Data Mining and Knowledge Discovery

Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

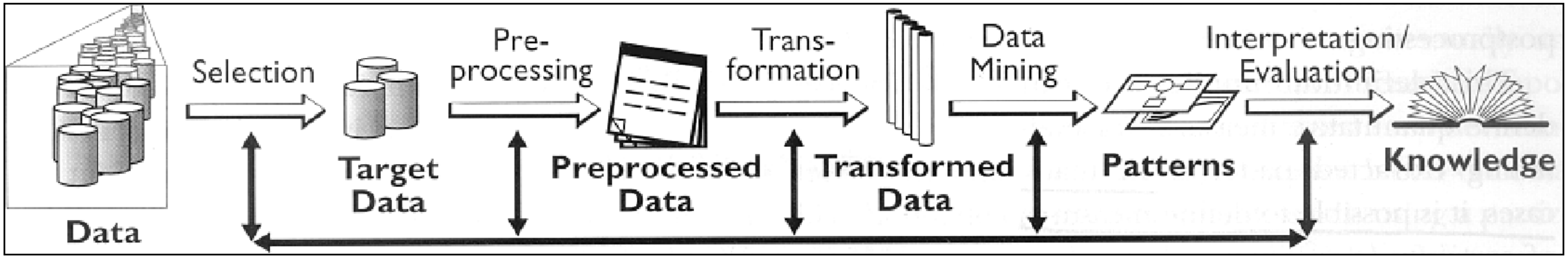
2011/11/08

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- Prof. Lavrač:
  - Data mining overview
  - Advanced topics

- Dr. Kralj Novak
  - Data mining basis

Keywords



- Data
  - Attribute, example, target variable, class, train set, test set, attribute-value data, market basket data
- Data mining
  - 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, heuristics vs. exhaustive search
- Evaluation
  - Accuracy, confusion matrix, cross validation, ROC space, error, leave-one-out

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Practice plan

- 2011/11/08: Predictive data mining 1
  - Decision trees
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - Hands on Weka 1: just a taste
- 2011/11/22: Predictive data mining 2
  - Discussion on decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers 2: Cross validation
  - Numeric prediction
  - Hands on Weka 2
- 2011/11/29: Descriptive data mining
  - Association rules
  - Descriptive data mining in Weka
  - Discussion about seminars and exam
  - Hands on Weka 3
- 2011/12/20: Written exam, seminar proposal presentations
- 2012/1/24 : Data mining seminar presentations

Decision tree induction

Given

- Attribute-value data with nominal target variable

Induce

- A decision tree and estimate its performance on new data

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

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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

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

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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!

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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**

Information gain

How much information do we gain by splitting the set S according to attribute A?

number of examples in the subset S_v

(probability of the branch)

set S

attribute A

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

entropy of set S

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) =

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

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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

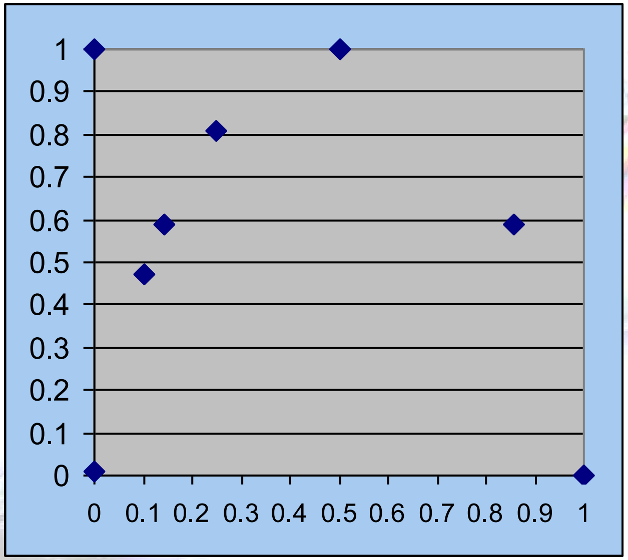
  E (1/4 , 3/4) = 0.81

  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

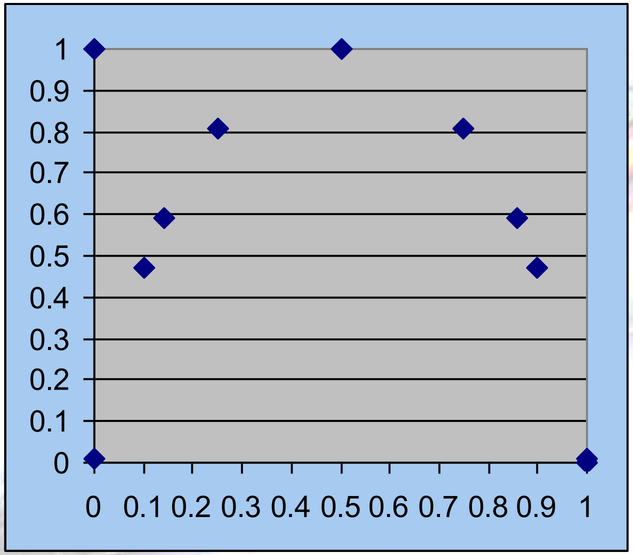
  E (1/4 , 3/4) = 0.81

  E (1/7 , 6/7) = 0.59

  E (6/7 , 1/7) = 0.59

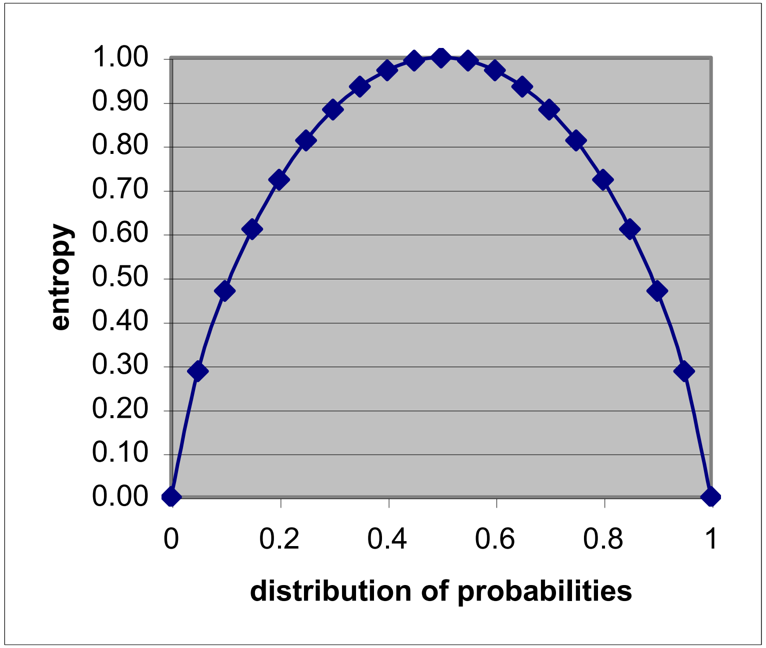
  E (0.1 , 0.9) = 0.47

  E (0.001 , 0.999) = 0.01



Entropy and information gain

probability of class 1	probability of class 2	
p_1	$p_2 = 1-p_1$	**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"

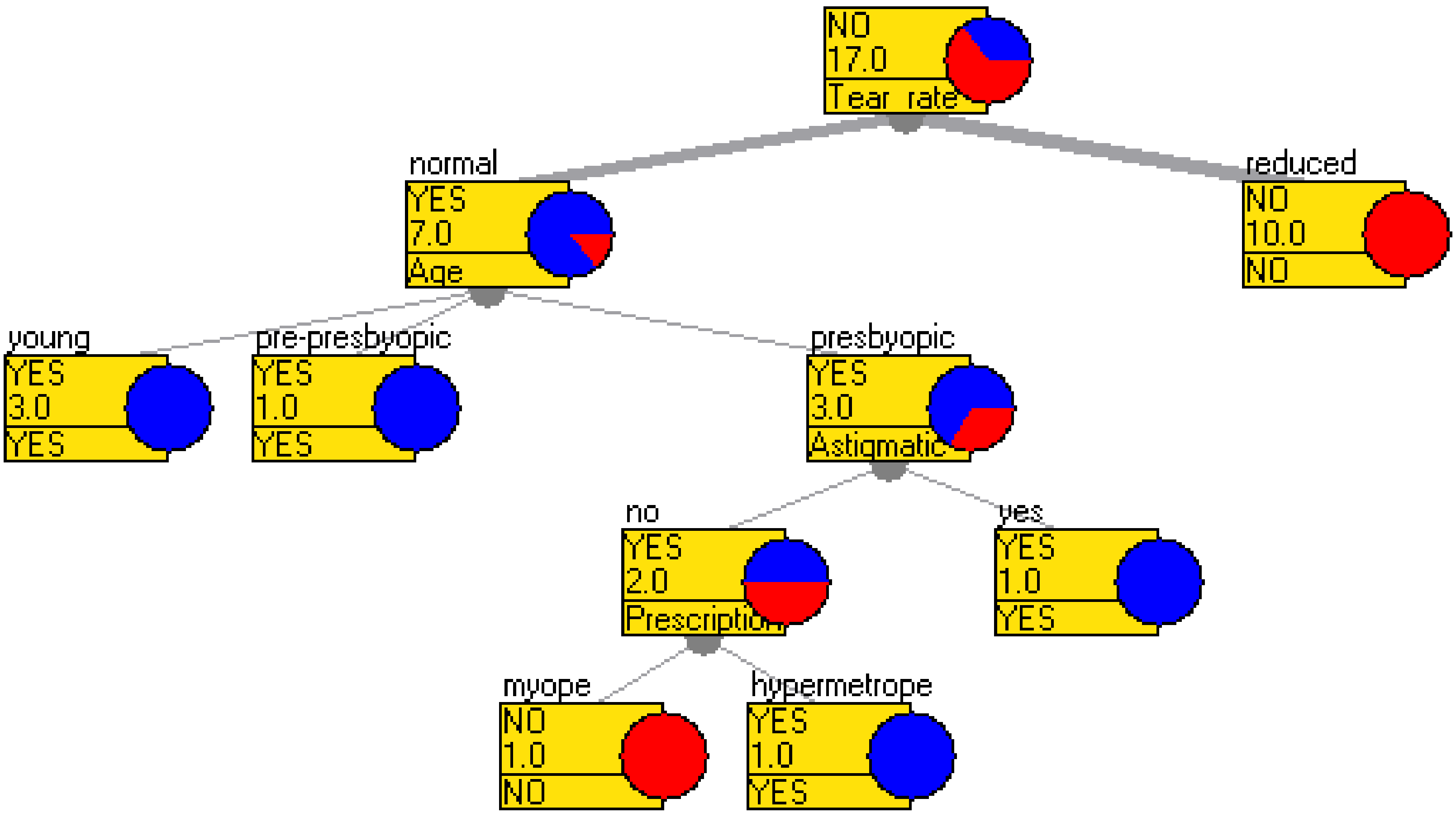
attribut A

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

set S

number of examples in set S

Decision tree



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Confusion matrix

predicted

actual

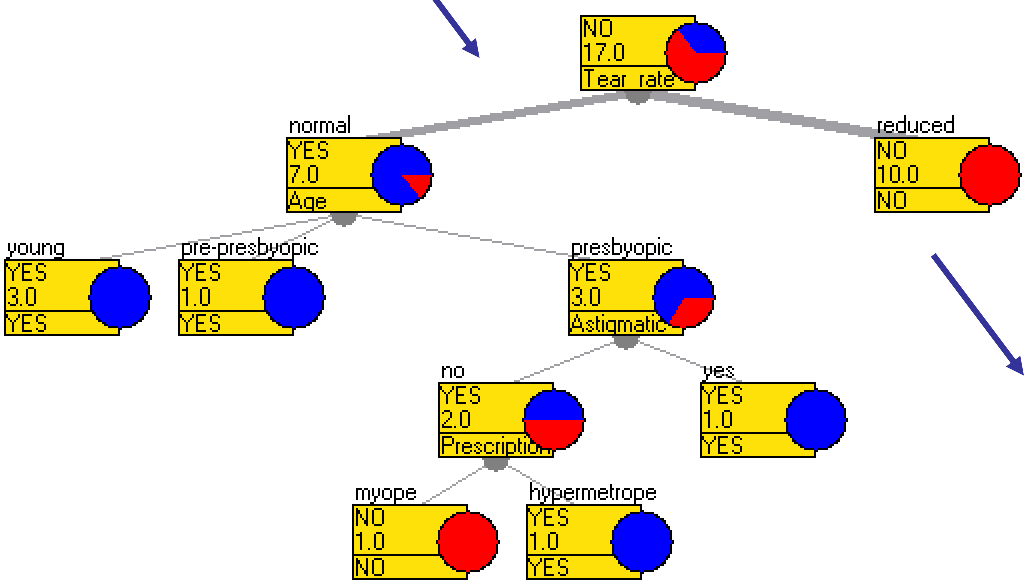
	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

- 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)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

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

TECHNOLOGIES
Jožef Stefan Institute

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 possible stopping criteria for building decision trees?
- In what circumstances is it impossible to achieve pure leaves in a decision tree?