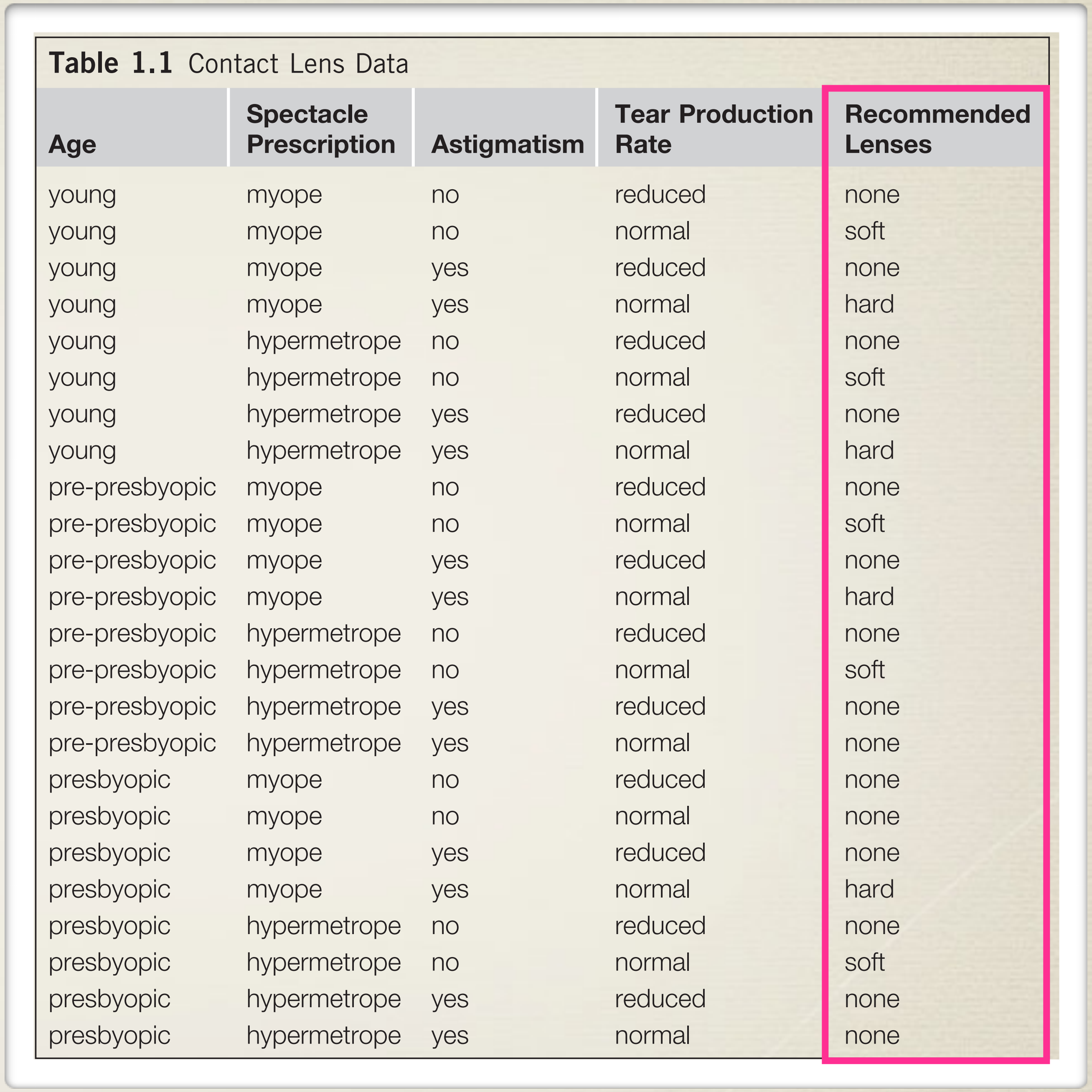


Table 1.1 Contact Lens Data

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

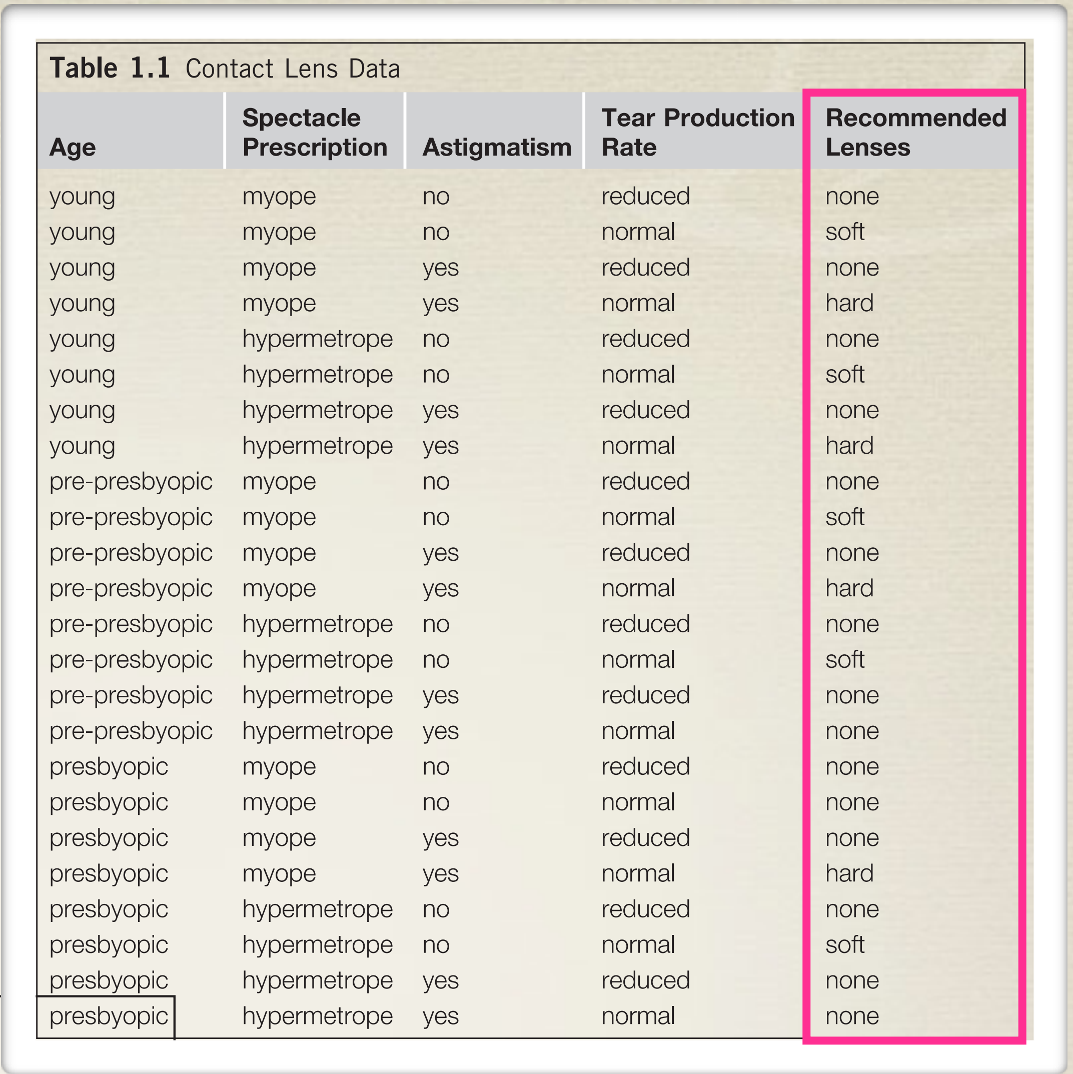
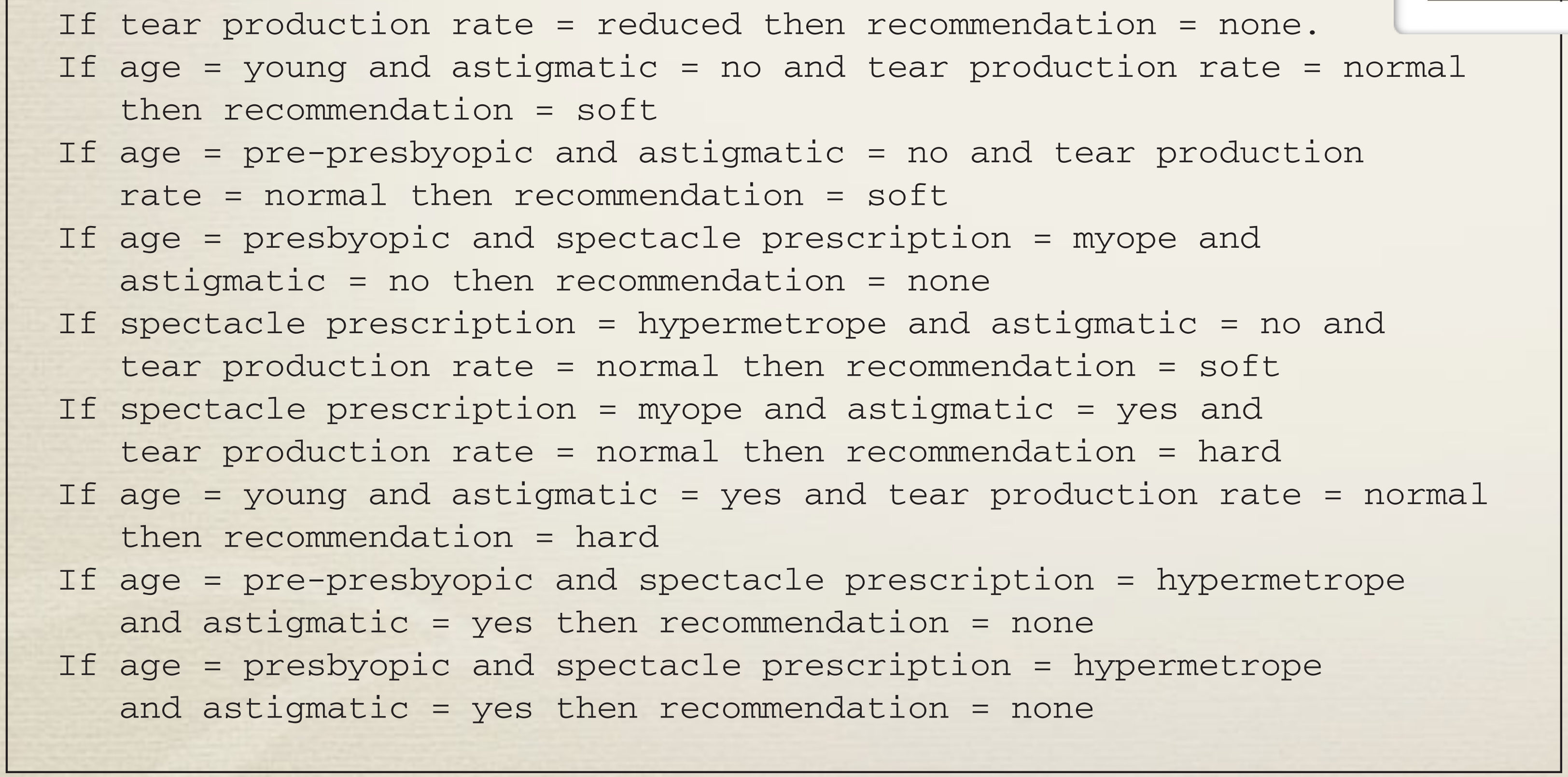


Table 1.1 Contact Lens Data

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

Rules



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

Simple examples: iris (5/7)

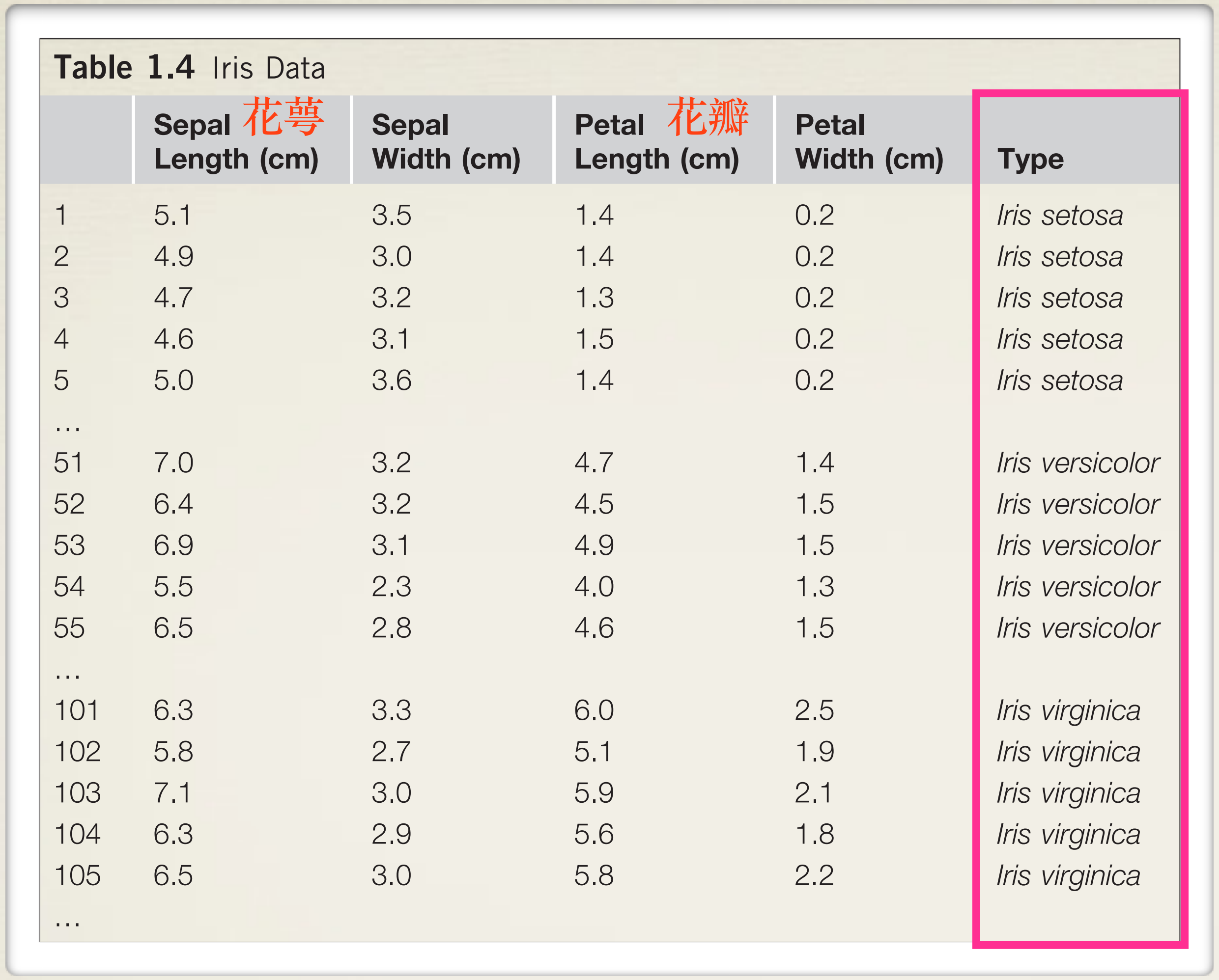


Table 1.4 Iris Data

	Sepal 花萼 Length (cm)	Sepal Width (cm)	Petal 花瓣 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*
…					

50 examples for each

Rules

```
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-length ≥ 5.85 and petal-length < 4.75 then Iris-versicolor
If  sepal-width < 2.55 and petal-length < 4.95 and
    petal-width < 1.55 then Iris-versicolor
```

Simple examples: CPU performance (6/7)

Numeric prediction

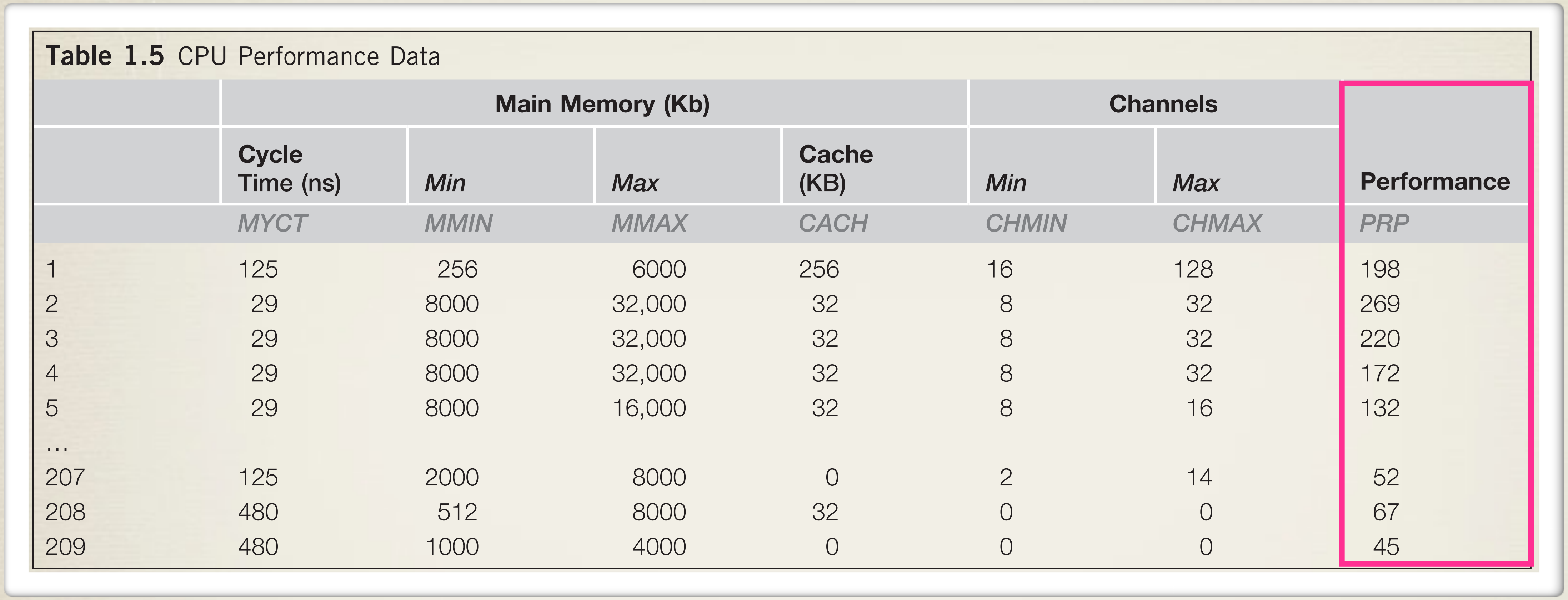


Table 1.5 CPU Performance Data

	Main Memory (Kb)				Channels		
	Cycle Time (ns)	*Min*	*Max*	Cache (KB)	*Min*	*Max*	Performance
	MYCT	*MMIN*	*MMAX*	*CACH*	*CHMIN*	*CHMAX*	*PRP*
1	125	256	6000	256	16	128	198
2	29	8000	32,000	32	8	32	269
3	29	8000	32,000	32	8	32	220
4	29	8000	32,000	32	8	32	172
5	29	8000	16,000	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

Regression equation

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

Simple examples: labor negotiations (7/7)

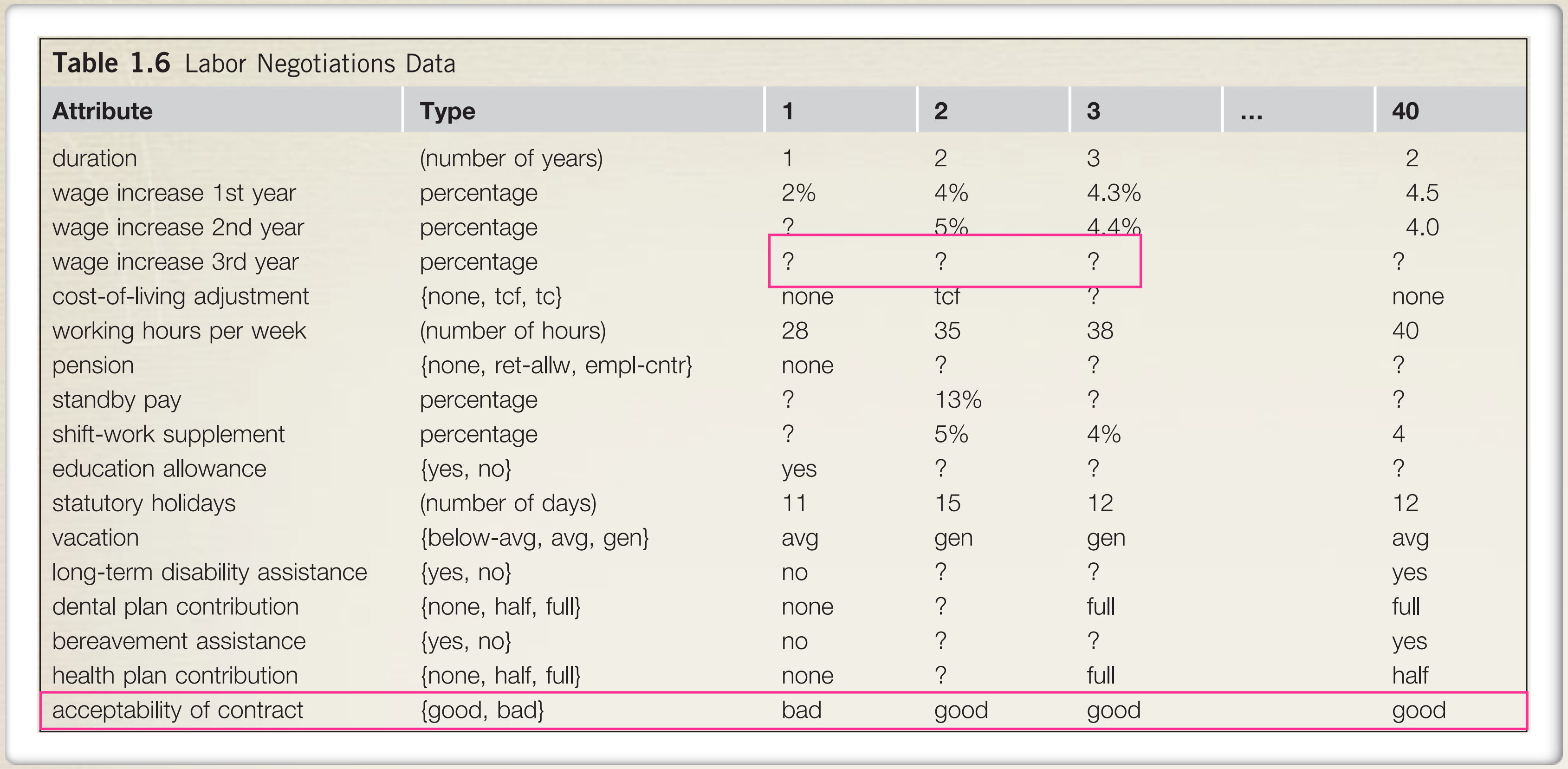
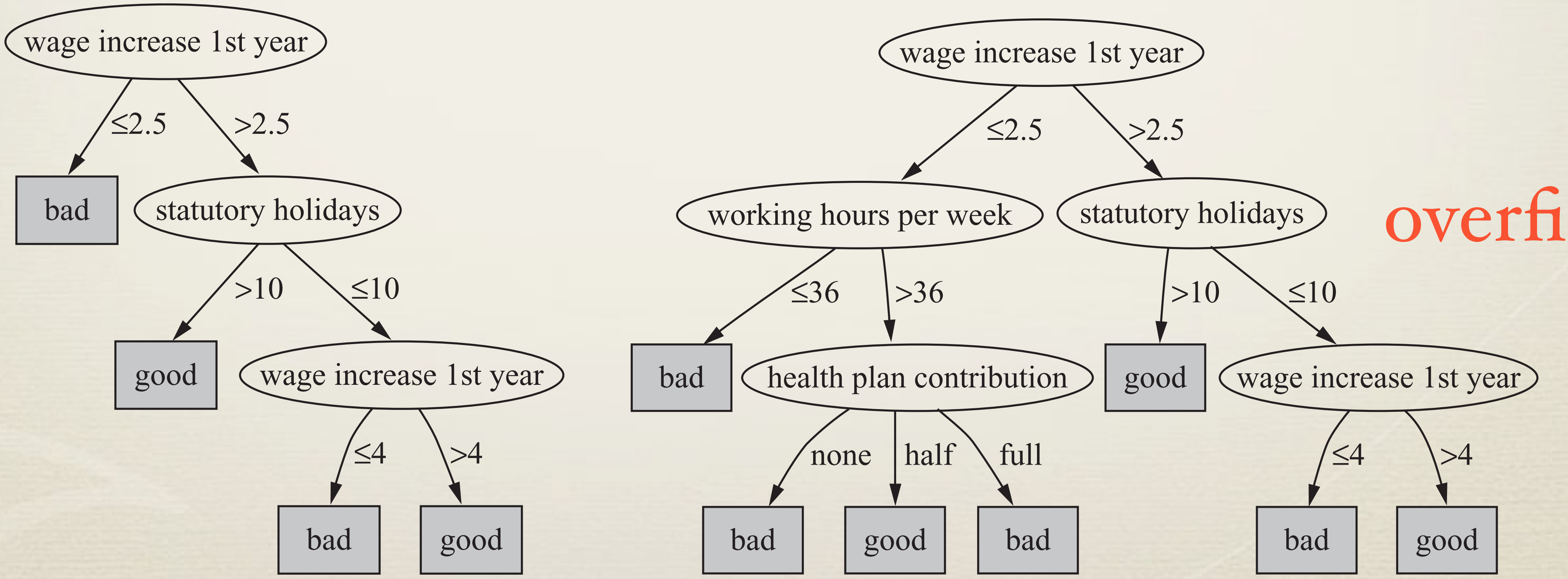


Table 1.6 Labor Negotiations Data

Attribute	**Type**	**1**	**2**	**3**	**…**	**40**
duration	(number of 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	(number of 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	(number of 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

missing or unknown



(a) (b)

overfitting

Fielded Applications (1/3)

- Web mining
  - Ranking the results of your search
  - Advanced query
  - Advertisements
  - e-commerce
    - Market basket analysis
    - Recommendations
  - Social network analysis

Fielded Applications (2/3)

- Decisions involving judgment
  - Loan companies
  - Credit card companies
- Screening images
  - Detect oil slicks from satellite images
- Load forecasting
  - In the electricity supply industry, it is important to determine future demand for power as far in advance as possible

Fielded Applications (3/3)

- Diagnosis
  - Preventative maintenance of electromechanical devices such as motors and generators
- Marketing and sales
  - Credit assessment
  - Customer loyalty
  - Market basket analysis
  - Direct marketing

Data Mining and Ethics (1/2)

- The use of data—particularly data about people—for data mining has serious ethical implications
- Re-identification techniques
  - 85% of Americans can be identified using five-digit zip code, birth date, and sex
  - 50% of Americans can be identified using city, birth date, and sex
  - If you really do remove all possible identification information from a database, you will probably be left with nothing useful

Data Mining and Ethics (2/2)

- When presented with data, you need to ask who is permitted to have access to it, for what purpose it was collected, and what kind of conclusions are legitimate to draw from it
- data -> information -> knowledge -> wisdom