Data Mining and Knowledge Discovery: Practice Notes

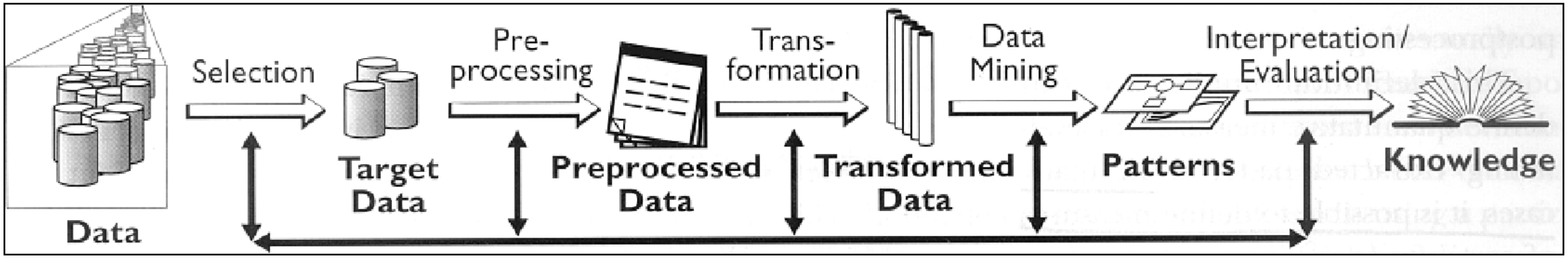
Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

2014/12/9

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Keywords



- Data
  - Attribute, example, attribute-value data, target variable, class, discretization
- Data mining
  - 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

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Practice plan

- 2014/11/11: Predictive data mining
  - Decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - Hands on Weka: Predictive data mining
- 2014/12/9: Numeric prediction and descriptive data mining
  - Discussion on classification
  - Numeric prediction and evaluation in Weka
  - Association rules
  - Hands on Weka: Numeric prediction
  - Hands on Weka: Descriptive data mining
  - Discussion about seminars and exam

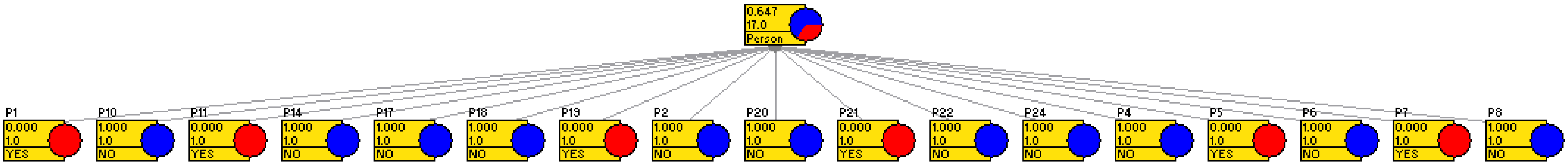
- 2014/12/16: Written exam, seminar proposal discussion
- 2014/1/21: Clowdflows platform and data mining seminar presentations

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

Discussion

→ 1. How much is the information gain for the "attribute" Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

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

1. How much is the information gain for the “attribute” Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

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

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

$= -\Sigma p_i * \log_2 p_i$

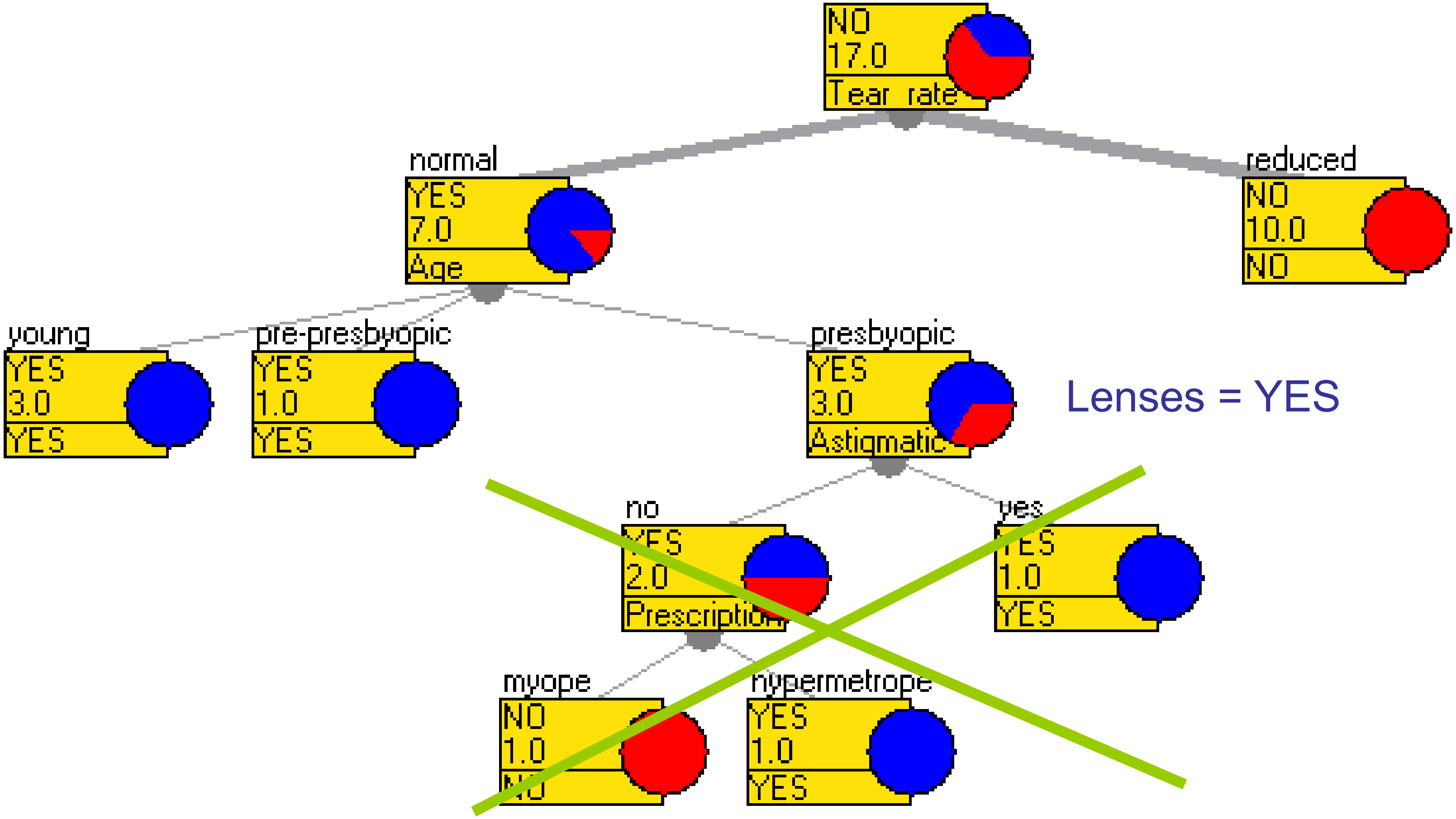
$= -4/22 * \log_2 4/22 - 5/22 * \log_2 5/22 - 13/22 * \log_2 13/22$

$= 1.38$

Discussion

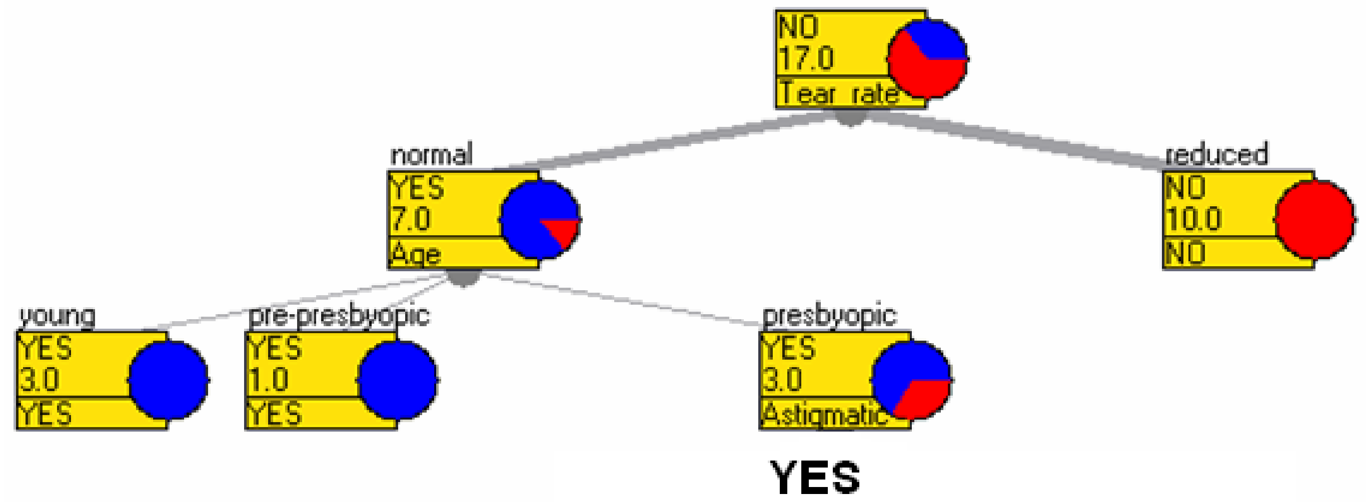
1. How much is the information gain for the "attribute" Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

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%

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

Discussion

1. How much is the information gain for the “attribute” Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

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

1. How much is the information gain for the "attribute" Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

Discussion

1. How much is the information gain for the “attribute” Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

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

DEPA
KN
TECHN

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

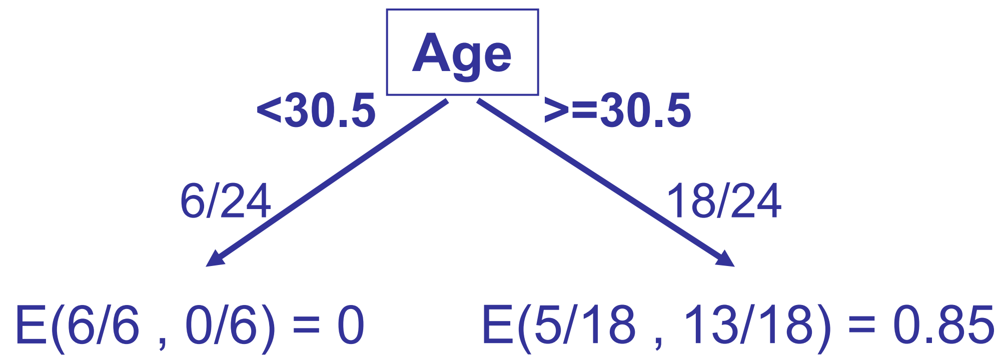
41.5

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 → 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

1. How much is the information gain for the “attribute” Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

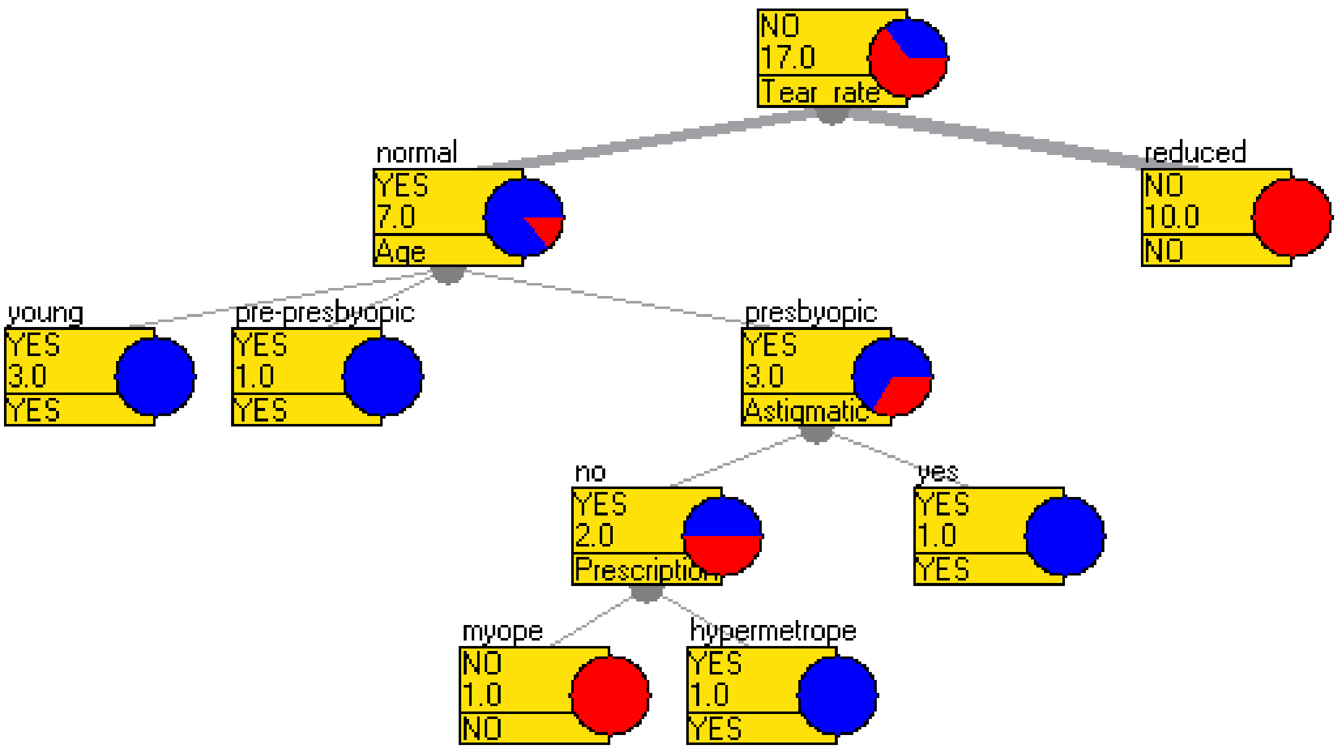
Comparison of naïve Bayes and decision trees

- Similarities
  - Classification
  - Same evaluation

- Differences
  - Missing values
  - Numeric attributes
  - Interpretability of the model

Comparison of naïve Bayes and decision trees: Handling missing values

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



Comparison of naïve Bayes and decision trees: Handling missing values

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

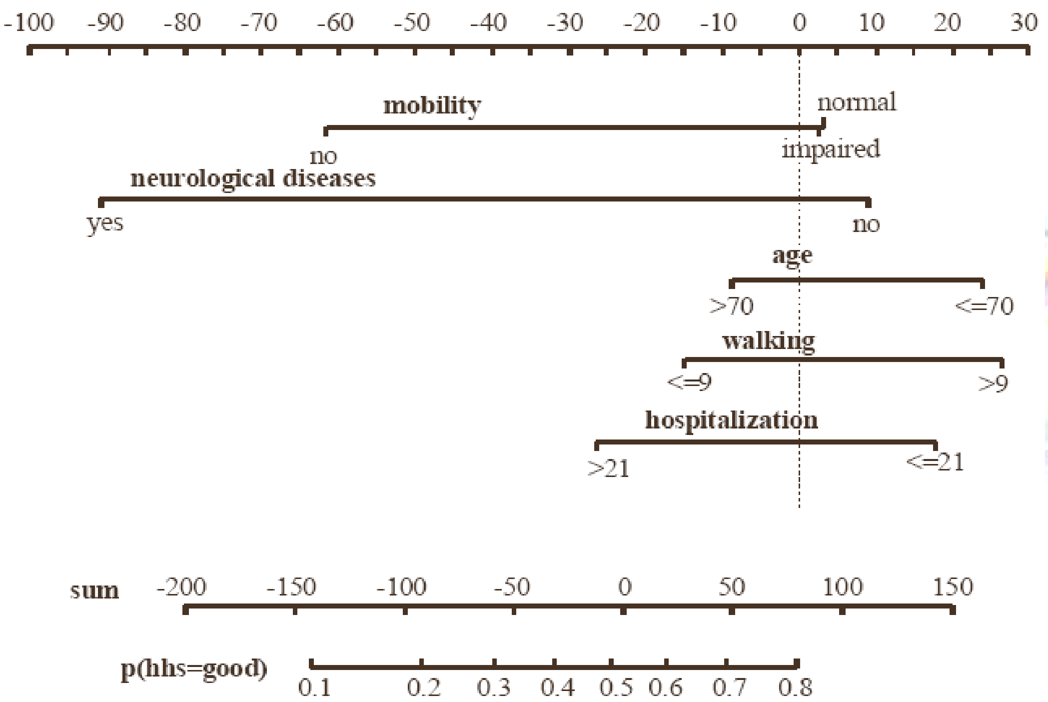
Comparison of naïve Bayes and decision trees: numeric attributes

- 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

(some implementations do handle)

Comparison of naïve Bayes and decision trees: Interpretability

- Decision trees are easy to understand and interpret (if they are of moderate size)
- Naïve bayes models are of the "black box type".
- Naïve bayes models have been visualized by nomograms.



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

Discussion

1. How much is the information gain for the "attribute" Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. → Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Comparison of cross validation and testing on a separate test set

- Both are methods for evaluating predictive models.
- Testing on a separate test set is simpler since we split the data into two sets: one for training and one for testing. We evaluate the model on the test data.
- Cross validation is more complex: It repeats testing on a separate test *n* times, each time taking 1/n of different data examples as test data. The evaluation measures are averaged over all testing sets therefore the results are more reliable.

Discussion

1. How much is the information gain for the “attribute” Person? How would it perform on the test set?
2. How do we compute entropy for a target variable that has three values? Lenses = {hard=4, soft=5, none=13}
3. What would be the classification accuracy of our decision tree if we pruned it at the node *Astigmatic*?
4. What are the stopping criteria for building a decision tree?
5. Why do we prune decision trees?
6. How would you compute the information gain for a numeric attribute?
7. Compare naïve Bayes and decision trees (similarities and differences) .
8. Can KNN be used for classification tasks?
9. Compare KNN and Naïve Bayes.
10. Compare cross validation and testing on a separate test set.
11. List 3 numeric prediction methods.
12. What is discretization.

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)

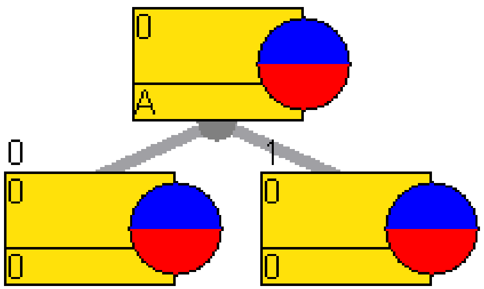
A	B	C	A xor B
1	1	0	0
0	0	1	0
1	0	0	1
0	0	0	0
0	1	0	1
1	1	1	0
1	0	1	1
0	0	1	0
0	1	0	1
0	1	0	1
1	0	1	1
1	1	1	0

- Three attributes:
  A, B and C
- Target variable is a logical combination attributes A and B

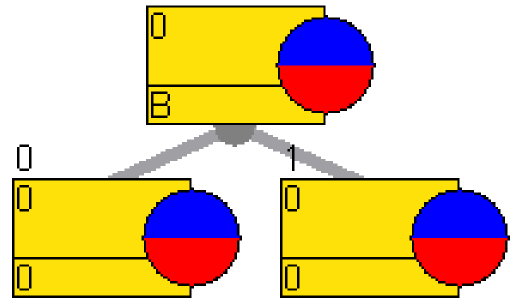
  class = A xor B
- Attribute C is random w.r.t. the target variable

Trees are shortsighted (2)

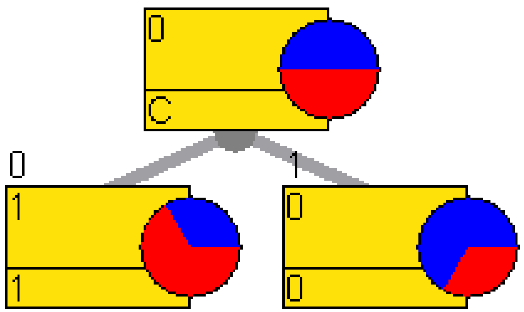
attribute A alone



attribute B alone



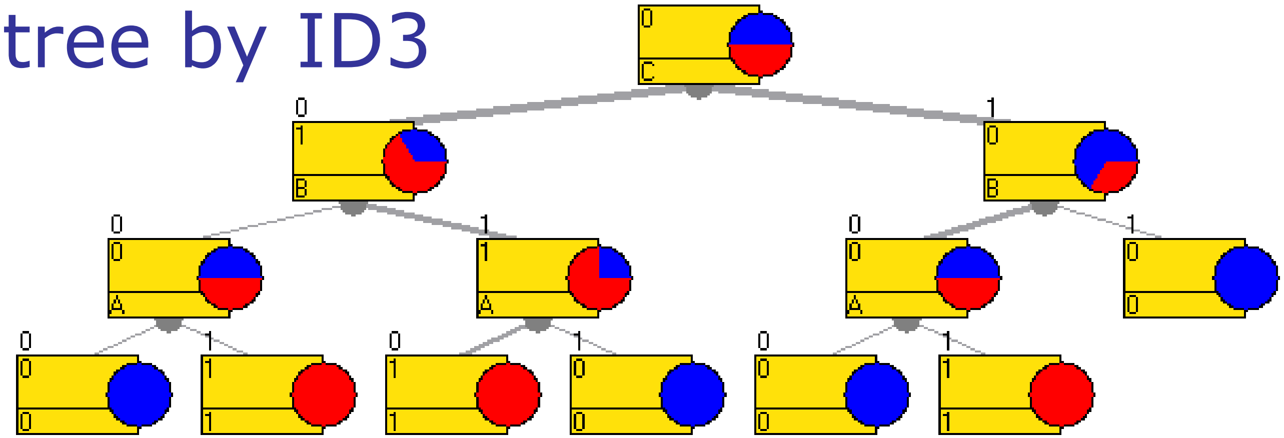
attribute C alone



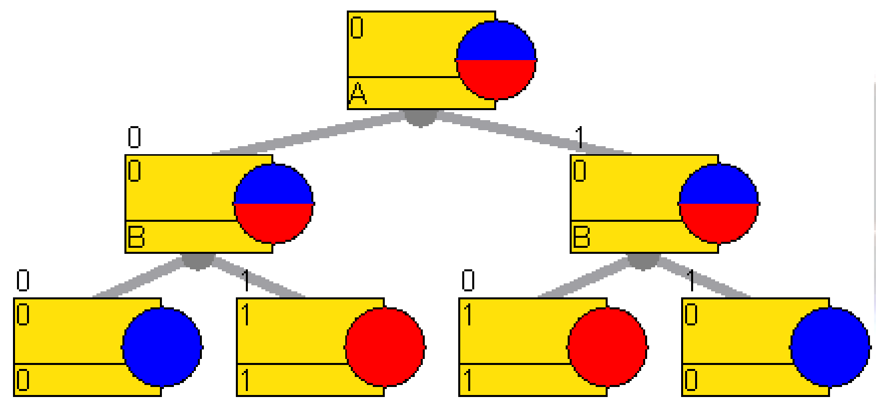
Attribute C has the highest information gain!

Trees are shortsighted (3)

- Decision tree by ID3



- The real model behind the data

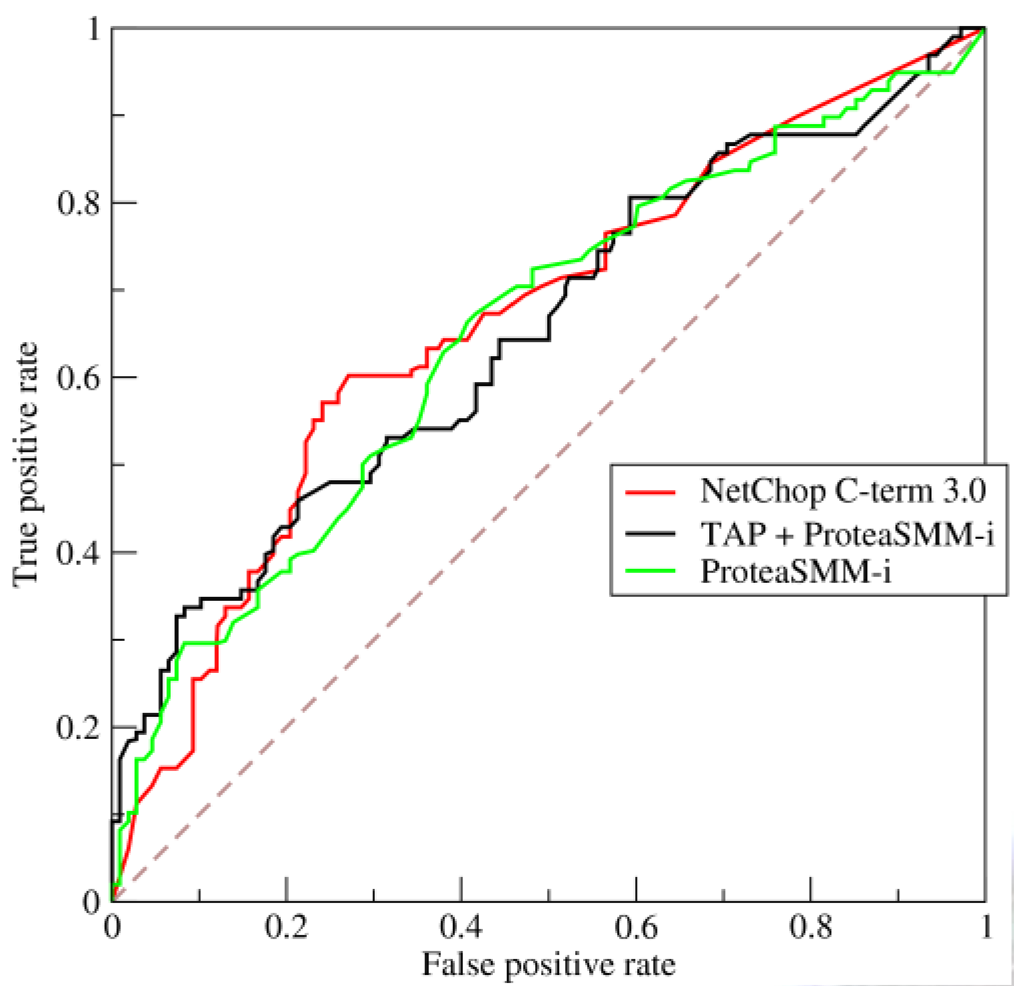


Overcoming shortsightedness of decision trees

- 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)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ROC - Receiver Operating Characteristic



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Practice plan

- 2014/11/11: Predictive data mining
  - Decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers 1: separate test set, confusion matrix, classification accuracy
  - Hands on Weka: Predictive data mining
- 2014/12/9: Numeric prediction and descriptive data mining
  - Discussion on classification
  - Numeric prediction and evaluation in Weka
  - Association rules
  - Hands on Weka: Numeric prediction
  - Hands on Weka: Descriptive data mining
  - Discussion about seminars and exam

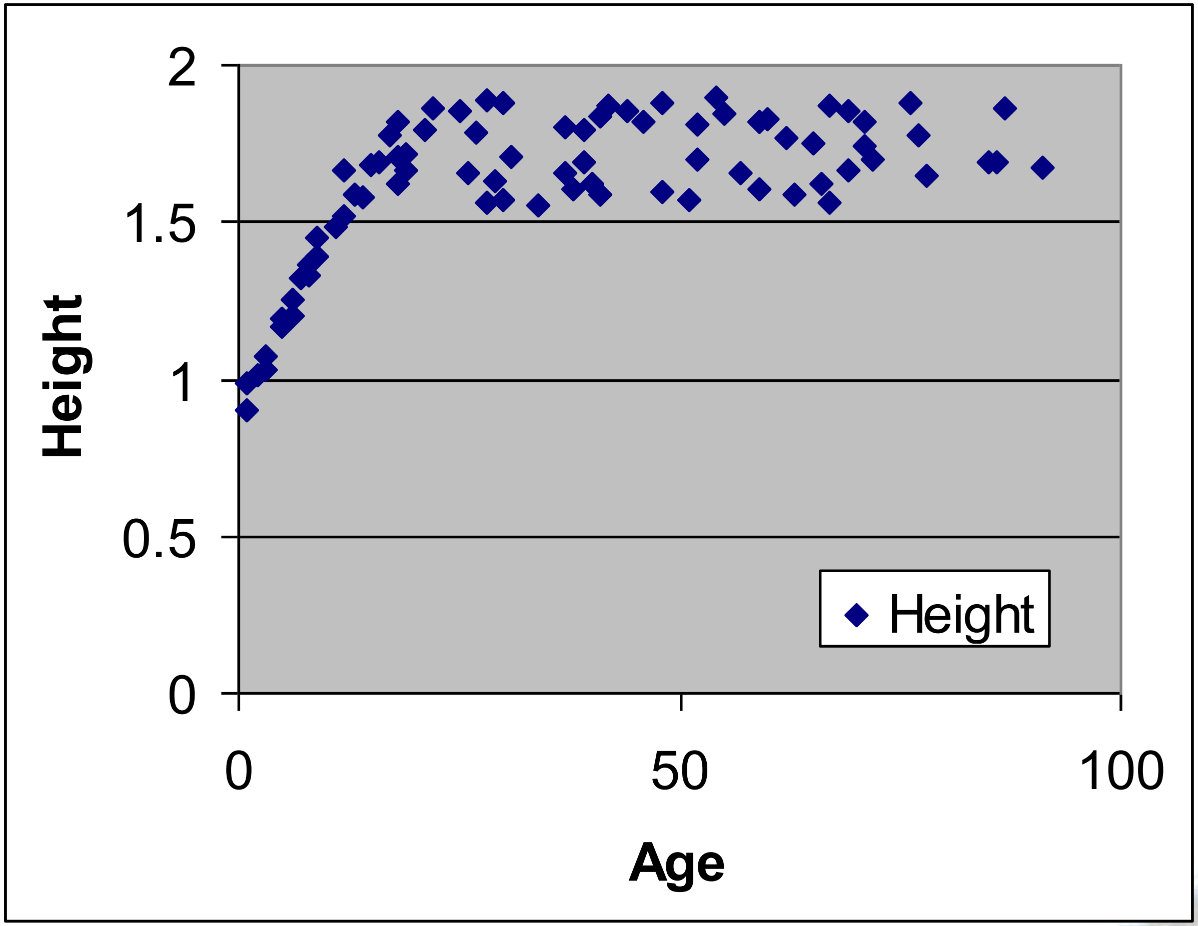
- 2014/12/16: Written exam, seminar proposal discussion
- 2014/1/21: Clowdflows platform and data mining seminar presentations

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

Numeric prediction

Example

- data about 80 people: Age and Height



Age	Height
3	1.03
5	1.19
6	1.26
9	1.39
15	1.69
19	1.67
22	1.86
25	1.85
41	1.59
48	1.60
54	1.90
71	1.82
...	...

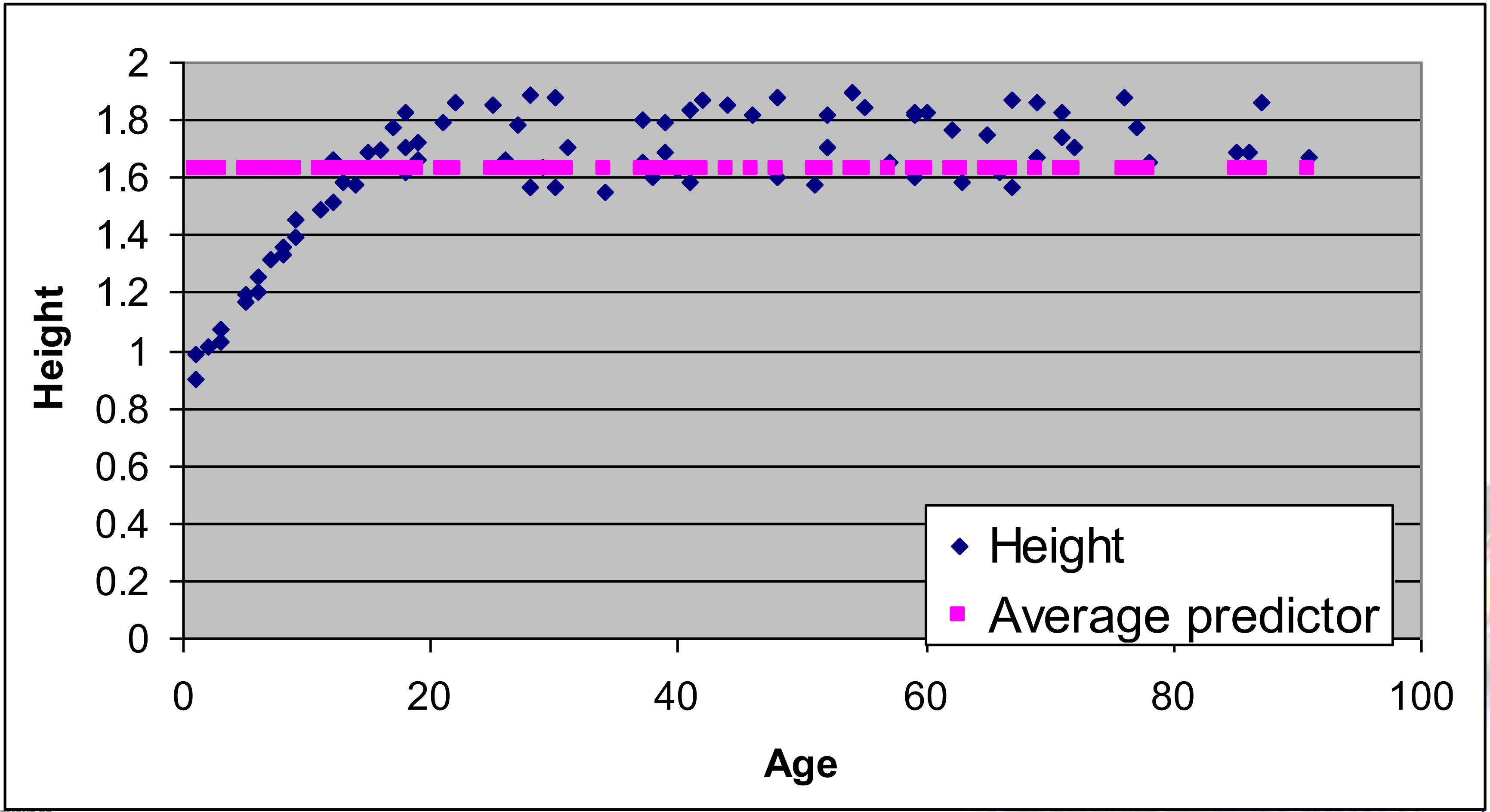
Test set

Age	Height
2	0.85
10	1.4
35	1.7
70	1.6

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Baseline numeric predictor

- Average of the target variable



Baseline predictor: prediction

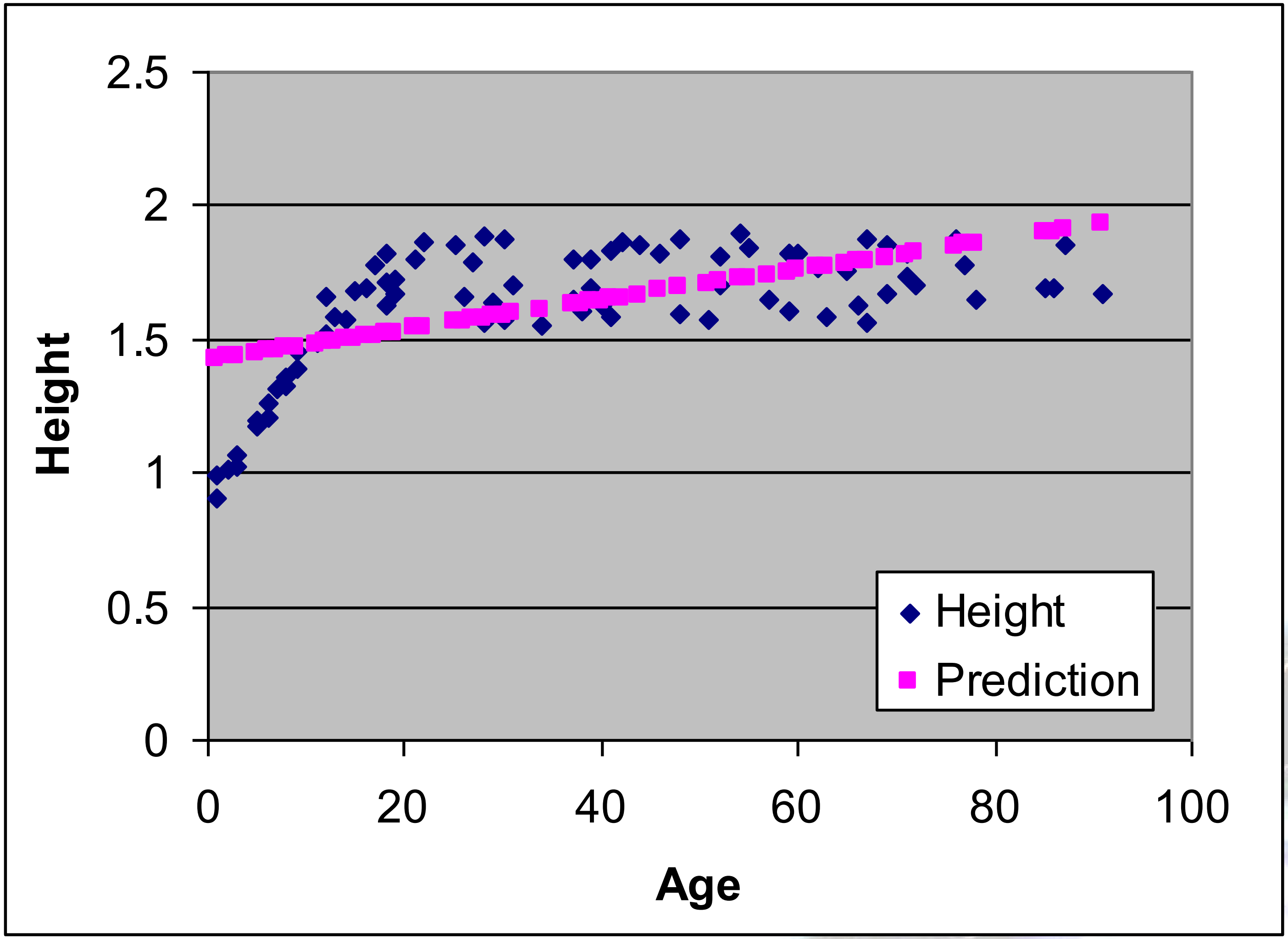
Average of the target variable is 1.63

Age	Height	Baseline
2	0.85	
10	1.4	
35	1.7	
70	1.6	

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Linear Regression Model

Height = 0.0056 * Age + 1.4181

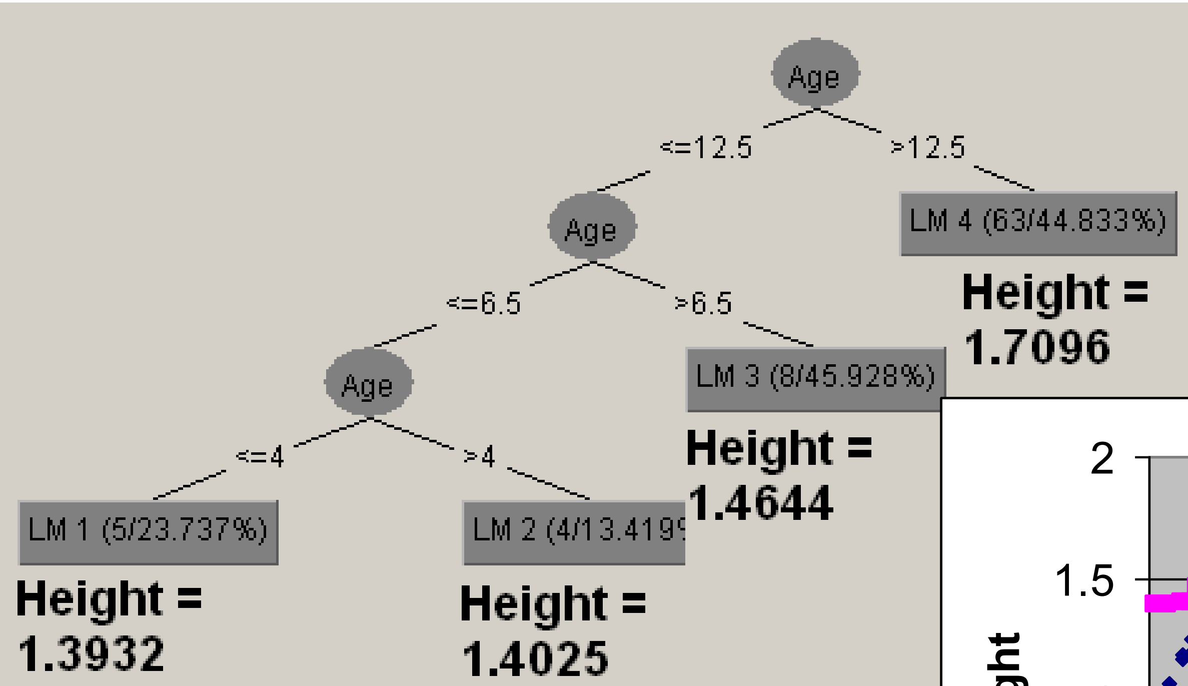


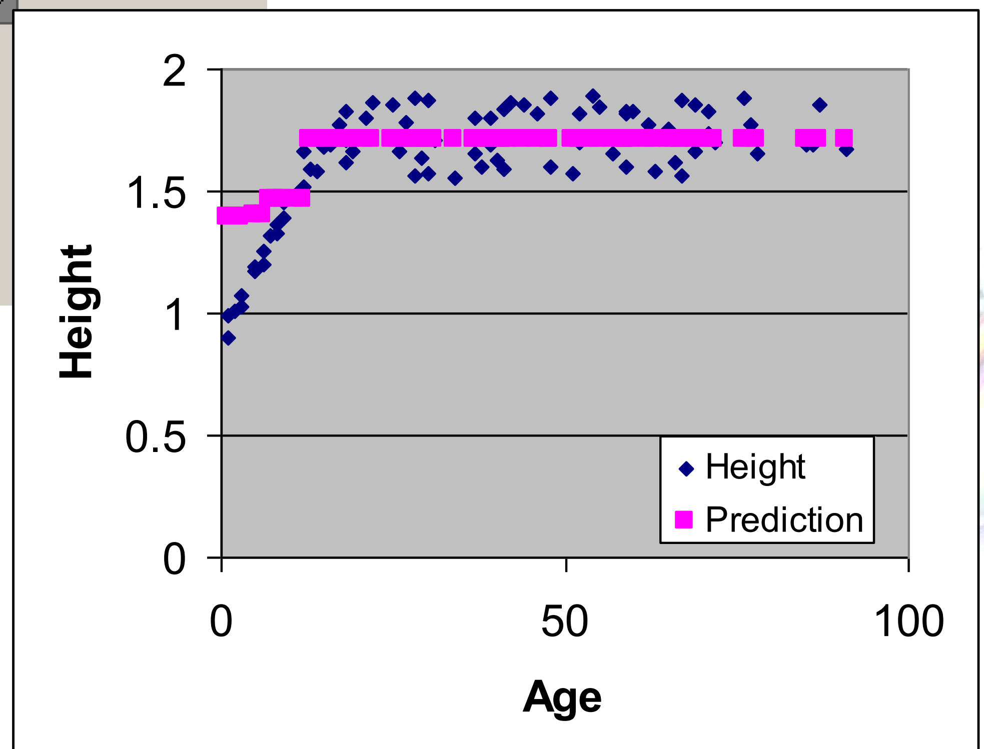
Linear Regression: prediction

Height = 0.0056 * Age + 1.4181

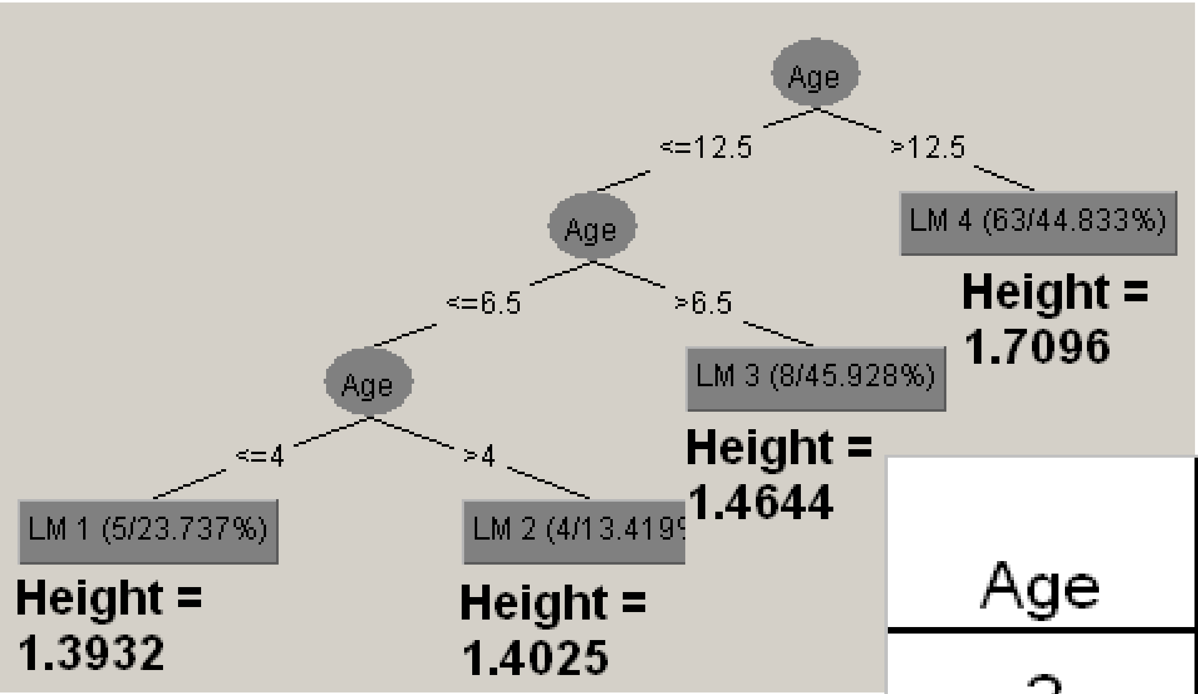
Age	Height	Linear regression
2	0.85	
10	1.4	
35	1.7	
70	1.6	

Regression tree



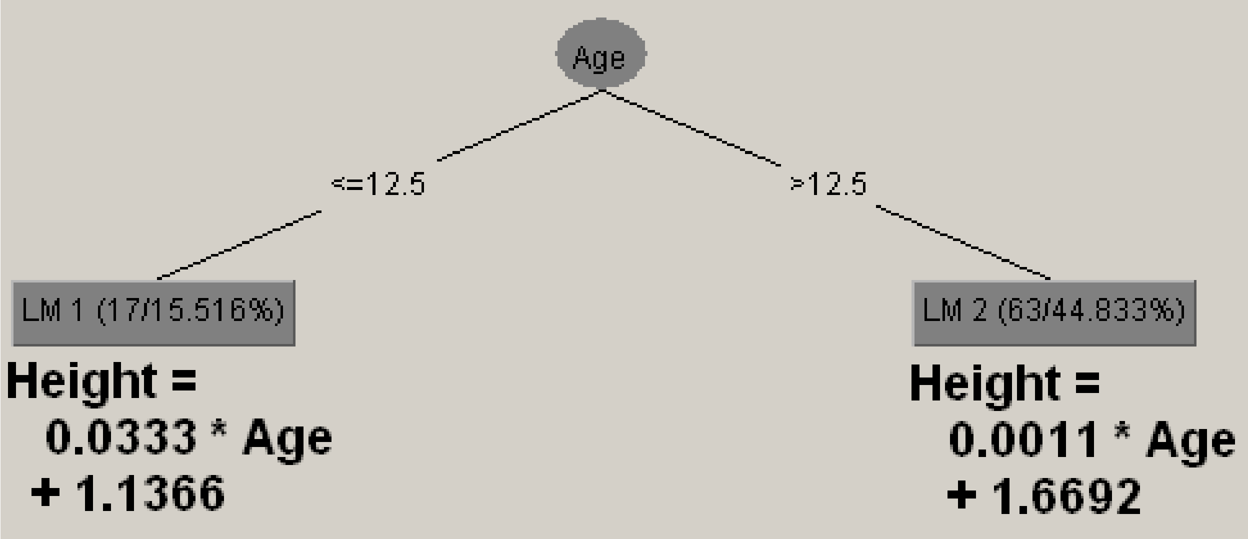


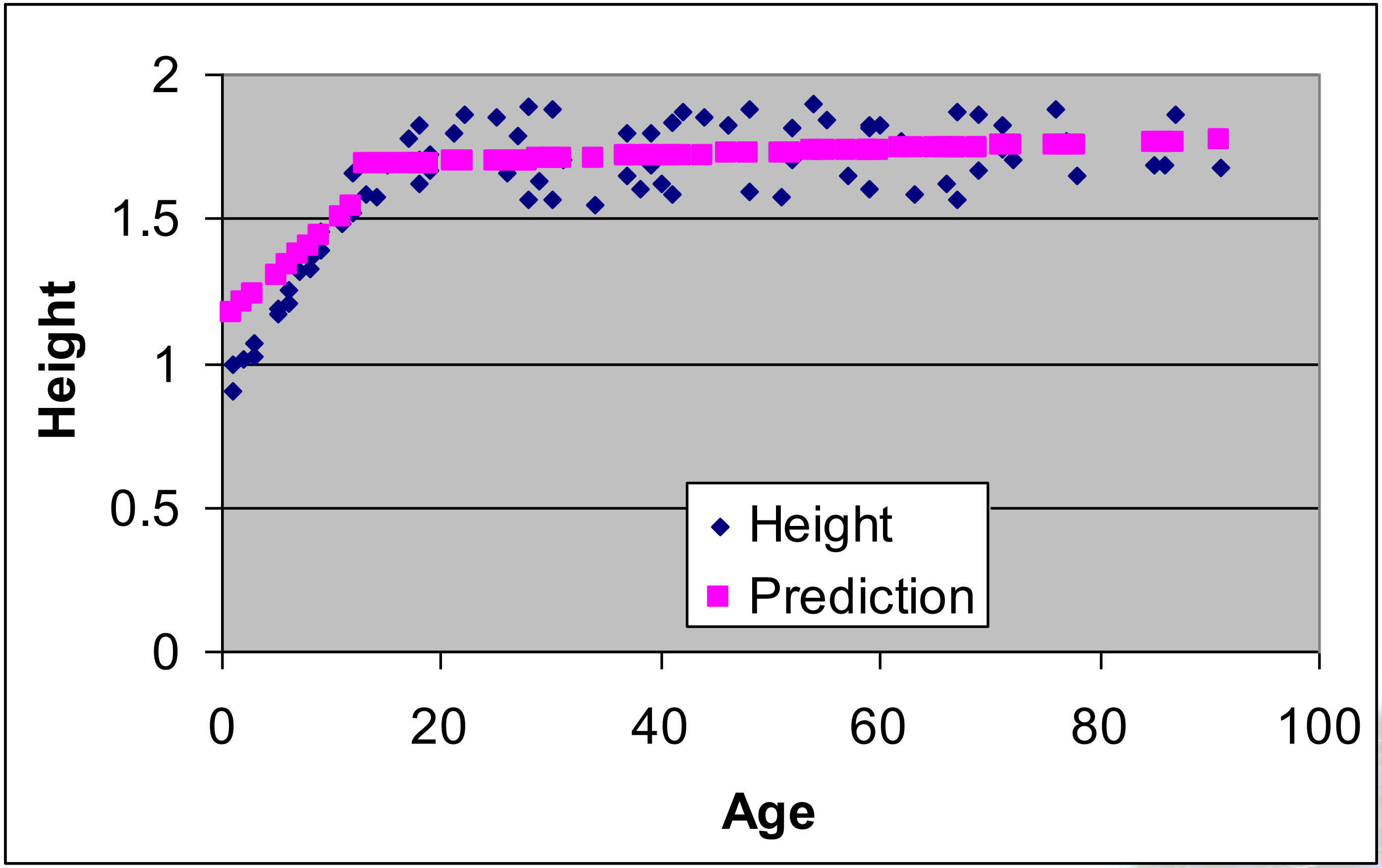
Regression tree: prediction



Age	Height	Regression tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	

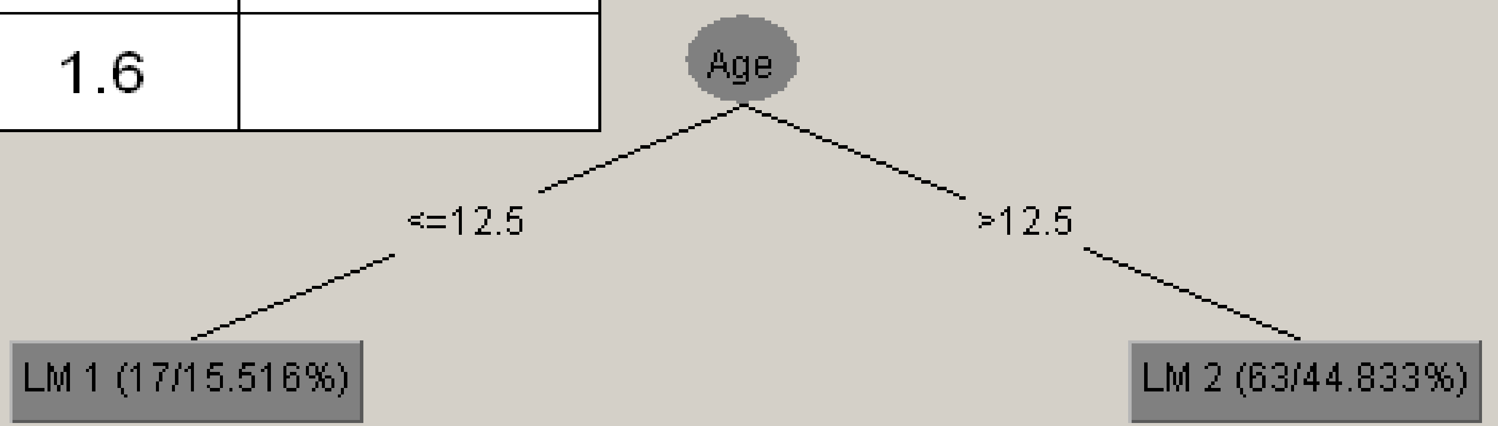
Model tree


Age
<=12.5
>12.5
LM 1 (17/15.516%)
LM 2 (63/44.833%)
Height = 0.0333 * Age + 1.1366
Height = 0.0011 * Age + 1.6692


2
1.5
1
0.5
0
0
20
40
60
80
100
Height
Age
Height
Prediction

Model tree: prediction

Age	Height	Model tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	

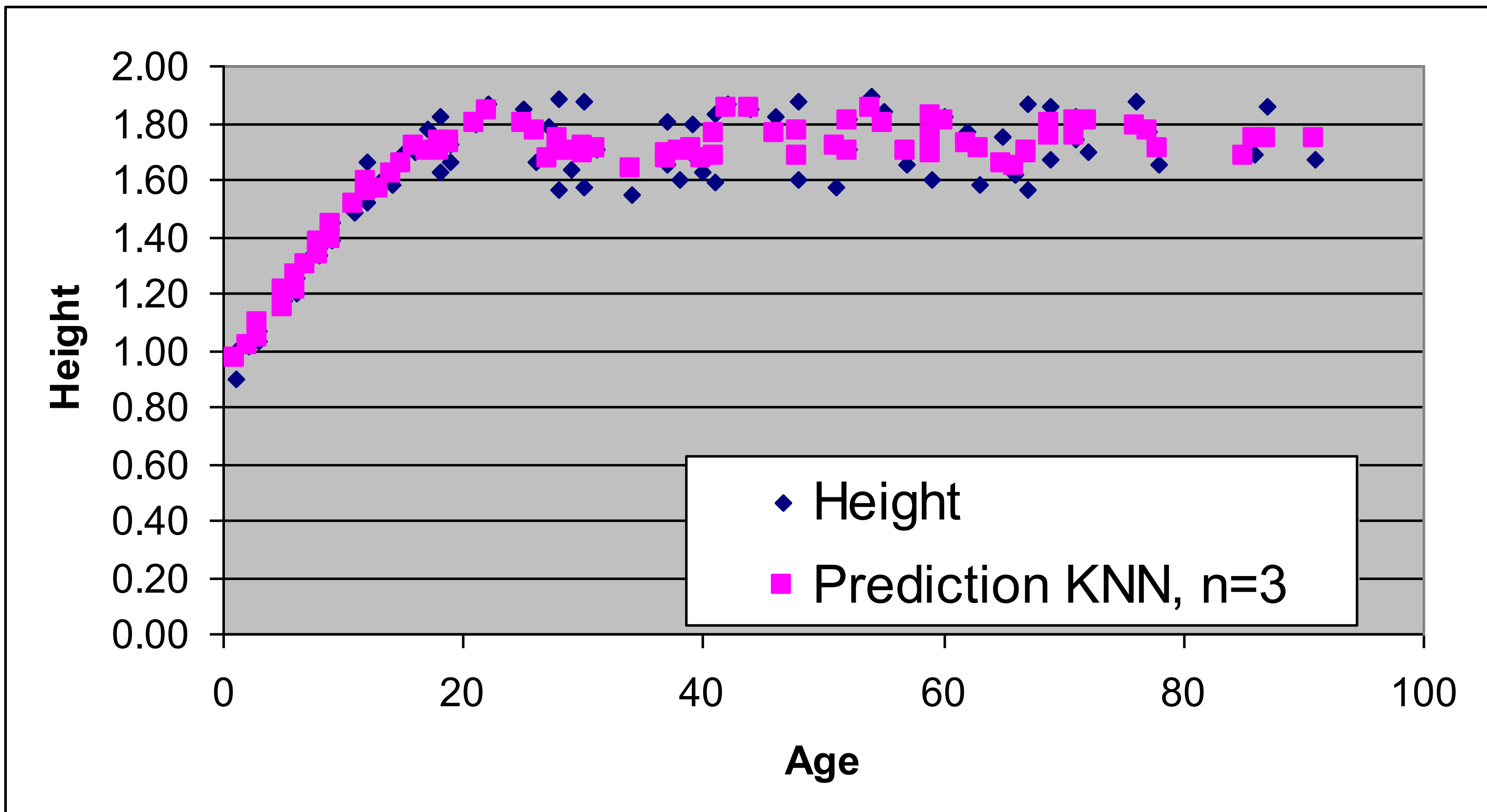


Height = 0.0333 * Age + 1.1366

Height = 0.0011 * Age + 1.6692

KNN – K nearest neighbors

- Looks at K closest examples (by non-target attributes) and predicts the average of their target variable
- In this example, K=3



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

KNN prediction

Age	Height
1	0.90
1	0.99
2	1.01
3	1.03
3	1.07
5	1.19
5	1.17

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

KNN prediction

Age	Height
8	1.36
8	1.33
9	1.45
9	1.39
11	1.49
12	1.66
12	1.52
13	1.59
14	1.58

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

KNN prediction

Age	Height
30	1.57
30	1.88
31	1.71
34	1.55
37	1.65
37	1.80
38	1.60
39	1.69
39	1.80

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

KNN prediction

Age	Height
67	1.56
67	1.87
69	1.67
69	1.86
71	1.74
71	1.82
72	1.70
76	1.88

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	

KNN video

- http://videolectures.net/aaai07_bosch_knnc

Which predictor is the best?

Age	Height	Baseline	Linear regression	Regression tree	Model tree	kNN
2	0.85	1.63	1.43	1.39	1.20	1.00
10	1.4	1.63	1.47	1.46	1.47	1.44
35	1.7	1.63	1.61	1.71	1.71	1.67
70	1.6	1.63	1.81	1.71	1.75	1.77

Evaluating numeric prediction

Performance measure	Formula
mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
relative squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i(p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_p=\frac{\sum_i(p_i-\bar{p})^2}{n-1}$, and $S_A=\frac{\sum_i(a_i-\bar{a})^2}{n-1}$

Numeric prediction	**Classification**
Data: attribute-value description	
Target variable: Continuous	**Target variable**: Categorical (nominal)
Evaluation: cross validation, separate test set, …	
Error: MSE, MAE, RMSE, …	**Error**: 1-accuracy
Algorithms: Linear regression, regression trees,…	**Algorithms**: Decision trees, Naïve Bayes, …
Baseline predictor: Mean of the target variable	**Baseline predictor**: Majority class

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

KNN for classification?

- Yes.
- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

Comparison of KNN and naïve Bayes

	Naïve Bayes	**KNN**
Used for		
Handle categorical data		
Handle numeric data		
Model interpretability		
Lazy classification		
Evaluation		
Parameter tuning		

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Comparison of KNN and naïve Bayes

	Naïve Bayes	**KNN**
Used for	Classification	Classification and numeric prediction
Handle categorical data	Yes	Proper distance function needed
Handle numeric data	Discretization needed	Yes
Model interpretability	Limited	No
Lazy classification	Partial	Yes
Evaluation	Cross validation,…	Cross validation,…
Parameter tuning	No	No

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

Comparison of regression and decision trees

1. Data
2. Target variable
3. Evaluation
4. Error
5. Algorithm
6. Heuristic
7. Stopping criterion

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Comparison of regression and decision trees

<table>
<tr><th>Regression trees</th><th>Decision trees</th></tr>
<tr><td colspan="2">Data: attribute-value description</td></tr>
<tr><td>Target variable:
Continuous</td><td>Target variable:
Categorical (nominal)</td></tr>
<tr><td colspan="2">Evaluation: cross validation, separate test set, …</td></tr>
<tr><td>Error:
MSE, MAE, RMSE, …</td><td>Error:
1-accuracy</td></tr>
<tr><td colspan="2">Algorithm:
Top down induction, shortsighted method</td></tr>
<tr><td>Heuristic:
Standard deviation</td><td>Heuristic :
Information gain</td></tr>
<tr><td>Stopping criterion:
Standard deviation< threshold</td><td>Stopping criterion:
Pure leafs (entropy=0)</td></tr>
</table>

Discussion

1. Can KNN be used for classification tasks?
2. Compare KNN and Naïve Bayes.
3. Compare decision trees and regression trees.
4. Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent

    1. Is this a classification or a numeric prediction problem?
    2. What if such a variable is an attribute, is it nominal or numeric?

Classification or a numeric prediction problem?

- Target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
- Classification: the **misclassification cost** is the same if “non sufficient” is classified as “sufficient” or if it is classified as “very good”
- Numeric prediction: The error of predicting “2” when it should be “1” is 1, while the error of predicting “5” instead of “1” is 4.
- If we have a variable with ordered values,

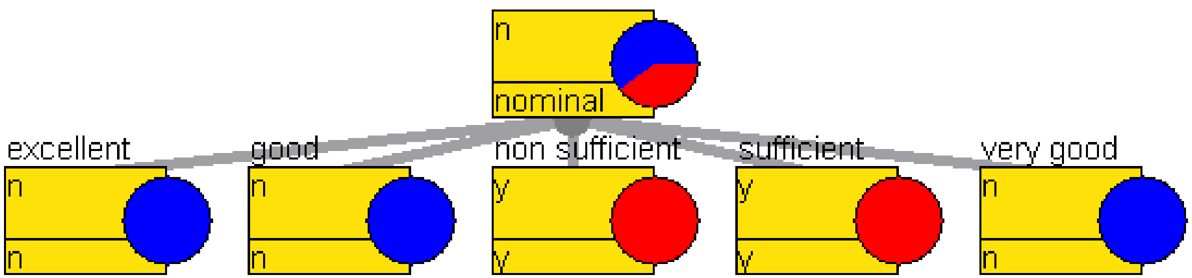
it should be considered numeric.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

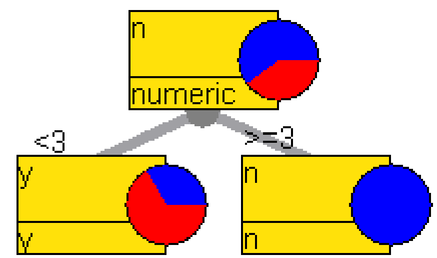
Nominal or numeric attribute?

- A variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. Excellent

Nominal:



Numeric:



- If we have a variable with **ordered** values, it should be considered numeric.

Association Rules

Association rules

- Rules **X → Y**, X, Y conjunction of items
- Task: Find **all** association rules that satisfy minimum support and minimum confidence constraints

- **Support**:

  $Sup(X \rightarrow Y) = \#XY/\#D \cong p(XY)$

- **Confidence**:

  $Conf(X \rightarrow Y) = \#XY/\#X \cong p(XY)/p(X) = p(Y|X)$

Association rules - algorithm

1. generate frequent itemsets with a minimum support constraint
2. generate rules from frequent itemsets with a minimum confidence constraint

* Data are in a transaction database

Association rules – transaction database

Items: **A**=apple, **B**=banana,
C=coca-cola, **D**=doughnut

- Client 1 bought: A, B, C, D
- Client 2 bought: B, C
- Client 3 bought: B, D
- Client 4 bought: A, C
- Client 5 bought: A, B, D
- Client 6 bought: A, B, C

Frequent itemsets

- Generate frequent itemsets with support at least 2/6

A	B	C	D
1	1	1	1
	1	1	
	1		1
1		1	
1	1		1
1	1	1	

Frequent itemsets algorithm

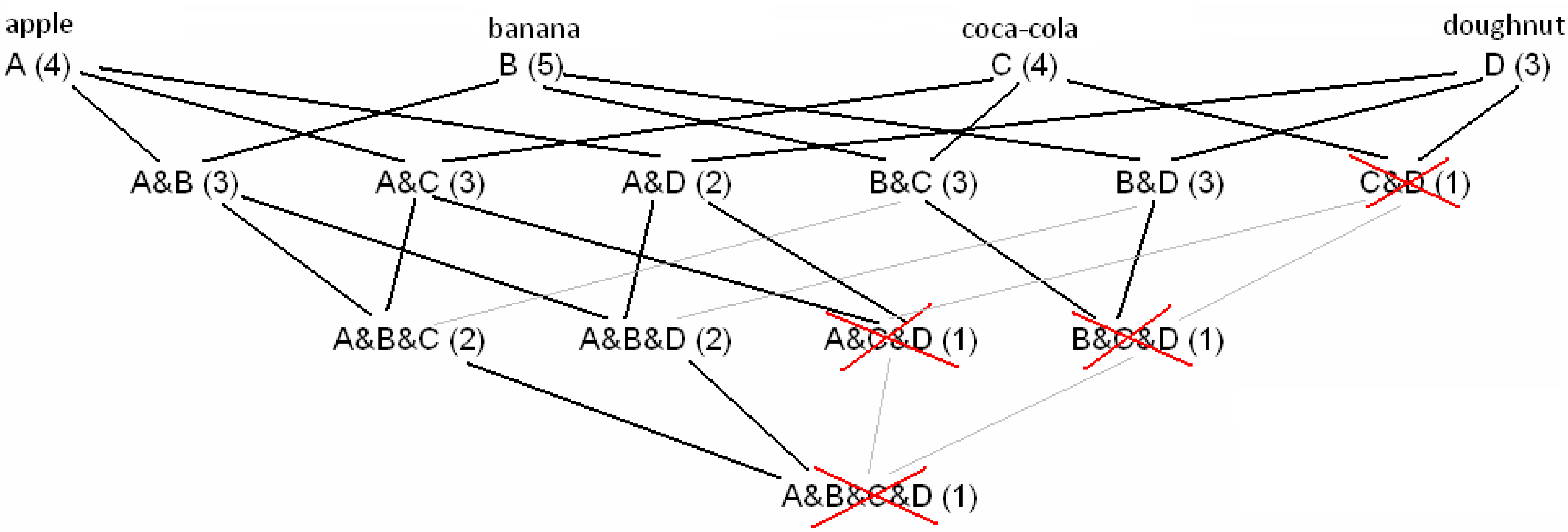
Items in an itemset should be **sorted** alphabetically.

1. Generate all 1-itemsets with the given minimum support.
2. Use 1-itemsets to generate 2-itemsets with the given minimum support.
3. From 2-itemsets generate 3-itemsets with the given minimum support as unions of 2-itemsets with the same item at the beginning.
4. …
5. From n-itemsets generate (n+1)-itemsets as unions of n-itemsets with the same (n-1) items at the beginning.

- To generate itemsets at level n+1 items from level n are used with a constraint: itemsets have to start with the same n-1 items.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Frequent itemsets lattice



Frequent itemsets:

- A&B, A&C, A&D, B&C, B&D
- A&B&C, A&B&D

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

Rules from itemsets

- A&B is a frequent itemset with support 3/6
- Two possible rules
  - A→B confidence = #(A&B)/#A = 3/4
  - B→A confidence = #(A&B)/#B = 3/5
- All the counts are in the itemset lattice!

Quality of association rules

Support(X) = #X / #D .... P(X)
Support(X→Y) = Support (XY) = #XY / #D .... P(XY)
Confidence(X→Y) = #XY / #X .... P(Y|X)

Lift(X→Y) = Support(X→Y) / (Support (X)*Support(Y))

Leverage(X→Y) = Support(X→Y) – Support(X)*Support(Y)

Conviction(X → Y) = 1-Support(Y)/(1-Confidence(X→Y))

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Quality of association rules

Support(X) = #X / #D .... P(X)

Support(X→Y) = Support (XY) = #XY / #D .... P(XY)

Confidence(X→Y) = #XY / #X .... P(Y|X)

Lift(X→Y) = Support(X→Y) / (Support (X)*Support(Y))

How many more times the items in X and Y occur together then it would be expected if the itemsets were statistically independent.

Leverage(X→Y) = Support(X→Y) – Support(X)*Support(Y)

Similar to lift, difference instead of ratio.

Conviction(X → Y) = 1-Support(Y)/(1-Confidence(X→Y))

Degree of implication of a rule.

Sensitive to rule direction.

Discussion

- Transformation of an attribute-value dataset to a transaction dataset.
- What would be the association rules for a dataset with two items A and B, each of them with support 80% and appearing in the same transactions as rarely as possible?
  - minSupport = 50%, min conf = 70%
  - minSupport = 20%, min conf = 70%
- What if we had 4 items: A, ¬A, B, ¬ B
- Compare decision trees and association rules regarding handling an attribute like “PersonID”. What about attributes that have many values (eg. Month of year)