


Data Mining and Knowledge Discovery

Practice notes: Decision trees

Data Mining and Knowledge Discovery: Practice Notes

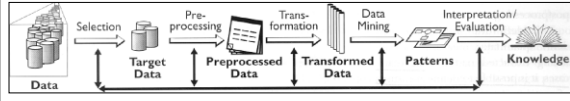
dr. Petra Kralj Novak
Petra.Kralj.Novak@ijs.si
 2016/01/05




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



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, naive 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




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	hypermetropic	no	normal	YES
P4	young	hypermetropic	no	reduced	NO
P5	young	myope	yes	normal	YES
P6	young	myope	yes	reduced	NO
P7	young	hypermetropic	yes	normal	YES
P8	young	hypermetropic	yes	reduced	NO
P9	pre-presbyopic	myope	no	normal	YES
P10	pre-presbyopic	myope	no	reduced	NO
P11	pre-presbyopic	hypermetropic	no	normal	YES
P12	pre-presbyopic	hypermetropic	no	reduced	NO
P13	pre-presbyopic	myope	yes	normal	YES
P14	pre-presbyopic	myope	yes	reduced	NO
P15	pre-presbyopic	hypermetropic	yes	normal	NO
P16	pre-presbyopic	hypermetropic	yes	reduced	NO
P17	presbyopic	myope	no	normal	NO
P18	presbyopic	myope	no	reduced	NO
P19	presbyopic	hypermetropic	no	normal	YES
P20	presbyopic	hypermetropic	no	reduced	NO
P21	presbyopic	myope	yes	normal	YES
P22	presbyopic	myope	yes	reduced	NO
P23	presbyopic	hypermetropic	yes	normal	NO
P24	presbyopic	hypermetropic	yes	reduced	NO

classes = values of the (nominal) target variable



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 $\text{Gain}(S, A)$
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 S_i

Test the model on the test set T

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



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 S_i

Test the model on the test set T

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

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

- 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

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

- Calculate the following entropies:

- $E(0,1) = 0$
- $E(1/2, 1/2) = 1$
- $E(1/4, 3/4) = 0.81$
- $E(1/7, 6/7) = 0.59$
- $E(6/7, 1/7) = 0.59$
- $E(0.1, 0.9) = 0.47$
- $E(0.001, 0.999) = 0.01$

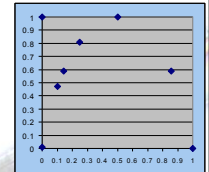


Entropy

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

- Calculate the following entropies:

- $E(0,1) = 0$
- $E(1/2, 1/2) = 1$
- $E(1/4, 3/4) = 0.81$
- $E(1/7, 6/7) = 0.59$
- $E(6/7, 1/7) = 0.59$
- $E(0.1, 0.9) = 0.47$
- $E(0.001, 0.999) = 0.01$

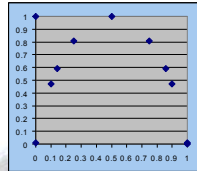


Entropy

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

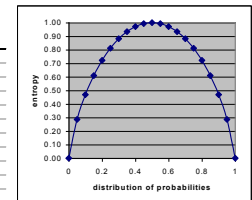
- Calculate the following entropies:

- $E(0,1) = 0$
- $E(1/2, 1/2) = 1$
- $E(1/4, 3/4) = 0.81$
- $E(1/7, 6/7) = 0.59$
- $E(6/7, 1/7) = 0.59$
- $E(0.1, 0.9) = 0.47$
- $E(0.001, 0.999) = 0.01$



Entropy and information gain

probability of class 1	probability of class 2	entropy $E(p_1, p_2) = -p_1 \cdot \log_2(p_1) - p_2 \cdot \log_2(p_2)$
p_1	$p_2 = 1 - p_1$	
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



$$Gain(S, A) = E(S) - \sum_{\text{val } \in \text{values}(A)} \frac{|S_{\text{val}}|}{|S|} E(S_{\text{val}})$$

Labels: attribute A, set S, number of examples in the subset, probability of the "branch", 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)

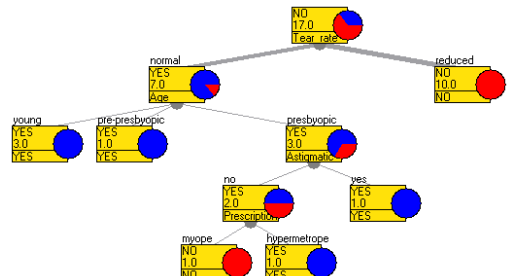
Induce a decision tree on training set S:

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 S_i

Test the model on the test set T



Decision tree



Data Mining and Knowledge Discovery

Practice notes: Decision trees

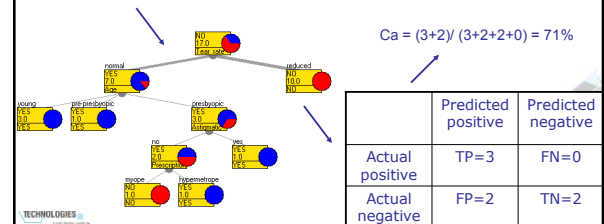
Confusion matrix

		predicted	
		Predicted positive	Predicted negative
actual	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
 - $= \#(\text{correctly classified examples}) / \#(\text{all examples})$
 - $= (TP+TN) / (TP+TN+FP+FN)$

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



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?