Data Mining and Knowledge Discovery

Knowledge Discovery and Knowledge Management in e-Science

Petra Kralj

Petra.Kralj@ijs.si

Practice, 2007/11/15

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion

- List evaluation methods for classification.
- 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 would have pruned it at he node *Astigmatic*?
- How would you compute the information gain of a numeric attribute?
- Compare the naïve Bayes classifier and decision trees regarding the handling of missing values.
- Compare the naïve Bayes classifier and decision trees regarding numeric attributes.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

List of evaluation methods

- Separate train and test set
- K-fold cross validation
- Leave one out
  - used with very small datasets (few 10 examples)
  - For each example *e*:
    - use *e* as test example and the rest for training
    - Count the correctly classified examples
- Optimistic estimate: test on training set

Discussion

- List evaluation methods for classification.
- 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 would have pruned it at he node *Astigmatic*?
- How would you compute the information gain of a numeric attribute?
- Compare the naïve Bayes classifier and decision trees regarding the handling of missing values.
- Compare the naïve Bayes classifier and decision trees regarding numeric attributes.

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

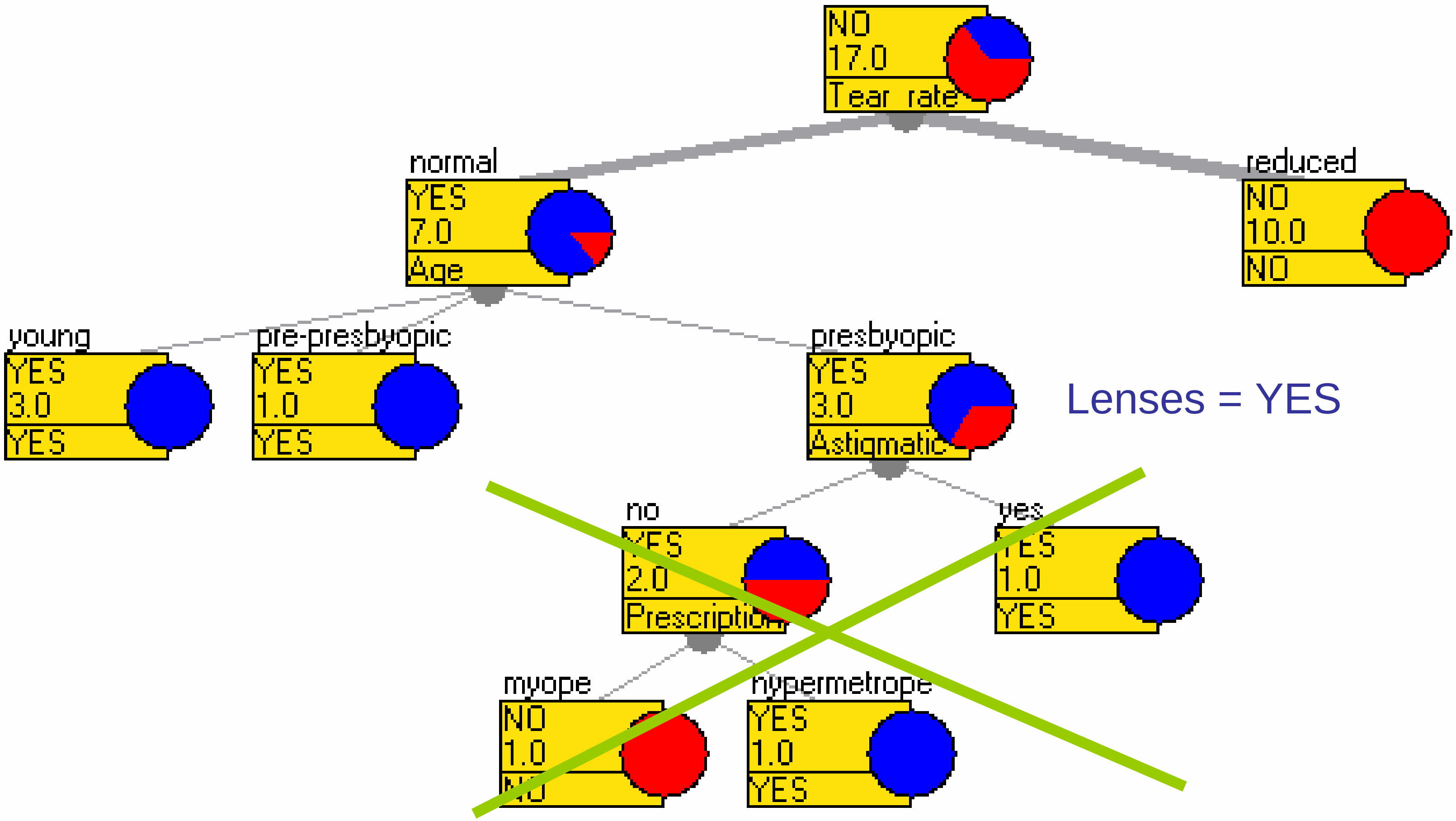
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

LEARNED MEDIA

Discussion

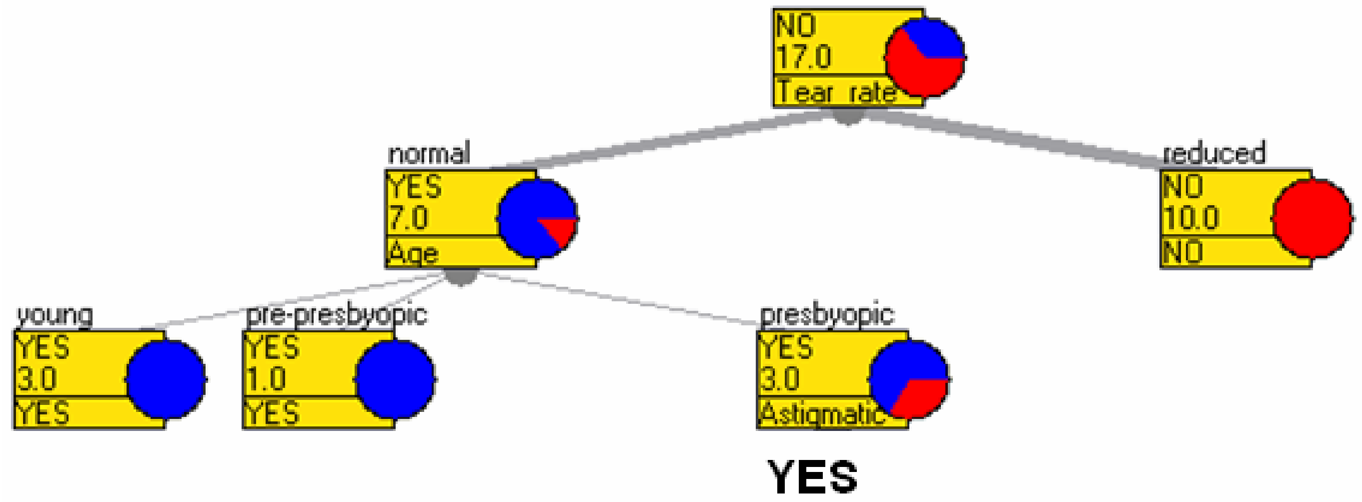
- List evaluation methods for classification.
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- What would be the classification accuracy of our decision tree if we would have pruned it at he node *Astigmatic*?
- How would you compute the information gain of a numeric attribute?
- Compare the naïve Bayes classifier and decision trees regarding the handling of missing values.
- Compare the naïve Bayes classifier and decision trees regarding numeric attributes.

Decision tree



Lenses = YES

These two trees are equivalent



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

LEARNED MEDIA

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) = 0,71%

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

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion

- List evaluation methods for classification.
- 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 would have pruned it at he node *Astigmatic*?
- How would you compute the information gain of a numeric attribute?
- Compare the naïve Bayes classifier and decision trees regarding the handling of missing values.
- Compare the naïve Bayes classifier and decision trees regarding numeric attributes.

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

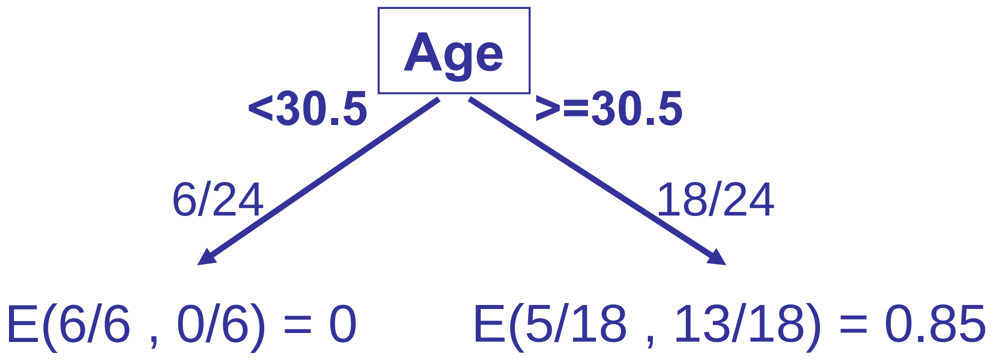
41.5

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

Candidate split points:

- 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): E(6/6 , 0/6) = 0
- **>=30.5** (18/24): E(5/18 , 13/18) = 0.85

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

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

= 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

Discussion

- List evaluation methods for classification.
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- How would you compute the information gain of a numeric attribute?
- What would be the classification accuracy of our decision tree if we would have pruned it at he node *Astigmatic*?
- Compare the naïve Bayes classifier and decision trees regarding the handling of missing values.
- Compare the naïve Bayes classifier and decision trees regarding numeric attributes.

Handling missing values: Naïve Bayes

Will the spider catch these two ants?

- Color = white, Time = night ← **missing value Size**
- Color = black, Size = large, Time = day

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

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

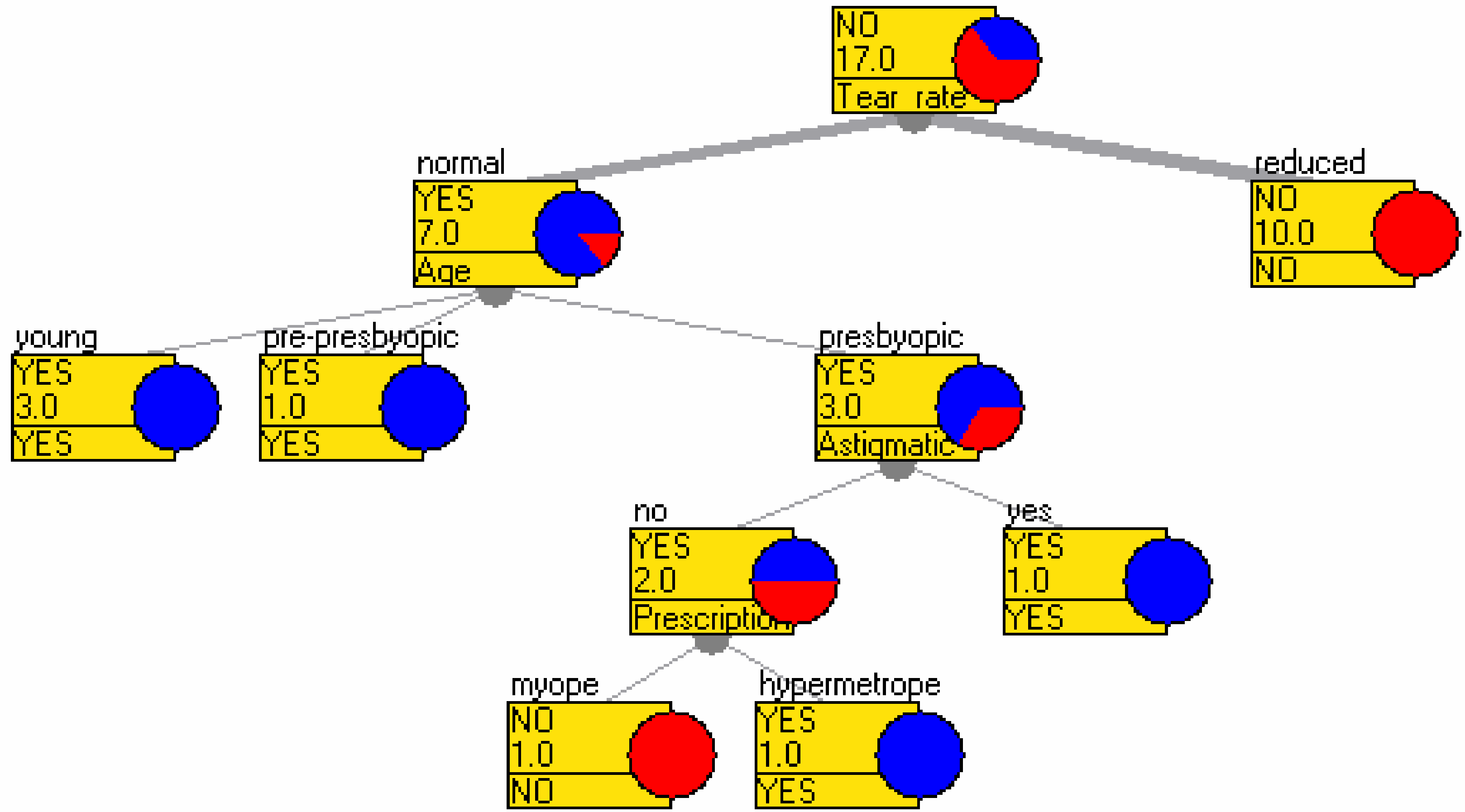
$$p(Caught = YES) * \frac{p(Caught = YES|Color = white)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = night)}{p(Caught = YES)} =$$

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

Naïve Bayes uses all the available information!

Handling missing values: Decision trees - 1

Age	Prescription	Astigmatic	Tear_Rate
?	hypermetrope	no	normal
pre-presbyopic	myope	?	normal



Handling missing values: Decision trees - 2

Algorithm **ID3**: does not handle missing values

Algorithm **C4.5** (J48)deals with two problems:

- Missing values in **train** data:
  - Missing values are not used in gain and entropy calculations
- Missing values in **test** data:
  - A missing **continuous** value is replaced with the median of the training set
  - A missing **categorical** values is replaced with the most frequent value

Discussion

- List evaluation methods for classification.
- How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
- How would you compute the information gain of a numeric attribute?
- What would be the classification accuracy of our decision tree if we would have pruned it at he node *Astigmatic*?
- Compare the naïve Bayes classifier and decision trees regarding the handling of missing values.
- Compare the naïve Bayes classifier and decision trees regarding numeric attributes.

Continuous attributes: decision trees & naïve bayes

- Decision trees **ID3** algorithm: does not handle continuous attributes → data need to be discretized
- Decision trees **C4.5** (J48 in Weka) algorithm: deals with continuous attributes as shown earlier
- **Naïve Bayes**: does not handle continuous attributes → data need to be discretized