More Data Mining with Weka

Class 3 – Lesson 3

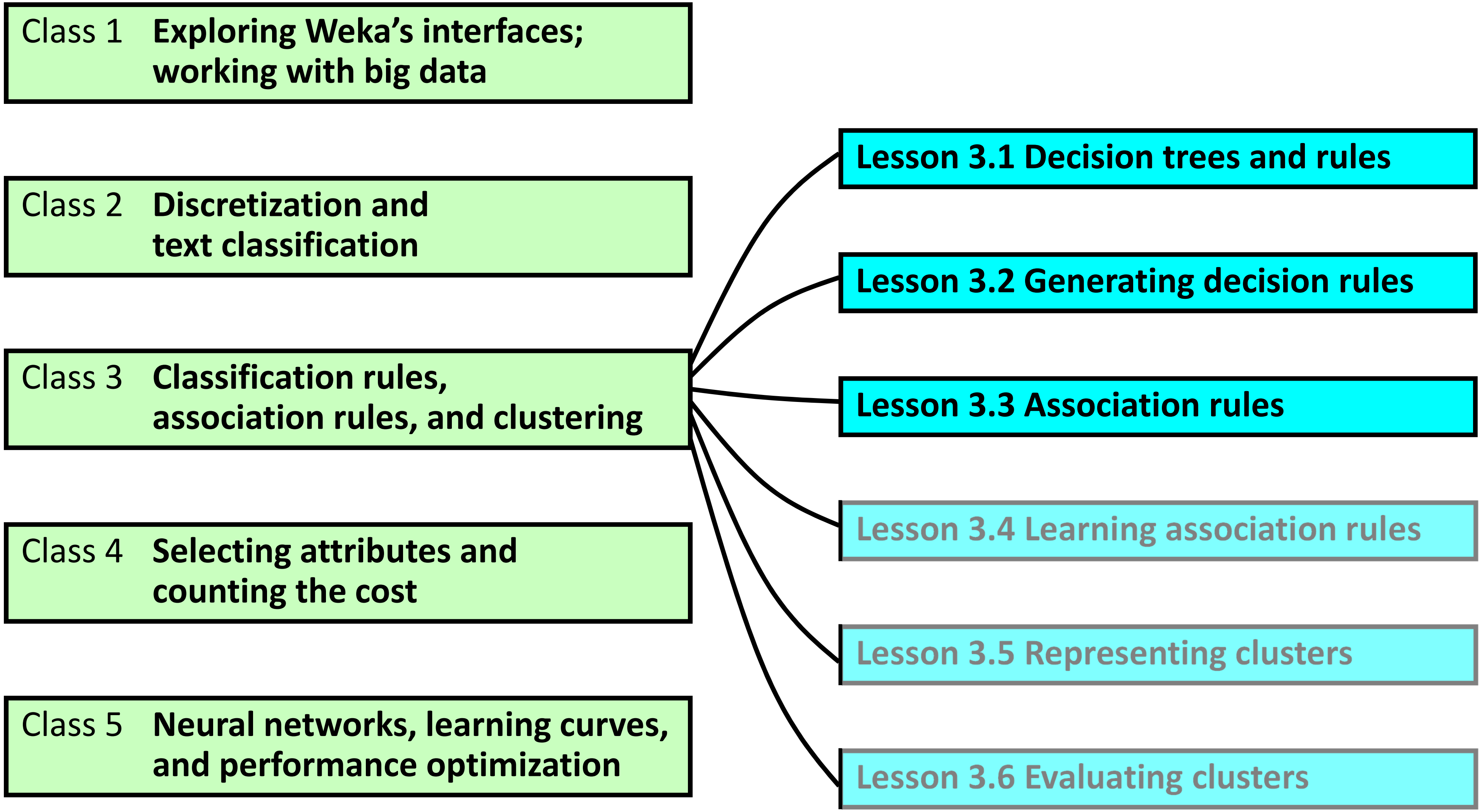
Association rules

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 3.3: Association rules



Class 1 **Exploring Weka's interfaces; working with big data**

Class 2 **Discretization and text classification**

Class 3 **Classification rules, association rules, and clustering**

Class 4 **Selecting attributes and counting the cost**

Class 5 **Neural networks, learning curves, and performance optimization**

Lesson 3.1 Decision trees and rules

Lesson 3.2 Generating decision rules

Lesson 3.3 Association rules

Lesson 3.4 Learning association rules

Lesson 3.5 Representing clusters

Lesson 3.6 Evaluating clusters

Lesson 3.3: Association rules

- With association rules, there is no “class” attribute
- Rules can predict any attribute, or combination of attributes
- Need a different kind of algorithm: “**Apriori**”

Here are some association rules for the weather data:

1. outlook = overcast ==> play = yes
2. temperature = cool ==> humidity = normal
3. humidity = normal & windy = false ==> play = yes
4. outlook = sunny & play = no ==> humidity = high
5. outlook = sunny & humidity = high ==> play = no
6. outlook = rainy & play = yes ==> windy = false
7. outlook = rainy & windy = false ==> play = yes
8. temperature = cool & play = yes ==> humidity = normal
9. outlook = sunny & temperature = hot ==> humidity = high
10. temperature = hot & play = no ==> outlook = sunny

Outlook	**Temp**	**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

Lesson 3.3: Association rules

- **Support:** number of instances that satisfy a rule
- **Confidence:** proportion of instances that satisfy the left-hand side for which the right-hand side also holds
- Specify minimum confidence, seek the rules with greatest support??

			support	confidence
1. outlook = overcast	==>	play = yes	4	100%
2. temperature = cool	==>	humidity = normal	4	100%
3. humidity = normal & windy = false	==>	play = yes	4	100%
4. outlook = sunny & play = no	==>	humidity = high	3	100%
5. outlook = sunny & humidity = high	==>	play = no	3	100%
6. outlook = rainy & play = yes	==>	windy = false	3	100%
7. outlook = rainy & windy = false	==>	play = yes	3	100%
8. temperature = cool & play = yes	==>	humidity = normal	3	100%
9. outlook = sunny & temperature = hot	==>	humidity = high	2	100%
10. temperature = hot & play = no	==>	outlook = sunny	2	100%

Lesson 3.3: Association rules

- **Itemset** set of attribute-value pairs, e.g.

  humidity = normal & windy = false & play = yes   support = 4

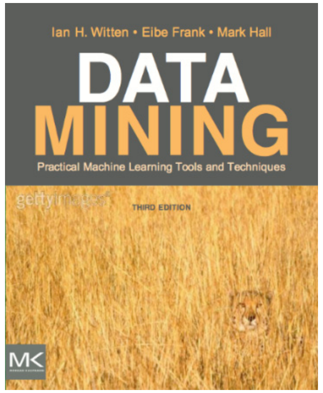
- 7 potential rules from this itemset:

	support	confidence
If humidity = normal & windy = false ==> play = yes	4	4/4
If humidity = normal & play = yes ==> windy = false	4	4/6
If windy = false & play = yes ==> humidity = normal	4	4/6
If humidity = normal ==> windy = false & play = yes	4	4/7
If windy = false ==> humidity = normal & play = yes	4	4/8
If play = yes ==> humidity = normal & windy = false	4	4/9
==> humidity = normal & windy = false & play = yes	4	4/14

- Generate high-support itemsets, get several rules from each
- Strategy: iteratively reduce the minimum support until the required number of rules is found with a given minimum confidence

Lesson 3.3: Association rules

- There are far more association rules than classification rules
  - need different techniques
- *Support* and *Confidence* are measures of a rule
- Apriori is the standard association-rule algorithm
- Want to specify minimum confidence value and seek rules with the most support
- Details? – see next lesson


Ian H. Witten • Eibe Frank • Mark Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Course text

- Section 4.5 *Mining association rules*

More Data Mining with Weka

Class 3 – Lesson 4

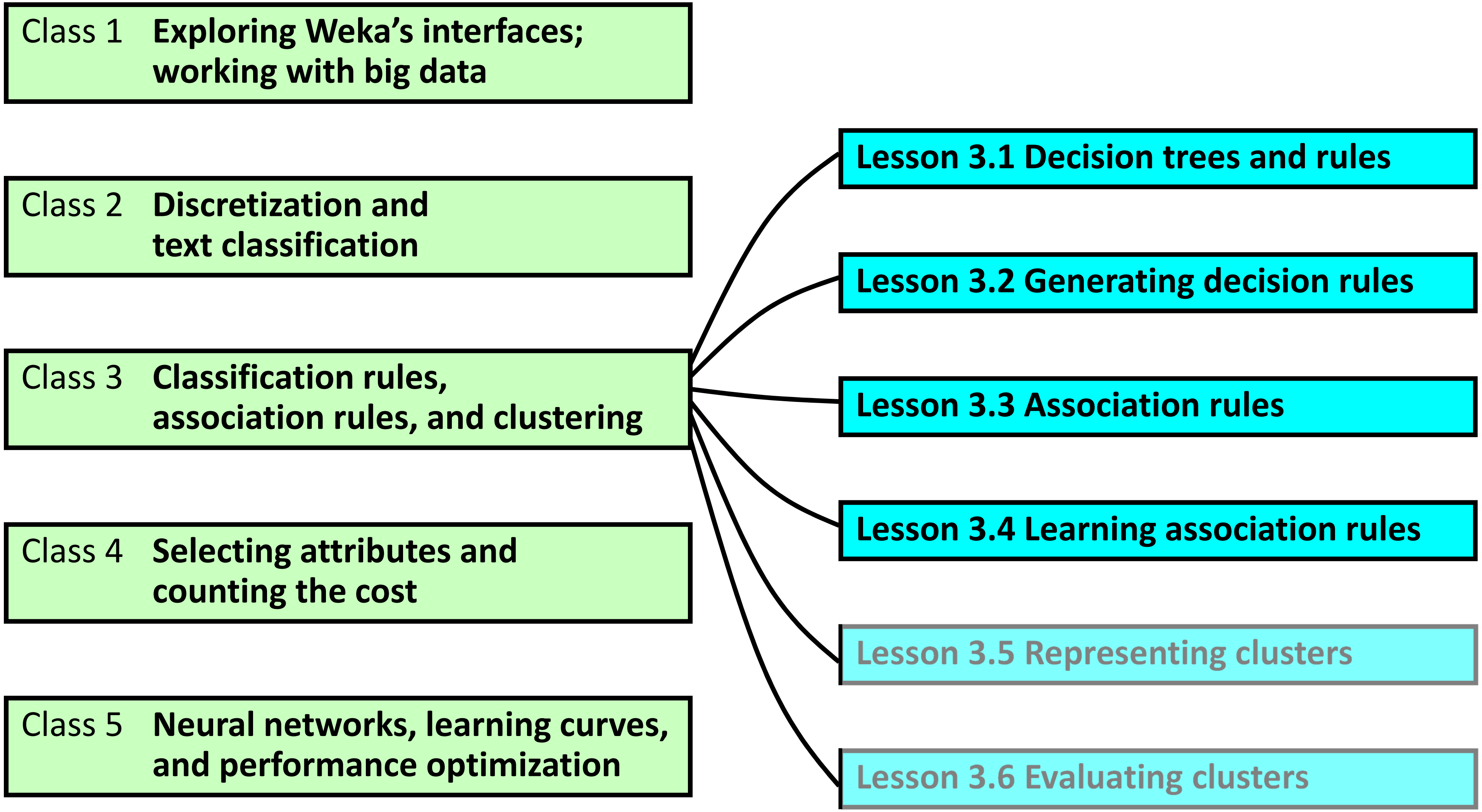
Learning association rules

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 3.4: Learning association rules



Class 1 **Exploring Weka's interfaces; working with big data**

Class 2 **Discretization and text classification**

Class 3 **Classification rules, association rules, and clustering**

Class 4 **Selecting attributes and counting the cost**

Class 5 **Neural networks, learning curves, and performance optimization**

Lesson 3.1 Decision trees and rules

Lesson 3.2 Generating decision rules

Lesson 3.3 Association rules

Lesson 3.4 Learning association rules

Lesson 3.5 Representing clusters

Lesson 3.6 Evaluating clusters

Lesson 3.4: Learning association rules

Strategy

- *specify minimum confidence*
- *iteratively reduce support until enough rules are found with > this confidence*

7 potential rules from a single itemset:

	support	confidence
If humidity = normal & windy = false ==> play = yes	4	4/4
If humidity = normal & play = yes ==> windy = false	4	4/6
If windy = false & play = yes ==> humidity = normal	4	4/6
If humidity = normal ==> windy = false & play = yes	4	4/7
If windy = false ==> humidity = normal & play = yes	4	4/8
If play = yes ==> humidity = normal & windy = false	4	4/9
==> humidity = normal & windy = false & play = yes	4	4/14

1. Generate itemsets with support 14 (none)
2. find rules with > min confidence level (Weka default: 90%)
3. continue with itemsets with support 13 (none)

… and so on, until sufficient rules have been generated

Lesson 3.4: Learning association rules

- ❖ Weather data has 336 rules with confidence 100%!
  - *but only 8 have support ≥ 3, only 58 have support ≥ 2*
- ❖ Weka: specify minimum confidence level (minMetric, default 90%)

  number of rules sought (numRules, default 10)
- ❖ Support is expressed as a proportion of the number of instances
- ❖ Weka runs Apriori algorithm several times

  starts at upperBoundMinSupport (usually left at 100%)

  decreases by delta at each iteration (default 5%)

  stops when numRules reached

  … or at lowerBoundMinSupport (default 10%)

Lesson 3.4: Learning association rules

```
Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17
Generated sets of large itemsets:
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 47
Size of set of large itemsets L(3): 39
Size of set of large itemsets L(4): 6

Best rules found:
 1. outlook = overcast 4 ==> play = yes 4
```

- 17 cycles of Apriori algorithm:
  - *support = 100%, 95%, 90%, …, 20%, 15%*
  - *14, 13, 13, …, 3, 2 instances*
  - *only 8 rules with conf > 0.9 & support ≥ 3*
- to see itemsets, set outputItemSets
  - *they're based on the final support value, i.e. 2*

12 one-item sets with support ≥ 2

```
outlook = sunny 5
outlook = overcast 4
  ...
play = no 5
```

47 two-item sets with support ≥ 2

```
outlook = sunny & temperature = hot 2
outlook = sunny & humidity = high 3
  ...
```

39 three-item sets with support ≥ 2

```
outlook = sunny & temperature = hot & humidity = high 2
outlook = sunny & humidity = high & play = no 3
outlook = sunny & windy = false & play = no 2
  ...
```

6 four-item sets with support ≥ 2

```
outlook = sunny & humidity = high & windy = false
   & play = no 2
 ...
```

Lesson 3.4: Learning association rules

Other parameters in Weka implementation

- car: always produce rules that predict the class attribute
  - *set the class attribute using classIndex*
- significanceLevel: filter rules according to a statistical test (χ^2)
  - *unreliable because with so many tests, significant results will be found just by chance*
  - *the test is inaccurate for small support values*
- metricType: different measures for ranking rules
  - *Confidence*
  - *Lift*
  - *Leverage*
  - *Conviction*
- removeAllMissingCols: removes attribute whose values are all “missing”

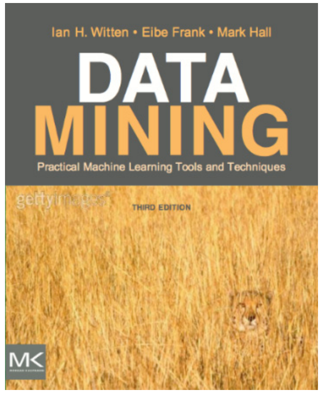
Lesson 3.4: Learning association rules

Market basket analysis

- Look at supermarket.arff
  - *collected from an actual New Zealand supermarket*
- 4500 instances, 220 attributes; 1M attribute values
- Missing values used to indicate that the basket did not contain that item
- 92% of values are missing
  - *average basket contains 220×8% = 18 items*
- Most popular items: bread-and-cake (3330), vegetables (2961), frozen foods (2717), biscuits (2605)

Lesson 3.4: Learning association rules

- Apriori makes multiple passes through the data
  - generates 1-item sets, 2-item sets, … with more than minimum support
  - turns each one into (many) rules and checks their confidence
- Fast and efficient (provided data fits into main memory)
- Weka invokes Apriori several times gradually reducing the support until sufficient high-confidence rules have been found
  - there are parameters to control this
- Activity: supermarket data


Ian H. Witten • Eibe Frank • Mark Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Course text

- Section 11.7 *Association-rule learners*