Data Mining and Knowledge Discovery: Practice Notes

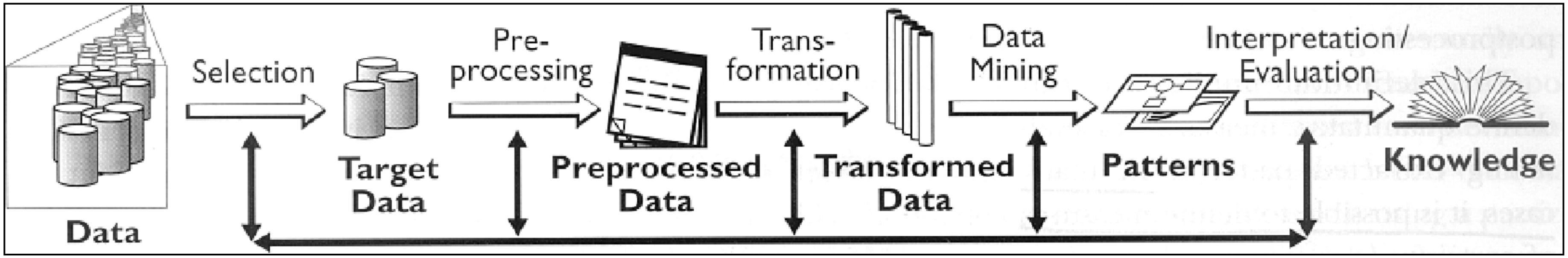
Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

2016/01/06

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Keywords



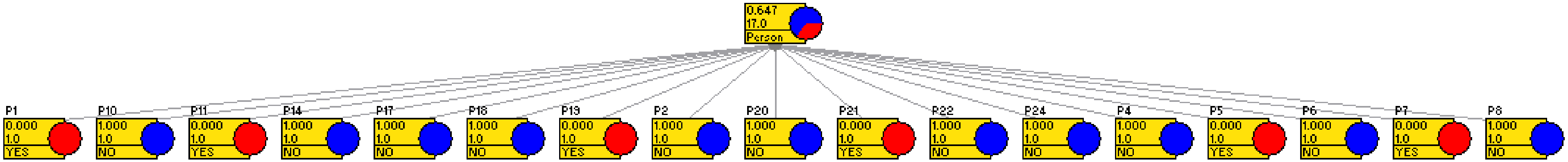
- Data
  - Attribute, example, attribute-value data, target variable, class, discretization
- Algorithms
  - Heuristics vs. exhaustive search, decision tree induction, entropy, information gain, overfitting, Occam's razor, model pruning, naïve Bayes classifier, KNN, association rules, support, confidence,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

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?

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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?

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

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

$= -\sum p_i * \log_2 p_i$

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

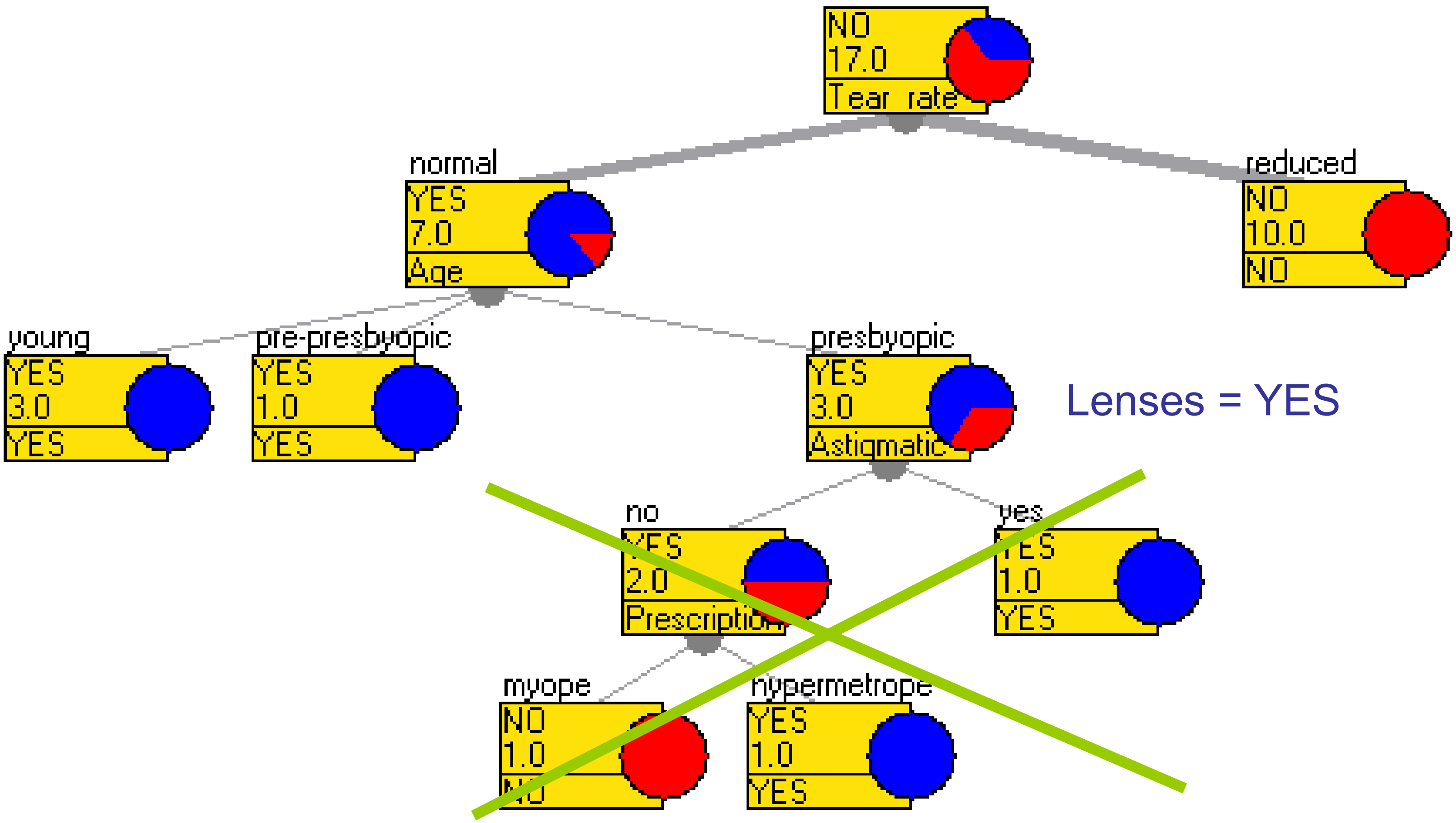
$= 1.38$

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion about decision trees

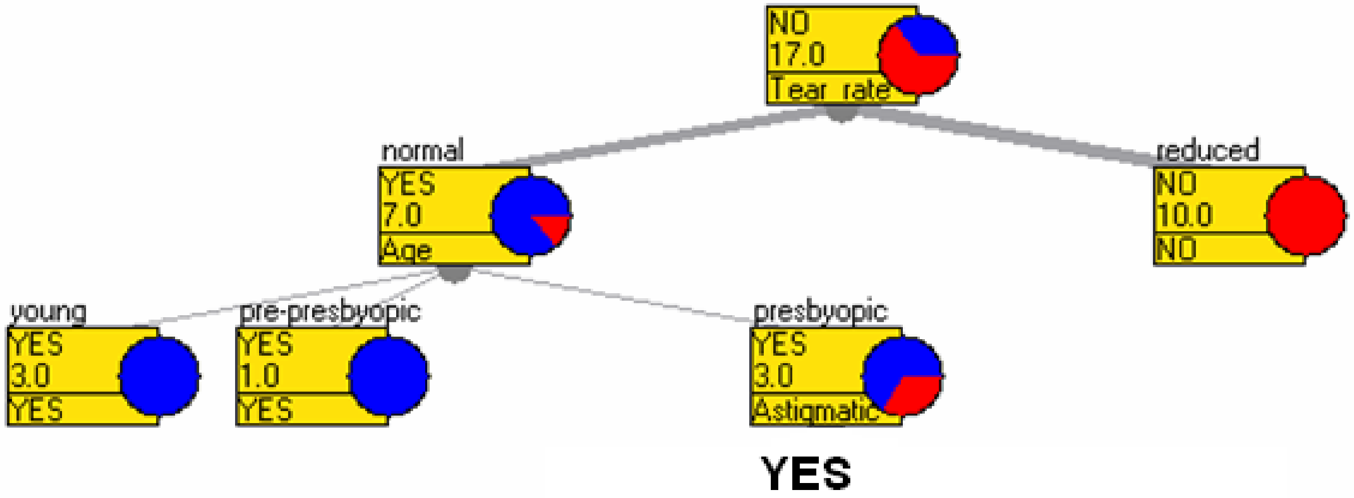
- How much is the information gain for the “attribute” Person? How would it perform on the test set?
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we pruned it at the node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

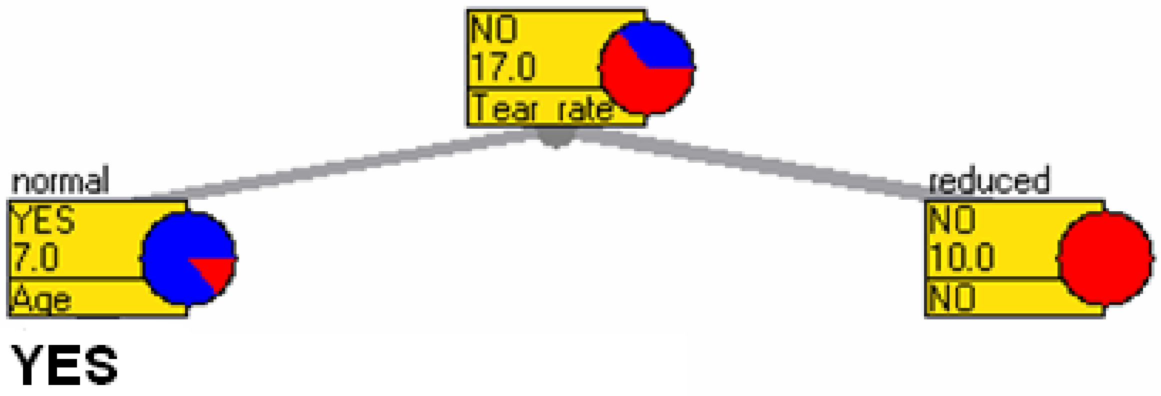
Decision tree pruning



Lenses = YES

These two trees are equivalent



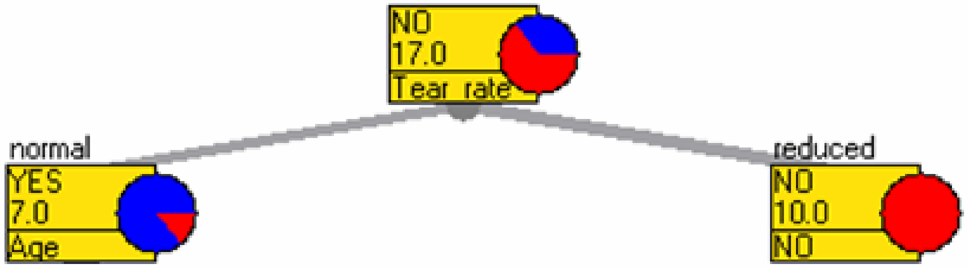


DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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%



YES

	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?

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

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 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

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

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
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**

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

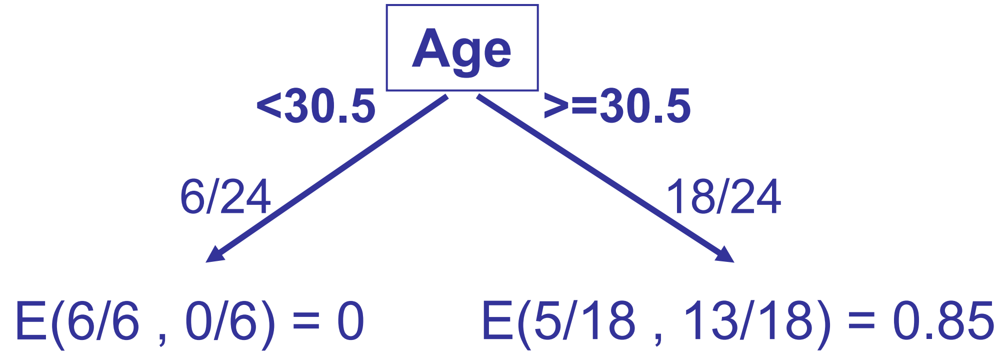
41.5

45.5

50.5

52.5

66



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

Age

<30.5 6/24

>=30.5 18/24

E(6/6 , 0/6) = 0

E(5/18 , 13/18) = 0.85

InfoGain (S, $\mathrm{Age}_{30.5}$)=

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

= 0.99 – (6/24*0 + 18/24*0.85)

= 0.35

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

Age: <30.5 | >=30.5

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

Age: <41.5 | >=41.5

Age: <45.5 | >=45.5

Age: <50.5 | >=50.5

Age: <52.5 | >=52.5

Age: <66 | >=66

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

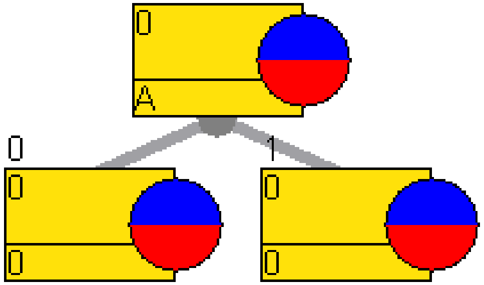
Trees are shortsighted (1)

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

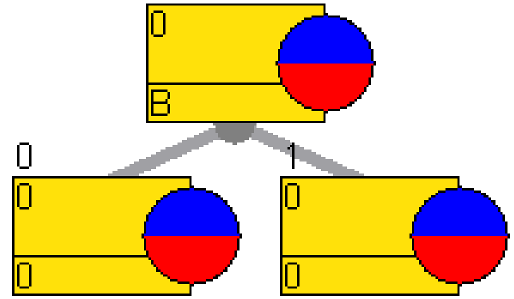
- Three attributes:
  A, B and C
- Target variable is a logical combination attributes A and B
  class = A xor B
- Attribute C is random w.r.t. the target variable

Trees are shortsighted (2)

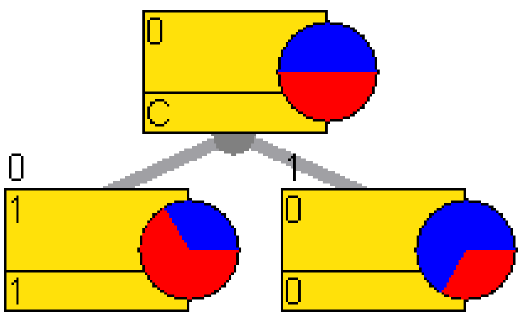
attribute A alone



attribute B alone



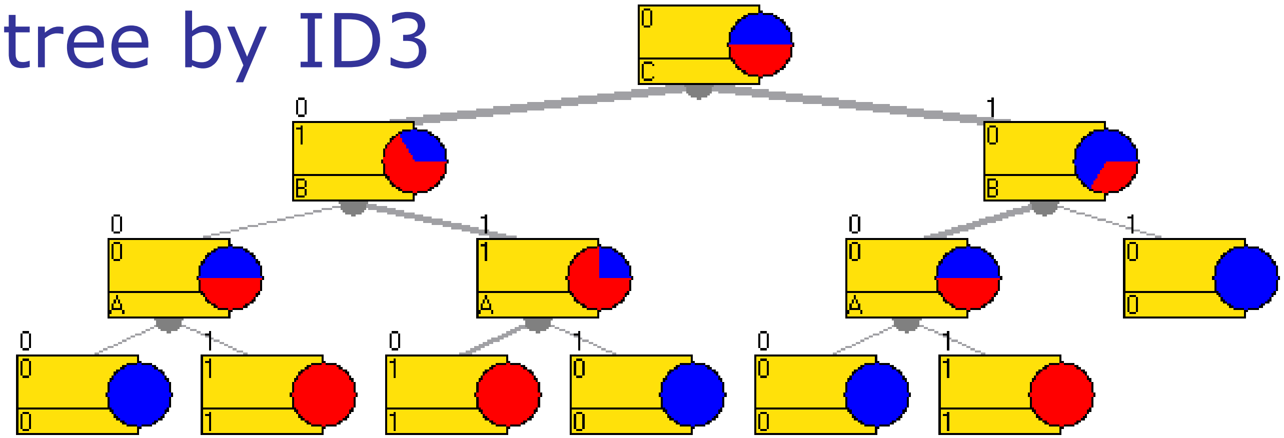
attribute C alone



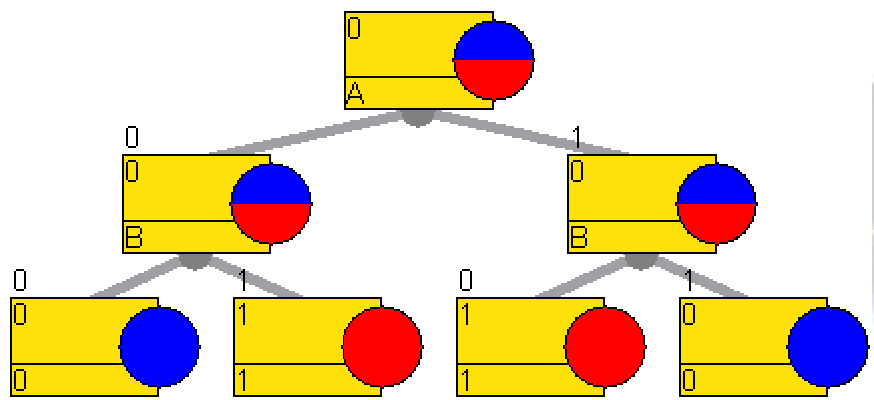
Attribute C has the highest information gain!

Trees are shortsighted (3)

- Decision tree by ID3



- The real model behind the data



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)

Predicting with Naïve Bayes

Given

- Attribute-value data with nominal target variable

Induce

- Build a Naïve Bayes classifier and estimate its performance on new data

Naïve Bayes classifier

$$P(c \mid a_1, a_2, \dots, a_n) = P(c) \prod_i \frac{P(c \mid a_i)}{P(c)}$$

Assumption: conditional independence of attributes given the class.

Will the spider catch these two ants?

- Color = white, Time = night
- Color = black, Size = large, Time = day

Color	Size	Time	Caught
black	large	day	YES
white	small	night	YES
black	small	day	YES
red	large	night	NO
black	large	night	NO
white	large	night	NO

Naïve Bayes classifier -example

Color	Size	Time	Caught
black	large	day	YES
white	small	night	YES
black	small	day	YES
red	large	night	NO
black	large	night	NO
white	large	night	NO

$$v_1 = \text{“}Color = white\text{”}$$

$$v_2 = \text{“}Time = night\text{”}$$

$$c_1 = YES$$

$$c_2 = NO$$

$$p(c_1|v_1, v_2) =$$

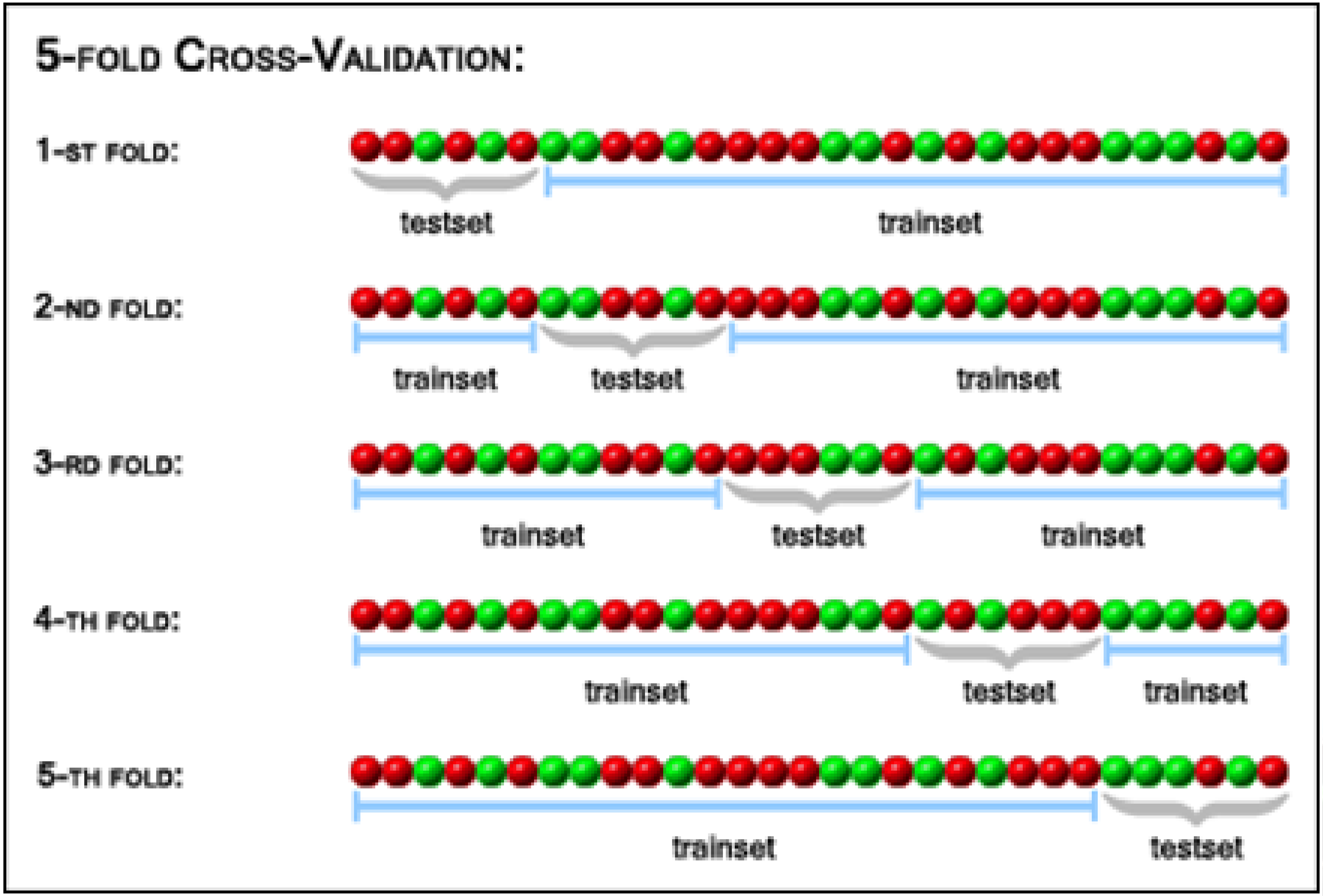
$$p(Caught = YES|Color = white, Time = night) =$$

$$p(Caught = YES) * \frac{p(Caught = YES|Color = white)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = night)}{p(Caught = YES)} =$$

$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4}$$

K-fold cross validation

1. The sample set is partitioned into K subsets ("folds") of about equal size
2. A single subset is retained as the validation data for testing the model (this subset is called the "testset"), and the remaining K - 1 subsets together are used as training data ("trainset").
3. A model is trained on the trainset and its performance (accuracy or other performance measure) is evaluated on the testset
4. Model training and evaluation is repeated K times, with each of the K subsets used exactly once as the testset.
5. The average of all the accuracy estimations obtained after each iteration is the resulting accuracy estimation.



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion

1. Compare naïve Bayes and decision trees (similarities and differences) .
2. Compare cross validation and testing on a different test set.
3. Why do we prune decision trees?
4. What is discretization.