Data Mining and Knowledge Discovery
Practice notes: Decision trees

Data Mining and Knowledge Discovery: Practice Notes

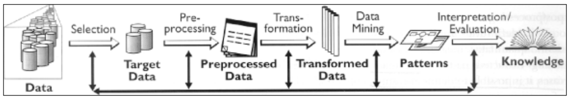
dr. Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
2016/11/15

1

- Prof. Nada Lavrač:
  - Data mining overview
  - Advanced topics
- Dr. Petra Kralj Novak
  - Data mining basis

2

Keywords



- Data
  - Attribute, example, attribute-value data, target variable, class, discretization
- Algorithms
  - Decision tree induction, entropy, information gain, overfitting, Occam's razor, model pruning, naïve Bayes classifier, KNN, association rules, support, confidence, numeric prediction, regression tree, model tree, heuristics vs. exhaustive search, predictive vs. descriptive DM
- Evaluation
  - Train set, test set, accuracy, confusion matrix, cross validation, true positives, false positives, ROC space, error, precision, recall

3

Decision tree induction

Given
- Attribute-value data with nominal target variable

Induce
- A decision tree and estimate its performance on new data

4

Attribute-value data

(nominal) target variable

attributes

examples

Person	Age	Prescription	Astigmatic	Tear_Rate	**Lenses**
P1	young	myope	no	normal	**YES**
P2	young	myope	no	reduced	**NO**
P3	young	hypermetrope	no	normal	**YES**
P4	young	hypermetrope	no	reduced	**NO**
P5	young	myope	yes	normal	**YES**
P6	young	myope	yes	reduced	**NO**
P7	young	hypermetrope	yes	normal	**YES**
P8	young	hypermetrope	yes	reduced	**NO**
P9	pre-presbyopic	myope	no	normal	**YES**
P10	pre-presbyopic	myope	no	reduced	**NO**
P11	pre-presbyopic	hypermetrope	no	normal	**YES**
P12	pre-presbyopic	hypermetrope	no	reduced	**NO**
P13	pre-presbyopic	myope	yes	normal	**YES**
P14	pre-presbyopic	myope	yes	reduced	**NO**
P15	pre-presbyopic	hypermetrope	yes	normal	**NO**
P16	pre-presbyopic	hypermetrope	yes	reduced	**NO**
P17	presbyopic	myope	no	normal	**NO**
P18	presbyopic	myope	no	reduced	**NO**
P19	presbyopic	hypermetrope	no	normal	**YES**
P20	presbyopic	hypermetrope	no	reduced	**NO**
P21	presbyopic	myope	yes	normal	**YES**
P22	presbyopic	myope	yes	reduced	**NO**
P23	presbyopic	hypermetrope	yes	normal	**NO**
P24	presbyopic	hypermetrope	yes	reduced	**NO**

classes = values of the (nominal) target variable

5

Decision tree induction (ID3)

Given:
Attribute-value data with nominal target variable
Divide the data into training set (S) and test set (T)

Induce a decision tree on training set S:
1. Compute the entropy E(S) of the set S
2. **IF** E(S) = 0
3. The current set is "clean" and therefore a leaf in our tree
4. **IF** E(S) > 0
5. Compute the information gain of each attribute Gain(S, A)
6. The attribute A with the highest information gain becomes the root
7. Divide the set S into subsets S_i according to the values of A
8. Repeat steps 1-7 on each Si

Test the model on the test set T

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106

6

Data Mining and Knowledge Discovery
Practice notes: Decision trees

Training and test set

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P1	young	myope	no	normal	**YES**
P2	young	myope	no	reduced	**NO**
P3	young	hypermetrope	no	normal	**YES**
P4	young	hypermetrope	no	reduced	**NO**
P5	young	myope	yes	normal	**YES**
P6	young	myope	yes	reduced	**NO**
P7	young	hypermetrope	yes	normal	**YES**
P8	young	hypermetrope	yes	reduced	**NO**
P9	pre-presbyopic	myope	no	normal	**YES**
P10	pre-presbyopic	myope	no	reduced	**NO**
P11	pre-presbyopic	hypermetrope	no	normal	**YES**
P12	pre-presbyopic	hypermetrope	no	reduced	**NO**
P13	pre-presbyopic	myope	yes	normal	**YES**
P14	pre-presbyopic	myope	yes	reduced	**NO**
P15	pre-presbyopic	hypermetrope	yes	normal	**NO**
P16	pre-presbyopic	hypermetrope	yes	reduced	**NO**
P17	presbyopic	myope	no	normal	**NO**
P18	presbyopic	myope	no	reduced	**NO**
P19	presbyopic	hypermetrope	no	normal	**YES**
P20	presbyopic	hypermetrope	no	reduced	**NO**
P21	presbyopic	myope	yes	normal	**YES**
P22	presbyopic	myope	yes	reduced	**NO**
P23	presbyopic	hypermetrope	yes	normal	**NO**
P24	presbyopic	hypermetrope	yes	reduced	**NO**

Put 30% of examples in a separate test set

Test set

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P3	young	hypermetrope	no	normal	**YES**
P9	pre-presbyopic	myope	no	normal	**YES**
P12	pre-presbyopic	hypermetrope	no	reduced	**NO**
P13	pre-presbyopic	myope	yes	normal	**YES**
P15	pre-presbyopic	hypermetrope	yes	normal	**NO**
P16	pre-presbyopic	hypermetrope	yes	reduced	**NO**
P23	presbyopic	hypermetrope	yes	normal	**NO**

Put these data away and do not look at them in the training phase!

Training set

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P1	young	myope	no	normal	**YES**
P2	young	myope	no	reduced	**NO**
P4	young	hypermetrope	no	reduced	**NO**
P5	young	myope	yes	normal	**YES**
P6	young	myope	yes	reduced	**NO**
P7	young	hypermetrope	yes	normal	**YES**
P8	young	hypermetrope	yes	reduced	**NO**
P10	pre-presbyopic	myope	no	reduced	**NO**
P11	pre-presbyopic	hypermetrope	no	normal	**YES**
P14	pre-presbyopic	myope	yes	reduced	**NO**
P17	presbyopic	myope	no	normal	**NO**
P18	presbyopic	myope	no	reduced	**NO**
P19	presbyopic	hypermetrope	no	normal	**YES**
P20	presbyopic	hypermetrope	no	reduced	**NO**
P21	presbyopic	myope	yes	normal	**YES**
P22	presbyopic	myope	yes	reduced	**NO**
P24	presbyopic	hypermetrope	yes	reduced	**NO**

Decision tree induction (ID3)

Given:
- Attribute-value data with nominal target variable
- Divide the data into training set (S) and test set (T)

Induce a decision tree on training set S:
1. Compute the entropy E(S) of the set S
2. **IF** E(S) = 0
3. The current set is "clean" and therefore a leaf in our tree
4. **IF** E(S) > 0
5. Compute the information gain of each attribute Gain(S, A)
6. The attribute A with the highest information gain becomes the root
7. Divide the set S into subsets S_i according to the values of A
8. Repeat steps 1-7 on each Si

Test the model on the test set T

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106

Information gain

number of examples in the subset S_v

(probability of the branch)

set S

attribute A

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot E(S_v)$$

number of examples in set S

Entropy

$$E(S) = -\sum_{c=1}^{N} p_c . \log_2 p_c$$

- Calculate the following entropies:
  - E (0,1) =
  - E (1/2 , 1/2) =
  - E (1/4 , 3/4) =
  - E (1/7 , 6/7) =
  - E (6/7 , 1/7) =
  - E (0.1 , 0.9) =
  - E (0.001 , 0.999) =

Data Mining and Knowledge Discovery
Practice notes: Decision trees

Entropy

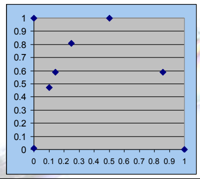
$$E(S) = -\sum_{c=1}^{N} p_c . \log_2 p_c$$

- Calculate the following entropies:
  - E (0,1) = 0
  - E (1/2 , 1/2) = 1
  - E (1/4 , 3/4) = 0.81
  - E (1/7 , 6/7) = 0.59
  - E (6/7 , 1/7) = 0.59
  - E (0.1 , 0.9) = 0.47
  - E (0.001 , 0.999) = 0.01

Entropy

$$E(S) = -\sum_{c=1}^{N} p_c . \log_2 p_c$$

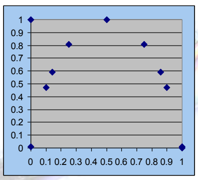
- Calculate the following entropies:
  - E (0,1) = 0
  - E (1/2 , 1/2) = 1
  - E (1/4 , 3/4) = 0.81
  - E (1/7 , 6/7) = 0.59
  - E (6/7 , 1/7) = 0.59
  - E (0.1 , 0.9) = 0.47
  - E (0.001 , 0.999) = 0.01



Entropy

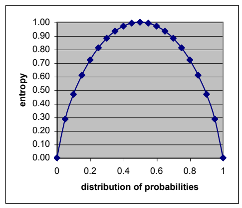
$$E(S) = -\sum_{c=1}^{N} p_c . \log_2 p_c$$

- Calculate the following entropies:
  - E (0,1) = 0
  - E (1/2 , 1/2) = 1
  - E (1/4 , 3/4) = 0.81
  - E (1/7 , 6/7) = 0.59
  - E (6/7 , 1/7) = 0.59
  - E (0.1 , 0.9) = 0.47
  - E (0.001 , 0.999) = 0.01



Entropy and information gain

probability of class 1 p_1	probability of class 2 $p_2 = 1-p_1$	entropy $E(p_1, p_2) = -p_1{*}\log_2(p_1) - p_2{*}\log_2(p_2)$
0	1	0.00
0.05	0.95	0.29
0.10	0.90	0.47
0.15	0.85	0.61
0.20	0.80	0.72
0.25	0.75	0.81
0.30	0.70	0.88
0.35	0.65	0.93
0.40	0.60	0.97
0.45	0.55	0.99
0.50	0.50	1.00
0.55	0.45	0.99
0.60	0.40	0.97
0.65	0.35	0.93
0.70	0.30	0.88
0.75	0.25	0.81
0.80	0.20	0.72
0.85	0.15	0.61
0.90	0.10	0.47
0.95	0.05	0.29
1	0	0.00



number of examples in the subset

probability of the "branch"

attribut A

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

set S

number of examples in set S

Decision tree induction (ID3)

Given:

Attribute-value data with nominal target variable

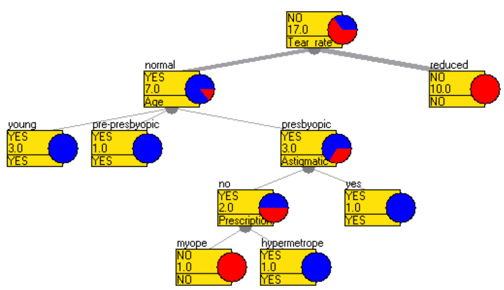
Divide the data into training set (S) and test set (T)

Induce a decision tree on training set S:

1. Compute the entropy E(S) of the set S
2. **IF** E(S) = 0
3. The current set is "clean" and therefore a leaf in our tree
4. **IF** E(S) > 0
5. Compute the information gain of each attribute Gain(S, A)
6. The attribute A with the highest information gain becomes the root
7. Divide the set S into subsets S_i according to the values of A
8. Repeat steps 1-7 on each Si

Test the model on the test set T

Decision tree



Data Mining and Knowledge Discovery
Practice notes: Decision trees

Confusion matrix

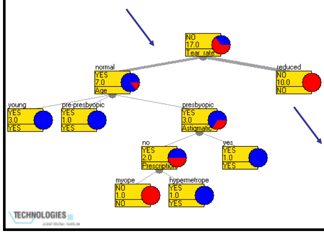
predicted

actual	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

- Confusion matrix is a matrix showing actual and predicted classifications
- Classification measures can be calculated from it, like classification accuracy
  = #(correctly classified examples) / #(all examples)
  = (TP+TN) / (TP+TN+FP+FN)

Evaluating decision tree accuracy

Person	Age	Prescription	Astigmatic	Tear_Rate	Lenses
P3	young	hypermetrope	no	normal	**YES**
P9	pre-presbyopic	myope	no	normal	**YES**
P12	pre-presbyopic	hypermetrope	no	reduced	**NO**
P13	pre-presbyopic	myope	yes	normal	**YES**
P15	pre-presbyopic	hypermetrope	yes	normal	**NO**
P16	pre-presbyopic	hypermetrope	yes	reduced	**NO**
P23	presbyopic	hypermetrope	yes	normal	**NO**



Ca = (3+2)/ (3+2+2+0) = 71%

	Predicted positive	Predicted negative
Actual positive	TP=3	FN=0
Actual negative	FP=2	TN=2

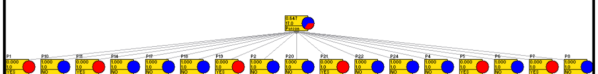
Discussion

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

Discussion about decision trees

→ • How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

Information gain of the "attribute" Person



On training set
- As many values as there are examples
- Each leaf has exactly one example
- E(1/1, 0/1) = 0 (entropy of each leaf is zero)
- The weighted sum of entropies is zero
- The information gain is maximum (as much as the entropy of the entire training set)

On testing set
- The values from the testing set do not appear in the tree

Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?

→ • How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

Data Mining and Knowledge Discovery
Practice notes: Decision trees

Entropy{hard=4, soft=5, none=13}=

= E(4/22, 5/22, 13/22)

= $-\Sigma p_i * \log_2 p_i$

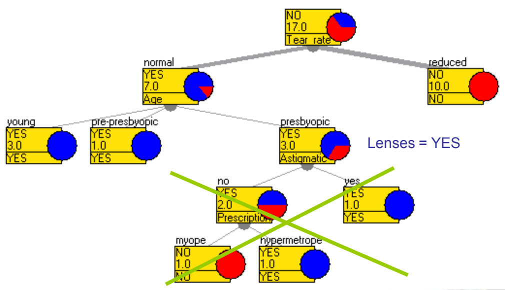
= $-4/22 * \log_2 4/22 - 5/22 * \log_2 5/22 - 13/22 * \log_2 13/22$

= 1.38

Discussion about decision trees

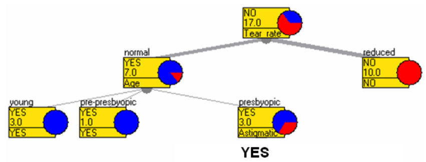
- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- → What would be the classification accuracy of our decision tree if we pruned it at the node Astigmatic?
- What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

Decision tree pruning



Lenses = YES

These two trees are equivalent



Classification accuracy of the pruned tree

Person	Age	Prescription	Astigmatic	Tear_rate	Lenses
P3	young	hypermetrope	no	normal	YES
P9	pre-presbyopic	myope	no	normal	YES
P12	pre-presbyopic	hypermetrope	no	reduced	NO
P13	pre-presbyopic	myope	yes	normal	YES
P15	pre-presbyopic	hypermetrope	yes	normal	NO
P16	pre-presbyopic	hypermetrope	yes	reduced	NO
P23	presbyopic	hypermetrope	yes	normal	NO

Ca = (3+2)/ (3+2+2+0) = 71%

	Predicted positive	Predicted negative
Actual positive	TP=3	FN=0
Actual negative	FP=2	TN=2

Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node Astigmatic?
- → What are the stopping criteria for building a decision tree?
- How would you compute the information gain for a numeric attribute?

Data Mining and Knowledge Discovery
Practice notes: Decision trees

Stopping criteria for building a decision tree

- ID3
  - "Pure" nodes (entropy =0)
  - Out of attributes
- J48 (C4.5)
  - Minimum number of instances in a leaf constraint

31

Discussion about decision trees

- How much is the information gain for the "attribute" Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node Astigmatic?
- What are the stopping criteria for building a decision tree?
- → How would you compute the information gain for a numeric attribute?

32

Information gain of a numeric attribute

Age	Lenses
67	YES
52	YES
63	NO
26	YES
65	NO
23	YES
65	NO
25	YES
26	YES
57	NO
49	NO
23	YES
39	NO
55	NO
53	NO
38	NO
67	YES
54	NO
29	YES
46	NO
44	YES
32	NO
39	NO
45	YES

33

Information gain of a numeric attribute

Age	Lenses
67	YES
52	YES
63	NO
26	YES
65	NO
23	YES
65	NO
25	YES
26	YES
57	NO
49	NO
23	YES
39	NO
55	NO
53	NO
38	NO
67	YES
54	NO
29	YES
46	NO
44	YES
32	NO
39	NO
45	YES

Sort by Age →

Age	Lenses
23	YES
23	YES
25	YES
26	YES
26	YES
29	YES
32	NO
38	NO
39	NO
39	NO
44	YES
45	YES
46	NO
49	NO
52	YES
53	NO
54	NO
55	NO
57	NO
63	NO
65	NO
65	NO
67	YES
67	YES

34

Information gain of a numeric attribute

Age	Lenses
67	YES
52	YES
63	NO
26	YES
65	NO
23	YES
65	NO
25	YES
26	YES
57	NO
49	NO
23	YES
39	NO
55	NO
53	NO
38	NO
67	YES
54	NO
29	YES
46	NO
44	YES
32	NO
39	NO
45	YES

Sort by Age →

Age	Lenses
23	YES
23	YES
25	YES
26	YES
26	YES
29	YES
32	NO
38	NO
39	NO
39	NO
44	YES
45	YES
46	NO
49	NO
52	YES
53	NO
54	NO
55	NO
57	NO
63	NO
65	NO
65	NO
67	YES
67	YES

Define possible splitting points →

Age	Lenses
23	YES
23	YES
25	YES
26	YES
26	YES
29	YES
32	NO
38	NO
39	NO
39	NO
44	YES
45	YES
46	NO
49	NO
52	YES
53	NO
54	NO
55	NO
57	NO
63	NO
65	NO
65	NO
67	YES
67	YES

Information gain of a numeric attribute

Age	Lenses	
23	YES	
23	YES	
25	YES	
26	YES	
26	YES	
29	YES	→ **30.5**
32	NO	
38	NO	
39	NO	
39	NO	→ **41.5**
44	YES	
45	YES	→ **45.5**
46	NO	
49	NO	→ **50.5**
52	YES	→ **52.5**
53	NO	
54	NO	
55	NO	
57	NO	
63	NO	
65	NO	
65	NO	→ **66**
67	YES	
67	YES	

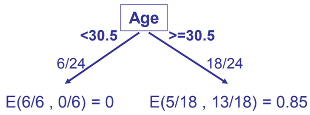
36

Data Mining and Knowledge Discovery
Practice notes: Decision trees

Information gain of a numeric attribute

Age	Lenses
23	YES
23	YES
25	YES
26	YES
26	YES
29	YES
32	NO
38	NO
39	NO
39	NO
44	YES
45	YES
46	NO
49	NO
52	YES
53	NO
54	NO
55	NO
57	NO
63	NO
65	NO
65	NO
67	YES
67	YES

30.5
41.5
45.5
50.5
52.5
66



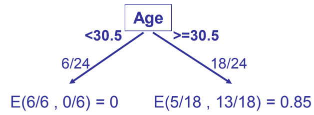
37

Information gain of a numeric attribute

Age	Lenses
23	YES
23	YES
25	YES
26	YES
26	YES
29	YES
32	NO
38	NO
39	NO
39	NO
44	YES
45	YES
46	NO
49	NO
52	YES
53	NO
54	NO
55	NO
57	NO
63	NO
65	NO
65	NO
67	YES
67	YES

30.5
41.5
45.5
50.5
52.5
66

E(S) = E(11/24 , 13/24) = 0.99



InfoGain (S, $Age_{30.5}$)=

= E(S) - $\sum p_v$ E(pv)

= 0.99 – (6/24*0 + 18/24*0.85)

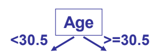
= 0.35

38

Information gain of a numeric attribute

Age	Lenses
23	YES
23	YES
25	YES
26	YES
26	YES
29	YES
32	NO
38	NO
39	NO
39	NO
44	YES
45	YES
46	NO
49	NO
52	YES
53	NO
54	NO
55	NO
57	NO
63	NO
65	NO
65	NO
67	YES
67	YES

30.5
41.5
45.5
50.5
52.5
66



InfoGain (S, $Age_{30.5}$) = 0.35

39

Decision trees

- Many possible decision trees

$$\sum_{i=0}^{k} 2^i (k-i) = -k + 2^{k+1} - 2$$

  - k is the number of binary attributes
- Heuristic search with information gain
- Information gain is short-sighted

40

Trees are shortsighted (1)

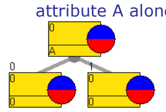
A	B	C	A xor B
1	1	0	0
0	0	1	0
1	0	0	1
0	0	0	0
0	1	0	1
1	1	1	0
1	0	1	1
0	0	1	0
0	1	0	1
0	1	0	1
1	0	1	1
1	1	1	0

- Three attributes: A, B and C
- Target variable is a logical combination attributes A and B
  class = A xor B
- Attribute C is random w.r.t. the target variable

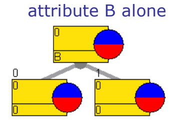
41

Trees are shortsighted (2)

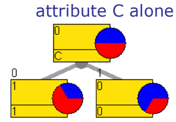
attribute A alone



attribute B alone



attribute C alone



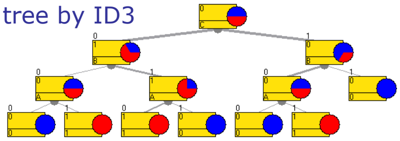
Attribute C has the highest information gain!

42

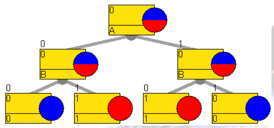
Data Mining and Knowledge Discovery
Practice notes: Decision trees

Trees are shortsighted (3)

- Decision tree by ID3



- The real model behind the data



43

Overcoming shortsightedness of decision trees

- Random forests
  (Breinmann & Cutler, 2001)
  - A random forest is a set of decision trees
  - Each tree is induced from a bootstrap sample of examples
  - For each node of the tree, select among a subset of attributes
  - All the trees vote for the classification
  - See also ensamble learning
- ReliefF for attribute estimation
  (Kononenko el al., 1997)

44