F20DL Data Mining and Machine Learning

Diana Bental
(with material from David Corne and slides from http://www.cs.waikato.ac.nz/ml/weka/book.html)

Lecture 2

- Machine learning
  - Some basic Terminology
  - Structural descriptions and
  - Numeric predictions
- Discussed in Witten Frank and Elbe "Data Mining and Machine Learning"
  - examples will recur

Data Mining – Basic Terminology

- Start (usually) from a flat table of data

Gender	weight	height	Age in mths	100m time
Male	52kg	1.71m	243	13.7s
Male	89kg	1.92m	388	22.3s
Female	48kg	1.67m	219	14.6s
Male	86kg	1.96m	274	9.58s
Male	80kg	1.88m	260	10.56s
etc …				

This is called a *data instance* or a *record* or just a *line of data*

Gender	weight	height	Age in mths	100m time
Male	52kg	1.71m	243	13.7s
Male	89kg	1.92m	388	22.3s
Female	48kg	1.67m	219	14.6s
Male	86kg	1.96m	274	9.58s
Male	80kg	1.88m	260	10.56s
etc …				

This is called an *attribute* or *field*; the *value* of the Age field in the 4th record is 274

- Start (usually) from a flat table of data

Gender	weight	height	Age in mths	100m time
Male	52kg	1.71m	243	13.7s
Male	89kg	1.92m	388	22.3s
Female	48kg	1.67m	219	14.6s
Male	86kg	1.96m	274	9.58s
Male	80kg	1.88m	260	10.56s
etc …				

Usually we are interested in predicting the value of a particular attribute, given the values of the other attributes. What we want to predict is called the *target class* (or *class attribute*)

Gender	weight	height	Age in mths	100m time
Male	52kg	1.71m	243	13.7s
Male	89kg	1.92m	388	22.3s
Female	48kg	1.67m	219	14.6s
Male	86kg	1.96m	274	9.58s
Male	80kg	1.88m	260	10.56s
etc …				

What's in an attribute?

- Possible attribute types (*levels of measurement*)
  - Nominal, ordinal, interval and ratio

Nominal

- Example:
  - the attribute **gender**
  - values: ***male*** / ***female***
- Values are distinct symbols
- Values serve only as labels or names
  - Nominal comes from the Latin word for name
- No relation is implied among nominal values
  - no ordering or distance measure
- Only tests for equality can be performed

Ordinal

- Example:
  - Attribute **temperature**
  - Values: ***hot*** > ***mild*** > ***cool***
- The values are in order
- But: there is no defined *distance* between values
- So addition and subtraction don't make sense
  - ***cool*** + ***mild*** = *???!*
- Example rule:
  if **temperature** < ***hot*** → **play** = ***yes***
- The distinction between nominal and ordinal is not always clear
  - E.g. colours, ordered by light wavelength

Interval

- Interval quantities are *ordered* and also measured in *fixed and equal units*
  - Example 1: attribute *temperature* expressed in degrees Fahrenheit
  - Example 2: attribute *year*
- Difference of two values makes sense
  - 2011 AD – 2005 AD = 6 years
- Sum or product doesn't make sense
  - 2011 AD + 2005 AD = ????
- Zero point is not defined!

Ratio

- 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
- But
  - is there a *really* a defined zero point?
  - answer depends on scientific knowledge (e.g. Fahrenheit knew no lower limit to temperature)

Attribute types used in practice

- Different attribute types are suitable for different machine learning techniques
- Many schemes use nominal and ordinal data
  - E.g. Decision trees, rules, association rules
  - Schemes require at least the class attribute to be nominal
- Some schemes use interval or ratio data
  - E.g. Regression, neural networks
- Some schemes can be used for both
  - E.g. Nearest neighbour
- Special case: dichotomy (*boolean* attribute)

Machine learning techniques

- Algorithms for acquiring structural descriptions from examples
- Structural descriptions represent the patterns explicitly
  - predict outcomes in new situations
  - understand and explain how prediction is derived
    - may be even more important
- Methods originate from artificial intelligence, statistics, and research on databases

Structural descriptions

- Example: Contact lens data

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

Contact lens data – in full

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 description: if-then rules

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

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

The contact lens data – 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
```


Complete and correct rules but..

- The rules just summarise the data set
- Is there a smaller set of rules that performs as well?
- Would that be better, and why?
- What if some combinations were not in the dataset?

Example: The "Weather Problem"

- Conditions for playing a game

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

Example: The "Weather Problem"

- Conditions for playing a game

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

```
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
```


Numeric and categorical attributes

- Weather data with mixed attributes

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```


Structural Descriptions: Classification Rules

- So far:
- Classification rules:
  - predict the value of one given attribute - the **class** attribute
  - other attributes may be numeric or categorical
  - class attribute is categorical
  - Eg. Weather game data - play *yes* **/** *no*

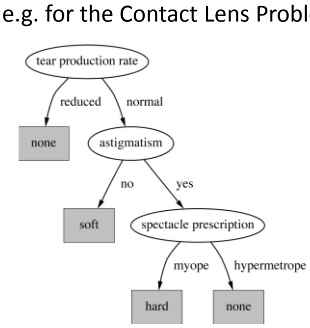
```
If outlook = sunny and humidity = high
   then play = no
```


Structural Descriptions: Association Rules

- Association rules:
  - predict the value of any attribute,
  - or a combination of attributes
  - categorical attributes

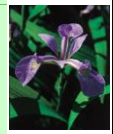
```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```


Structural descriptions: Decision Trees e.g. for the Contact Lens Problem



Numeric data: classifying iris flowers

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	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
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					



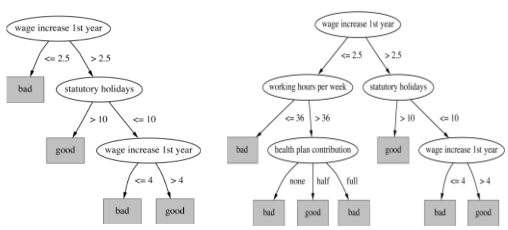
```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```


A more realistic example

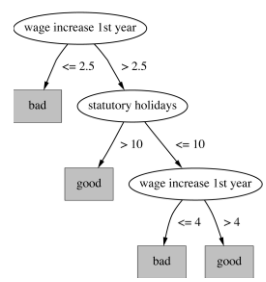
- Canadian labour contract negotiations 1987/8

Attribute	Type	1	2	3	…	40
Duration	(Number of years)	1	2	3		2
Wage increase first year	Percentage	2%	4%	4.3%		4.5
Wage increase second year	Percentage	?	5%	4.4%		4.0
Wage increase third 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

Decision trees for the Canadian labour data



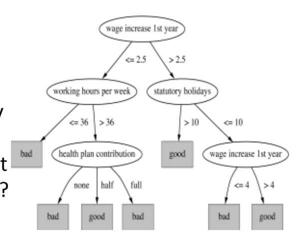
Decision trees for the Canadian labour data



- Simple
- Approximate
  - Will predict *bad* for some outcomes that are really *good*
- But it makes intuitive sense in terms of what is good for the employees

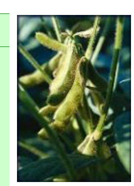
Decision trees for the Canadian labour data

- More complex
- Not intuitive
- If hours exceed 36, why are no health plan or a full health plan bad, but half a health plan good?
- Good for thinking in depth about the structure of the data



Early example of ML success : Soybean disease classification (1970s)

	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	4	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



Early example of ML success : Soybean disease classification (1970s)

- Used 300 carefully selected examples as training data
- Examples selected to be very different from each other
- Automatically derived rules
- Also a plant expert created rules
- And the computer-generated rules performed better than the expert rules on the rest of the data – 97.5% for the machine vs. 72% for the expert rules

Two derived 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
```


Two derived 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
  - If Leaf condition is normal then Leaf malformation must be absent
  - So Leaf malformation is a special case of Leaf abnormality
  - So second rule only applies of there is a different Leaf abnormality
  - Not obvious!

Classifying iris flowers – clustering

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	
2	4.9	3.0	1.4	0.2	
…					
51	7.0	3.2	4.7	1.4	
52	6.4	3.2	4.5	1.5	
…					
101	6.3	3.3	6.0	2.5	
102	5.8	2.7	5.1	1.9	
…					

"Unsupervised" learning

- So far - these were classifications
  - Weather problem – play? yes / no
  - Contact lens – hard / soft / none
  - Iris type - Setosa; Versicolor / Virginica
  - Labour negotiation outcomes - good / bad
  - Soybean diseases – normal / root rot (etc)
  - All predict categories
    - Non-class attributes may be numbers
- What if we're trying to predict a number?

Numeric predictions: predicting CPU performance PRP

- 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Numeric predictions

- Structural Description: Linear regression function

```
PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
      + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

- Structural Description: Neural networks (later)

Take away

- Different structural descriptions suit different data
- Use unsupervised learning if the data has not been pre classified data
- Examples are discussed in Witten Frank and Elbe “Data Mining and Machine Learning”
  - examples will recur