Input:

Concepts, Instances, and Attributes

Contents

Machine Learning & Data Mining

- Data mining:
  - Discovering implicit, previously unknown, and potentially useful information (or knowledge) from data
  - Entire knowledge discovery process including data cleansing, data integration, data transformation, and model building
- Machine learning:
  - ML algorithms acquire structural descriptions from examples
  - Structural descriptions represent patterns underlying the data
    - Can be used to predict outcome in new situation
    - Can be used to understand and explain how prediction is derived
      - ✓ Some learning techniques such as neural nets do not produce explicit description of what is learned

Machine Learning & Data Mining

- Description of the structural patterns:
  - Example from contact lens data (Table 1.1)
  - Can be rules, decision trees, or others (Figures 1.1, 1.2)

If tear production rate = reduced

then recommendation = none

Otherwise, if age = young and astigmatic = no

then recommendation = soft

Machine Learning & Data Mining

Table 1.1 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
Prepresbyopic	Myope	No	Reduced	None
Prepresbyopic	Myope	No	Normal	Soft
Prepresbyopic	Myope	Yes	Reduced	None
Prepresbyopic	Myope	Yes	Normal	Hard
Prepresbyopic	Hypermetrope	No	Reduced	None
Prepresbyopic	Hypermetrope	No	Normal	Soft
Prepresbyopic	Hypermetrope	Yes	Reduced	None
Prepresbyopic	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

Machine Learning & Data Mining

```
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.

Machine Learning & Data Mining

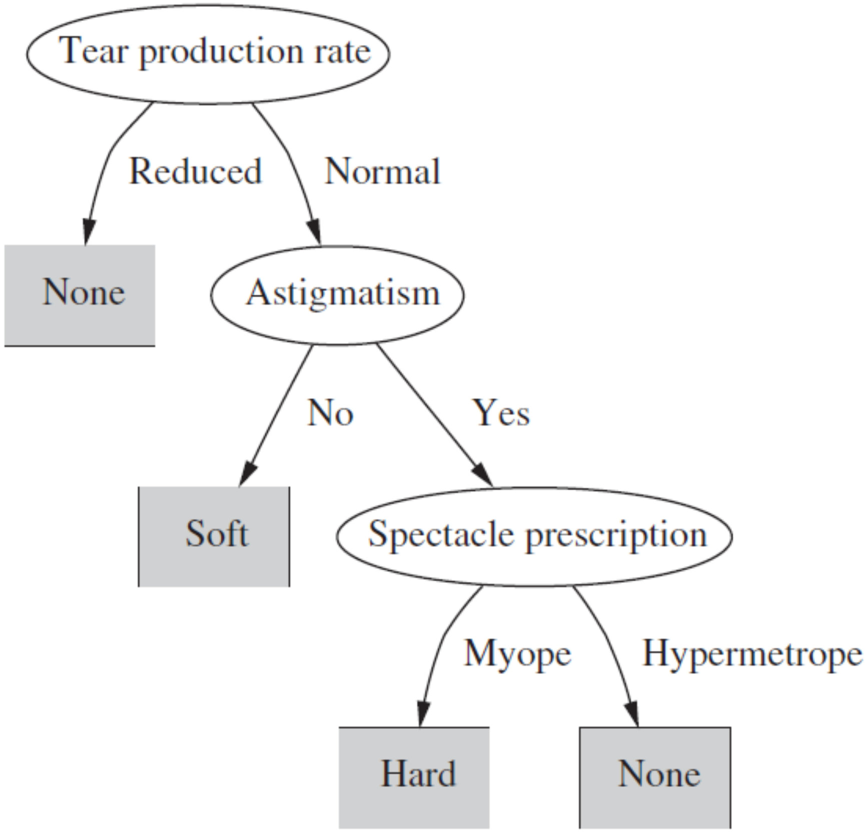

Figure 1.2 Decision tree for the contact lens data.

The Data Mining Process

- The life cycle of a data mining project: (Figure 1.4)
  - Business understanding phase
    - Investigate the business objectives and requirements, and decides whether data mining can be applied to meet them
    - Determine what kind of data can be collected to build a deployable model
  - Data understanding phase
    - Establish an initial dataset and see whether it is suitable for further processing
    - If the data quality is poor, it may be necessary to collect new data based on more stringent criteria

The Data Mining Process

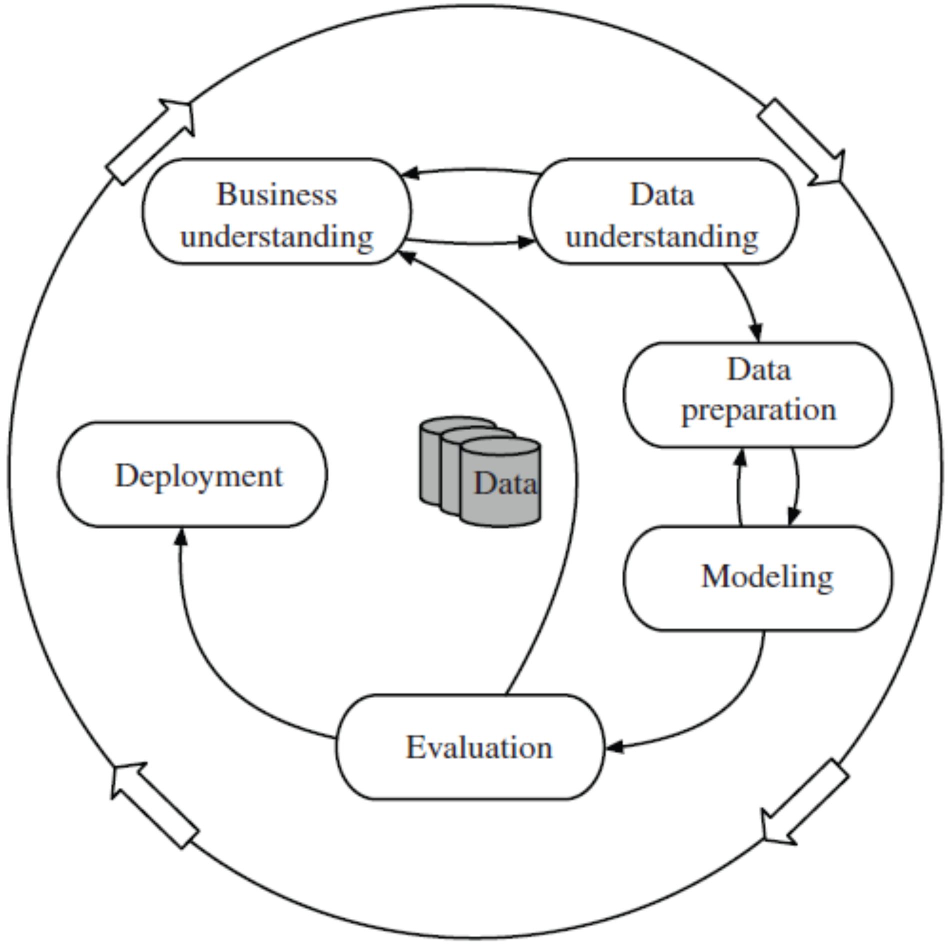

Figure 1.4 Life cycle of a data mining project.

The Data Mining Process

- The life cycle of a data mining project:
  - Data preparation phase
    - Preprocess the raw data so that machine learning algorithms can produce a model
      (often include model building activities as well, e.g., for outlier detection or feature selection)
  - Modelling phase
    - Build models by applying learning algorithms
    - Data preparation and modeling usually go hand in hand: Results obtained during modeling provide new insights that affect the choice of preprocessing techniques

The Data Mining Process

- The life cycle of a data mining project:
  - Evaluation phase
    - Estimate the predictive performance of models built by machine learning
    - If the model is poor, you may need to return to the business understanding step to identify more fruitful business objectives or avenues for data collection
  - Deployment phase
    - Integrate the models into a larger software system

Generalization as Search

- Learning (generalization)
  - Can be viewed as a search through a space of possible concept descriptions (hypotheses) to find one that fits the data
- Impractical to enumerate all possible descriptions:
  - Search space for weather example (Table 1.2)
    - $4 \times 4 \times 3 \times 3 \times 2 = 288$ possibilities for each rule
    - No more than 14 rules $\rightarrow$ about 2.7×10^{34} possible rule sets
- Generalization as a hill-climbing search is a practical option:
  - Heuristic search with some preference criterion
  - No guarantee to find an optimal description

Generalization as Search

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

Bias

- Important decisions for a successful search:
  - How to select a language that can best describe the target concept? (language bias)
  - What is the most efficient search heuristic? (search bias)
    - Greedy search, local search
    - General-to-specific search, specific-to-general search
  - How to avoid overfitting to the particular training data? (overfitting-avoidance bias, a kind of search bias)
    - Simplest-first ordering (preprunning, forward pruning)
    - Postprunning, backward prunning
- Bias is the only means of making search feasible

What Is a Concept?

- Concept: Anything to be learned
  - Classification learning
    - Concept description (e.g., rules or decision trees) can be used to classify unseen data
  - Numeric prediction (Table 1.5)
  - Clustering (unsupervised learning)
    - Natural groups of examples that belong together (Table 2.1)
    - Success is subjectively measured by usefulness

Instances → Learning Algorithm → Concept Description

What Is a Concept?

Table 1.5 The CPU Performance Data

	Cycle Time (ns)	Main Memory (Kb)		Cache (KB)	Channels		Performance
		Min	Max		Min	Max	
	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

What Is a Concept?

Table 2.1 Iris Data as a Clustering Problem

	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
…				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
…				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3.0	5.8	2.2
…				

What Is in an Example?

- The input to a machine learning scheme is a set of instances
  - Each instance is an individual, independent example of the concept to be learned
  - Instances are characterized by the values of a set of predetermined attributes (features)
- Flat file representation:
  - Most machine learning schemes require that the input data be expressed as a table (flat file) of independent instances of the concept to be learned
  - ARFF (attribute-relation file format) file is a standard way of representing datasets (Figure 2.2)

What Is in an Example?

```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
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
```

Figure 2.2 ARFF file for the weather data.

What Is in an Attribute?

- Attributes:
  - A fixed, predefined set of features whose values characterize each instance
- Attribute types
  - Nominal (categorical, discrete) attributes

    If age = young and astigmatic = no and
      tear production range = normal
    then recommendation = soft

    - Boolean (yes/no, true/false) is a special case
    - Sometimes nominal values are coded as integers
      - ✓ E.g., postal zip code
  - Numeric (continuous) attributes

Preparing the Input

- Data preparation is usually the most time consuming and costly process in data mining
  - Real data is often disappointingly low in quality
  - Low-quality data will lead to low-quality mining results

- Data cleansing removes noise and correct inconsistencies in data
- Data integration merges data from multiple sources into a coherent one
- Data transformation eliminates redundant features or discretizes continuous attributes

Sparse Data

- In text mining, for example, the columns represent documents and rows represent how many times a particular word appears in a particular document
  – Most entries are 0 because most documents have a rather small vocabulary
- Instead of representing each value in order, like this:

  0, X, 0, 0, 0, 0, Y, 0, 0, 0, “class A”

  0, 0, 0, W, 0, 0, 0, 0, 0, 0, “class B”

- The nonzero attributes can be explicitly identified by attribute number and their value stated:

  {1 X, 6 Y, 10 “class A”}

  {3 W, 10 “class B”}

Missing Values

- Different kinds of missing values:
  - Unknown vs. unrecorded vs. irrelevant values
  - May be indicated by different out-of-range values (-1, -2, etc.)
- Sometimes the reason for missing is important
  - Due to some random event?
  - Any significance in itself?
  - Intentionally not tested?
- Most machine learning schemes assume a missing value as unknown

Inaccurate Values

- Typographical errors
  - Numeric errors cause outliers
- Data duplication
  - Affects the results of learning algorithms
- Deliberate errors
  - Previously refused insurance applicants may adjust their names
- Stale data
  - Need to check whether the data is still current

Unbalanced Data

- When one class is far more prevalent than the others, raw accuracy may not be a meaningful measure of performance
  - Predicting the majority class for every instance can achieve 99% accuracy if the minority class spans only 1% of the whole data
  - A better way of predicting the minority outcome will inevitably make some errors on some cases with the majority outcome (false alarm)
- In practice, different costs may be associated with the two types of error
  - Cost sensitive evaluation, classification, and learning are needed