Chapter 2

Data Mining

A Closer Look

2.1 Data Mining Strategies

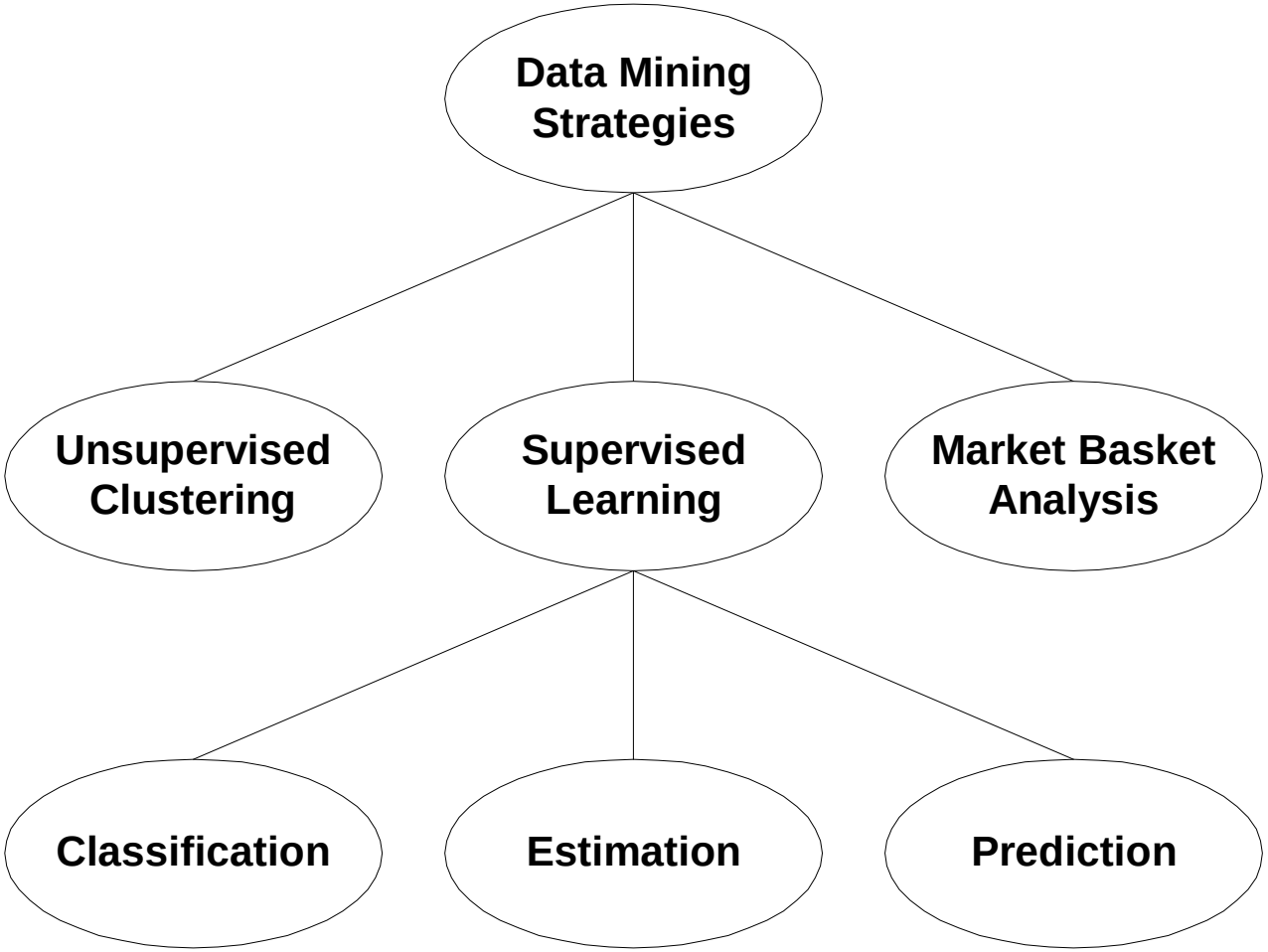

Figure 2.1 A hierarchy of data mining strategies

Data Mining Strategies: Classification

- Learning is supervised.
- The dependent variable is categorical.
- Well-defined classes.
- Current rather than future behavior.

Data Mining Strategies: Estimation

- Learning is supervised.
- The dependent variable is numeric.
- Well-defined classes.
- Current rather than future behavior.

Data Mining Strategies: Prediction

- The emphasis is on predicting future rather than current outcomes.
- The output attribute may be categorical or numeric.

Classification, Estimation or Prediction?

The nature of the data determines whether a model is suitable for classification, estimation, or prediction.

The Cardiology Patient Dataset

This dataset contains 303 instances. Each instance holds information about a patient who either has or does not have a heart condition.

The Cardiology Patient Dataset

- 138 instances represent patients with heart disease.
- 165 instances contain information about patients free of heart disease.

Table 2.1 • **Cardiology Patient Data**

Attribute Name	Mixed Values	Numeric Values	Comments
Age	Numeric	Numeric	Age in years
Gender	Male, Female	1, 0	Patient gender
Chest Pain Type	Angina, Abnormal Angina, NoTang, Asymptomatic	1–4	NoTang =Nonanginal pain
Blood Pressure	Numeric	Numeric	Resting blood pressure upon hospital admission
Cholesterol	Numeric	Numeric	Serum cholesterol
Fasting Blood Sugar < 120	True, False	1, 0	Is fasting blood sugar less than 120?
Resting ECG	Normal, Abnormal, Hyp	0, 1, 2	Hyp = Left ventricular hypertrophy
Maximum Heart Rate	Numeric	Numeric	Maximum heart rate achieved
Induced Angina?	True, False	1, 0	Does the patient experience angina as a result of exercise?
Old Peak	Numeric	Numeric	ST depression induced by exercise relative to rest
Slope	Up, flat, down	1–3	Slope of the peak exercise ST segment
Number Colored Vessels	0, 1, 2, 3	0, 1, 2, 3	Number of major vessels colored by fluorosopy
Thal	Normal fix, rev	3, 6, 7	Normal, fixed defect, reversible defect
Concept Class	Healthy, Sick	1, 0	Angiographic disease status

Table 2.2 • **Typical and Atypical Instances from the Cardiology Domain**

Attribute Name	Typical Healthy Class	Atypical Healthy Class	Typical Sick Class	Atypical Sick Class
Age	52	63	60	62
Gender	Male	Male	Male	Female
Chest Pain Type	NoTang	Angina	Asymptomatic	Asymptomatic
Blood Pressure	138	145	125	160
Cholesterol	223	233	258	164
Fasting Blood Sugar < 120	False	True	False	False
Resting ECG	Normal	Hyp	Hyp	Hyp
Maximum Heart Rate	169	150	141	145
Induced Angina?	False	False	True	False
Old Peak	0	2.3	2.8	6.2
Slope	Up	Down	Flat	Down
Number of Colored Vessels	0	0	1	3
Thal	Normal	Fix	Rev	Rev

Classification, Estimation or Prediction?

Classification, Estimation or Prediction?

- Can we say that either of the two rules seen in the slides that follow are predictive?

- Why or why not?

Here is a rule for the Healthy Class

IF Maximum Heart Rate >=158.333

THEN Class = Healthy

Rule precision: 75.60%

Rule coverage: 40.60%

Rule Precision & Rule Coverage

- Rule coverage specifies the percent of data instances that satisfy the rule's preconditions.
- Rule Precision tells us the percent of the covered instances that are in the class specified by the rule.

Here is a rule for the Sick Class

IF Thal = Rev

THEN Class = Sick

Rule precision:76.30%

Rule coverage: 38.90%

Data Mining Strategies: Unsupervised Clustering

Unsupervised Clustering can be used to:

- Detect Fraud
- Evaluate the likely performance of a supervised model.
- Determine a best set of input attributes for supervised learning.
- Detect Outliers
- Much more..

Data Mining Strategies: Market Basket Analysis

- Find interesting relationships among retail products.
- Uses association rule algorithms.

2.2 Supervised Data Mining Techniques

The Credit Card Promotion Database

Table 2.3 • **The Credit Card Promotion Database**

Income Range ($)	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Gender	Age
40–50K	Yes	No	No	No	Male	45
30–40K	Yes	Yes	Yes	No	Female	40
40–50K	No	No	No	No	Male	42
30–40K	Yes	Yes	Yes	Yes	Male	43
50–60K	Yes	No	Yes	No	Female	38
20–30K	No	No	No	No	Female	55
30–40K	Yes	No	Yes	Yes	Male	35
20–30K	No	Yes	No	No	Male	27
30–40K	Yes	No	No	No	Male	43
30–40K	Yes	Yes	Yes	No	Female	41
40–50K	No	Yes	Yes	No	Female	43
20–30K	No	Yes	Yes	No	Male	29
50–60K	Yes	Yes	Yes	No	Female	39
40–50K	No	Yes	No	No	Male	55
20–30K	No	No	Yes	Yes	Female	19

A Hypothesis for the Credit Card Promotion Database

A combination of one or more of the dataset attributes differentiate Acme Credit Card Company card holders who have taken advantage of the life insurance promotion and those card holders who have chosen not to participate in the promotional offer.

Supervised Data Mining Techniques: Production Rules

Two Predictive Rules for the Credit Card Promotion Database

IF *Gender = Female*
THEN *Life Insurance Promotion = Yes*
Rule precision: 85.7%
Rule coverage: 46.7%

IF *Gender = Male* and *Income Range = 40–50K*
THEN *Life Insurance Promotion = No*
Rule precision: 100.00%
Rule coverage: 20.00%

Supervised Data Mining Techniques: Neural Networks

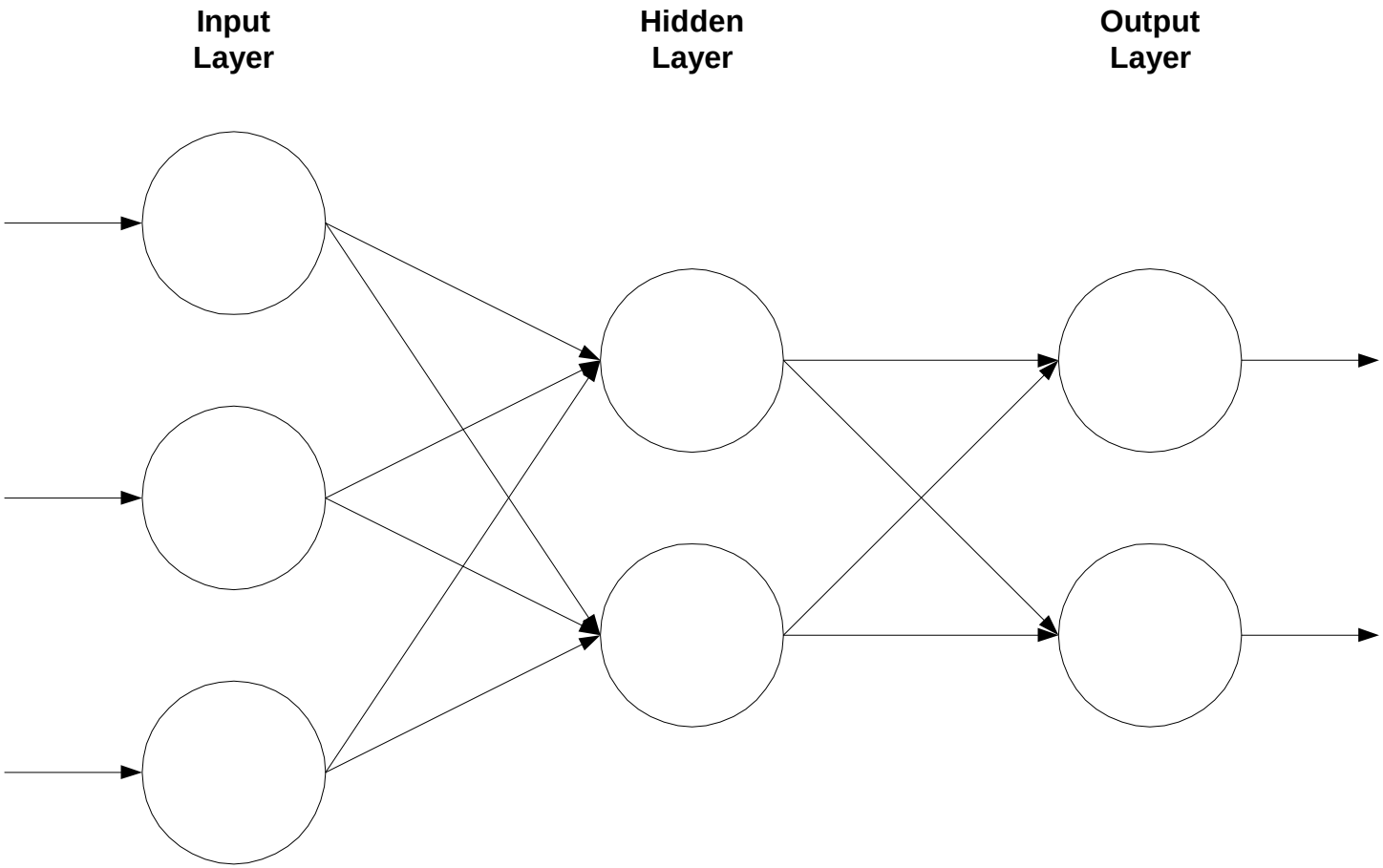

Figure 2.2 A fully connected multilayer neural network

Table 2.4 • **Neural Network Training: Actual and Computed Output**

Instance Number	Life Insurance Promotion	Computed Output
1	0	0.024
2	1	0.998
3	0	0.023
4	1	0.986
5	1	0.999
6	0	0.050
7	1	0.999
8	0	0.262
9	0	0.060
10	1	0.997
11	1	0.999
12	1	0.776
13	1	0.999
14	0	0.023
15	1	0.999

Supervised Data Mining Techniques: Statistical Regression

Life insurance promotion =

0.5909 (credit card insurance) -

0.5455 (gender) + 0.7727

2.3 Association Rules

Comparing Association Rules & Production Rules

- Association rules can have one or several output attributes. Production rules are limited to one output attribute.
- With association rules, an output attribute for one rule can be an input attribute for another rule.

Two Association Rules for the Credit Card Promotion Database

IF *Gender = Female and Age = over40 and Credit Card Insurance = No*
THEN *Life Insurance Promotion = Yes*

IF *Gender = Female and Age = over40*
THEN *Credit Card Insurance = No* and *Life Insurance Promotion = Yes*

2.4 Clustering Techniques

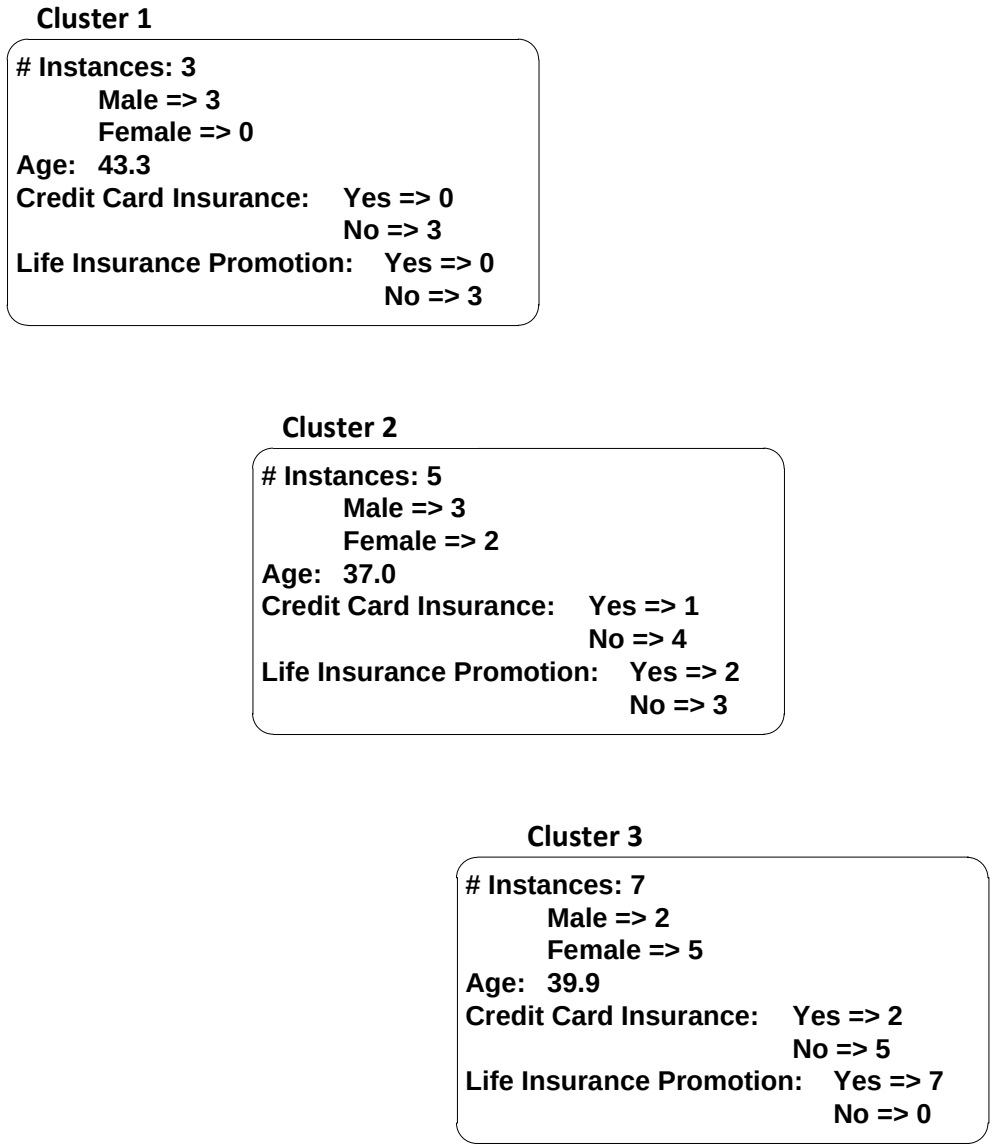

Figure 2.3 An unsupervised clustering of the credit card database

2.5 Evaluating Performance

Evaluating Supervised Learner Models

Confusion Matrix

- A matrix used to summarize the results of a supervised classification.
- Entries along the main diagonal are correct classifications.
- Entries other than those on the main diagonal are classification errors.

Table 2.5 • **A Three-Class Confusion Matrix**

		Computed Decision		
		c_1	c_2	c_3
True Class	c_1	C_{11}	C_{12}	C_{13}
	c_2	C_{21}	C_{22}	C_{23}
	c_3	C_{31}	C_{32}	C_{33}

Two-Class Error Analysis

Table 2.6 • **A Simple Confusion Matrix**

	Computed Accept	**Computed Reject**
Accept	True Accept	False Reject
Reject	False Accept	True Reject

Table 2.7 • **Two Confusion Matrices Each Showing a 10% Error Rate**

Model A	Computed Accept	Computed Reject	Model B	Computed Accept	Computed Reject
Accept	600	25	Accept	600	75
Reject	75	300	Reject	25	300

Evaluating Numeric Output

- Mean absolute error
- Mean squared error
- Root mean squared error

Mean Absolute Error

The average absolute difference between classifier predicted output and actual output.

Mean Squared Error

The average of the sum of squared differences between classifier predicted output and actual output.

Root Mean Squared Error

The square root of the mean squared error.

Comparing Models by Measuring Lift

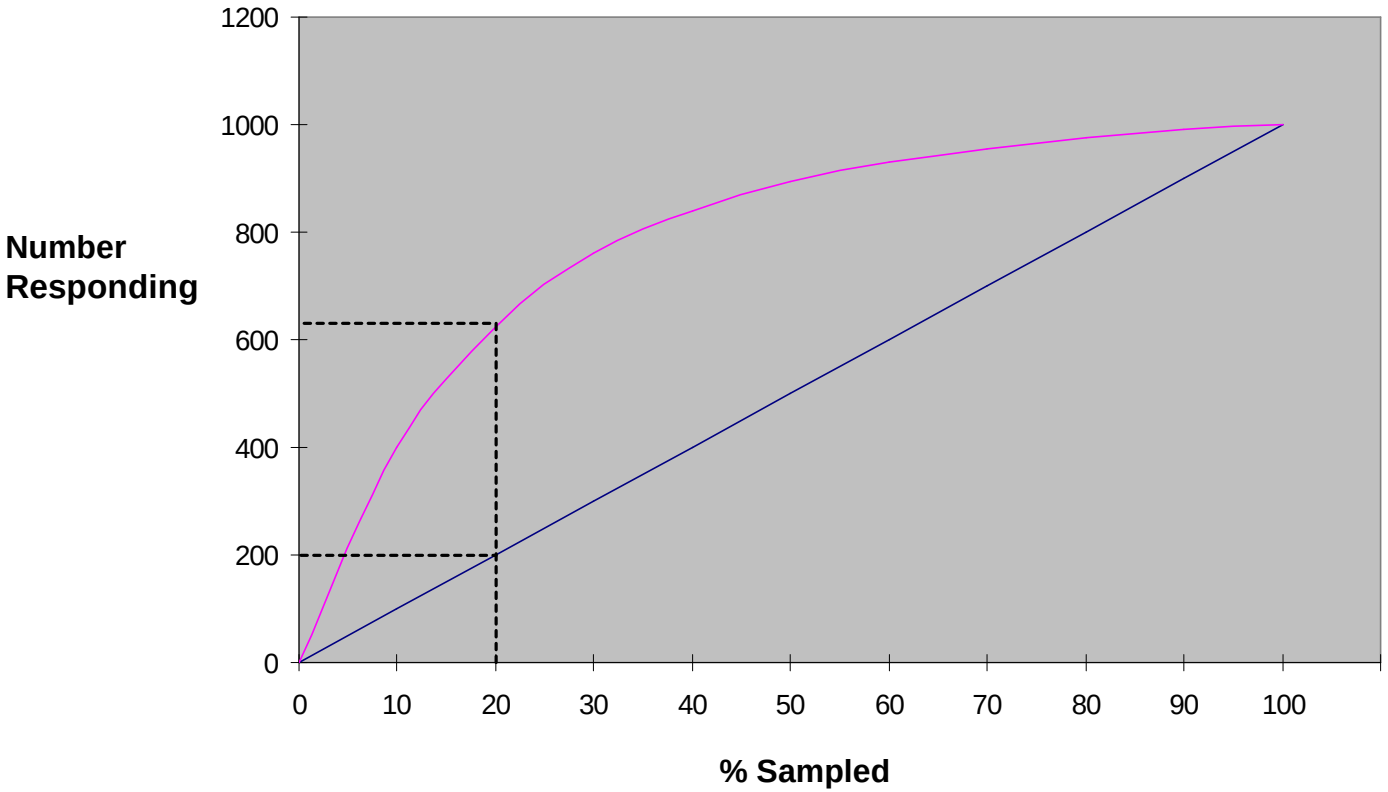

Figure 2.4 Targeted vs. mass mailing

Computing Lift

$$Lift = \frac{P(C_i \mid Sample)}{P(C_i \mid Population)}$$

Table 2.8 • **Two Confusion Matrices: No Model and an Ideal Model**

No Model	Computed Accept	Computed Reject	Ideal Model	Computed Accept	Computed Reject
Accept	1,000	0	Accept	1,000	0
Reject	99,000	0	Reject	0	99,000

Table 2.9 • **Two Confusion Matrices for Alternative Models with Lift Equal to 2.25**

Model X	Computed Accept	Computed Reject	Model Y	Computed Accept	Computed Reject
Accept	540	460	Accept	450	550
Reject	23,460	75,540	Reject	19,550	79,450

Unsupervised Model Evaluation

Unsupervised Model Evaluation (cluster quality)

- All clustering techniques compute some measure of cluster quality.
- One evaluation method is to calculate the sum of squared error differences between the instances of each cluster and their cluster center.
- Smaller values indicate clusters of higher quality.

Supervised Learning for Unsupervised Model Evaluation

- Designate each formed cluster as a class and assign each class an arbitrary name.
- Choose a random sample of instances from each class for supervised learning.
- Build a supervised model from the chosen instances. Employ the remaining instances to test the correctness of the model.