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

Practice plan

- 2007/11/8: Classification
  - Decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers (confusion matrix, classification accuracy)
  - Predictive data mining in Weka
- 2007/11/15: Numeric prediction and descriptive data mining
  - Models for numeric prediction
  - Association rules
  - Regression models and evaluation in Weka
  - Descriptive data mining in Weka
  - Discussion about seminars and exam
- 2007/11/29: Written examination and seminar proposal presentations

Numeric prediction

Baseline,

Linear Regression,

Regression tree,

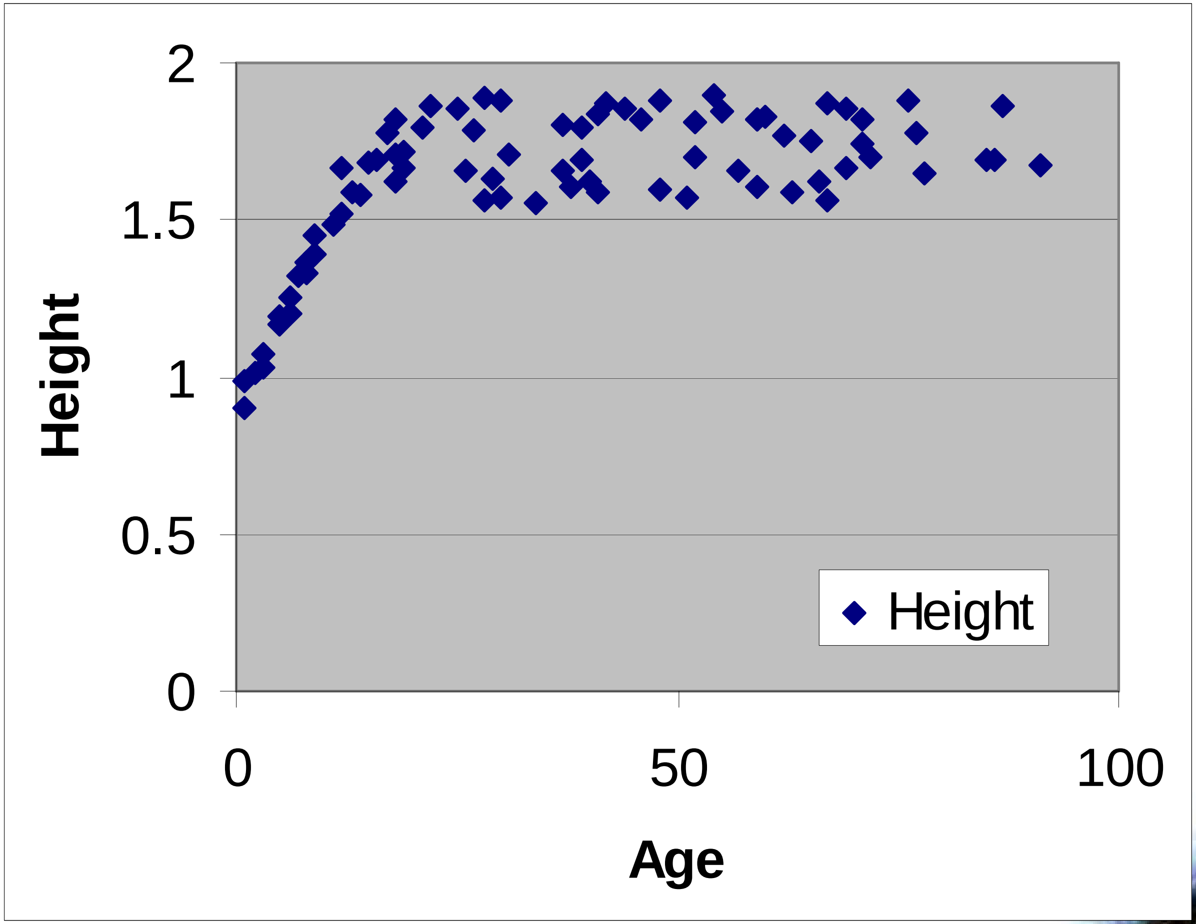
Model Tree,

KNN

Regression	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

Example

- data about 80 people: Age and Height



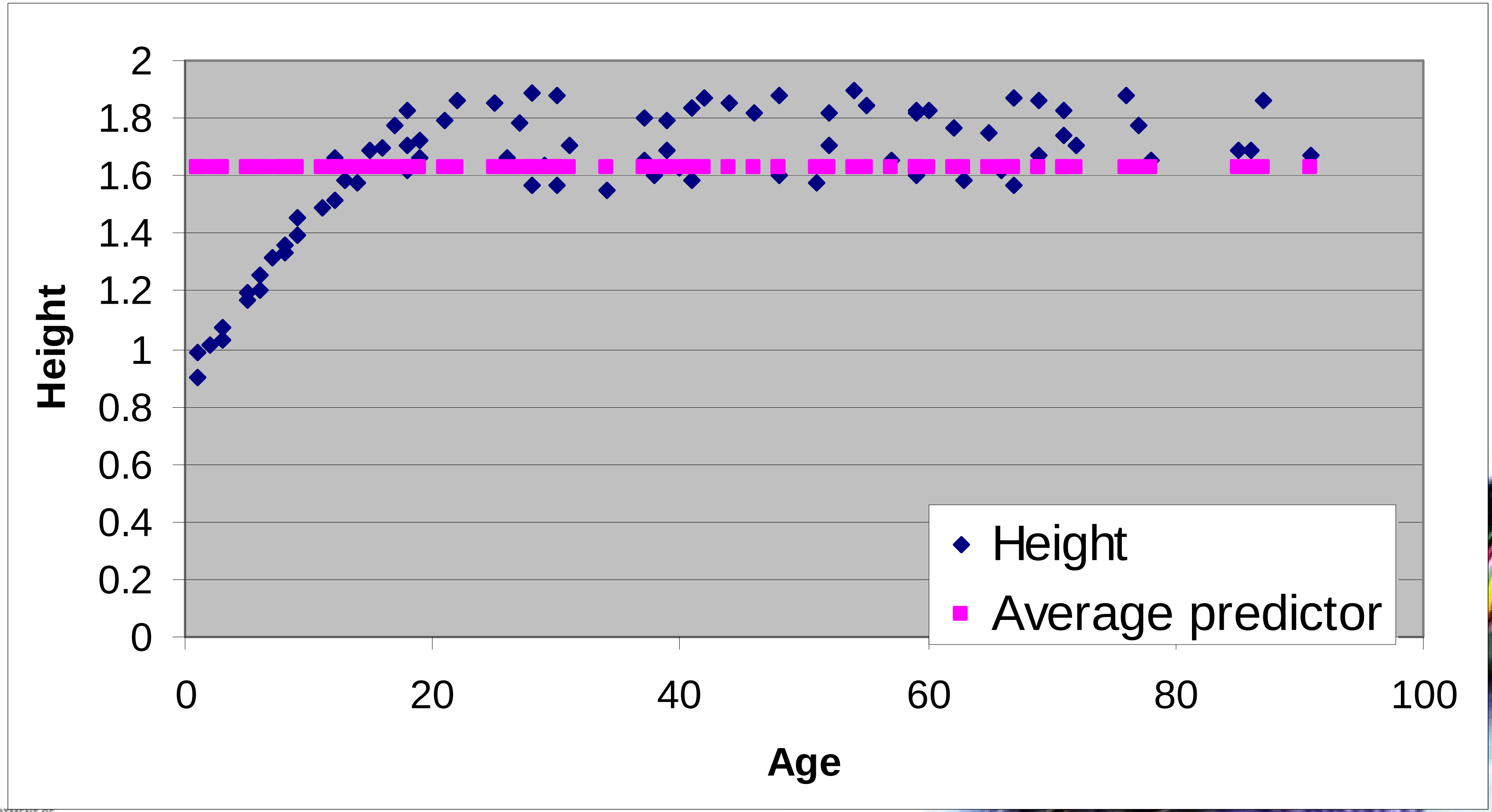
Age	Height
3	1.03
5	1.19
6	1.26
9	1.39
15	1.69
19	1.67
22	1.86
25	1.85
41	1.59
48	1.60
54	1.90
71	1.82
...	...

Test set

Age	Height
2	0.85
10	1.4
35	1.7
70	1.6

Baseline numeric predictor

- Average of the target variable



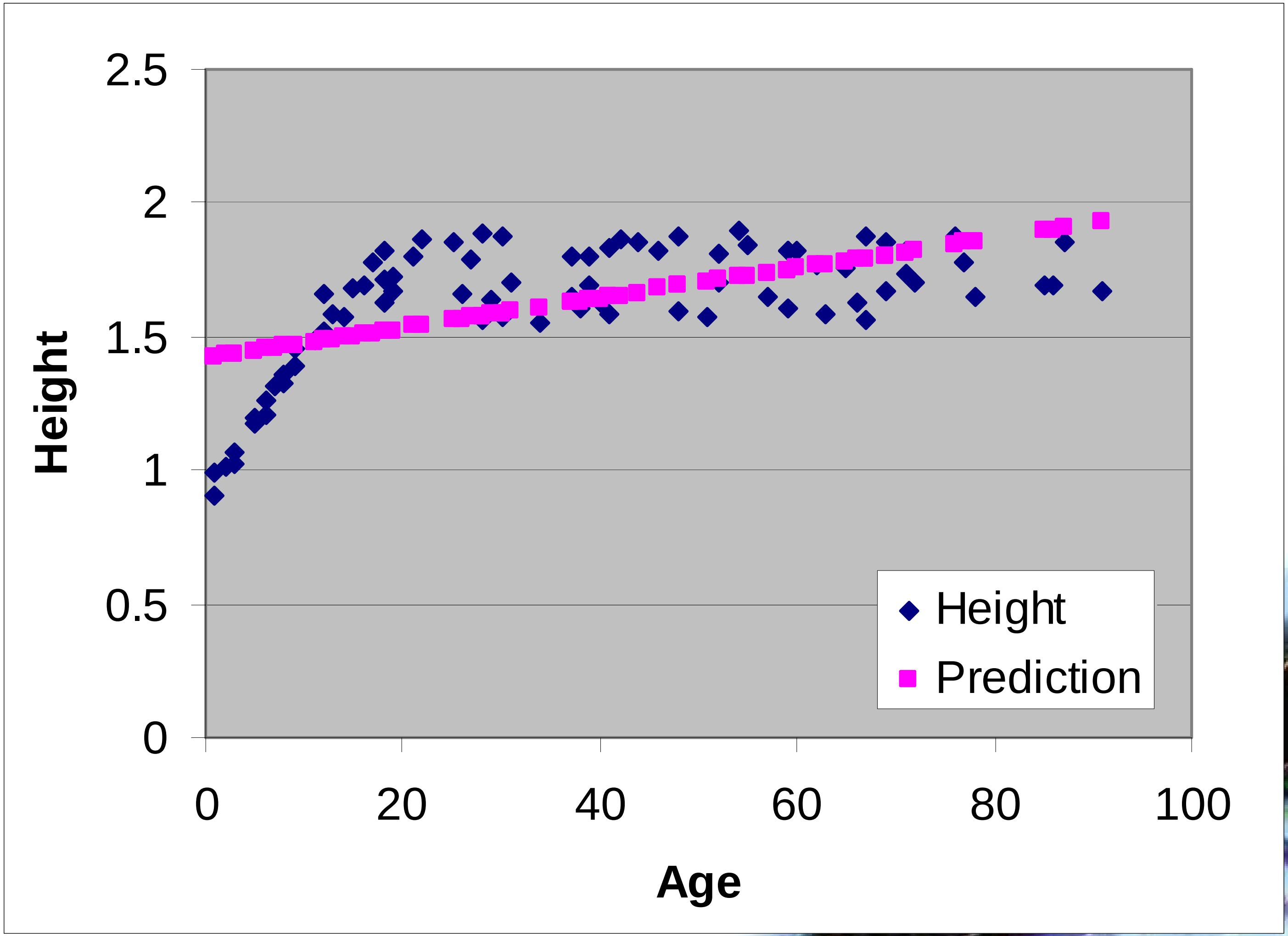
Baseline predictor: prediction

Average of the target variable is 1.63

Age	Height	Baseline
2	0.85	
10	1.4	
35	1.7	
70	1.6	

Linear Regression Model

Height = 0.0056 * Age + 1.4181



Linear Regression: prediction

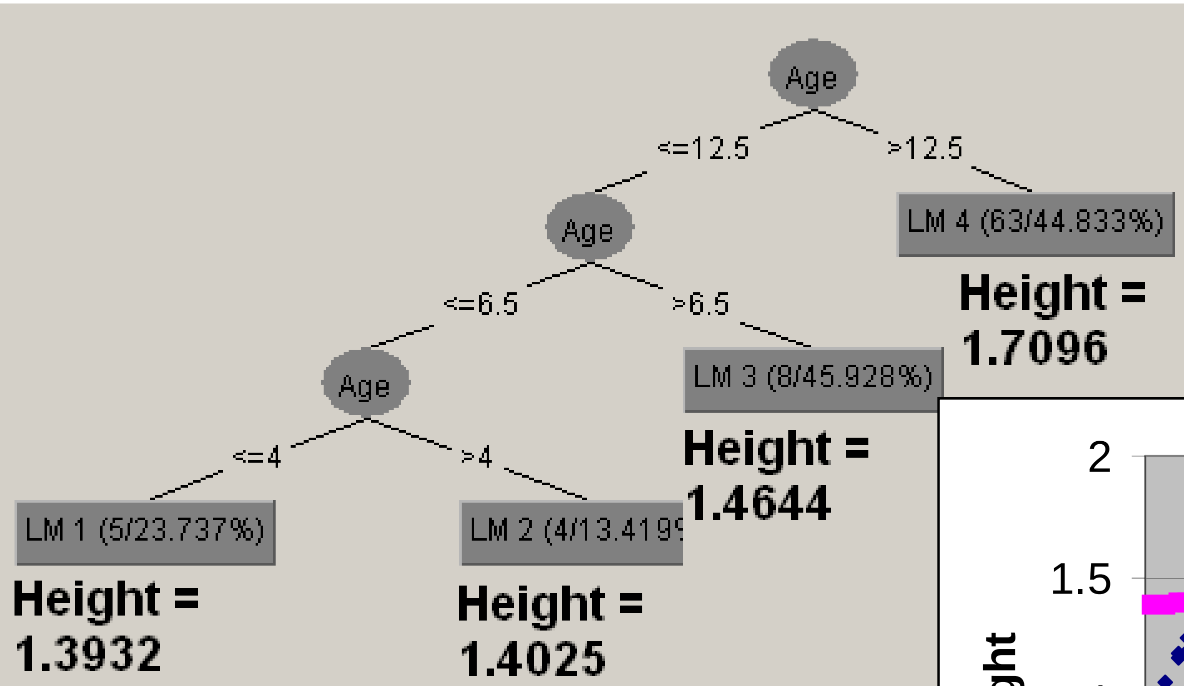
Height = 0.0056 * Age + 1.4181

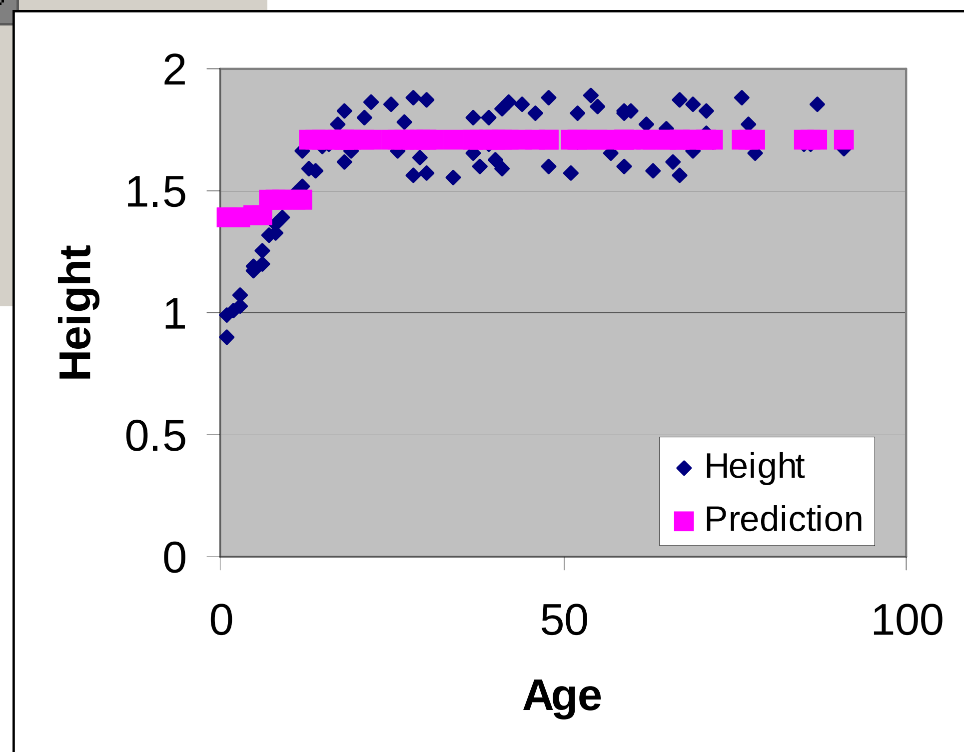
Age	Height	Linear regression
2	0.85	
10	1.4	
35	1.7	
70	1.6	

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

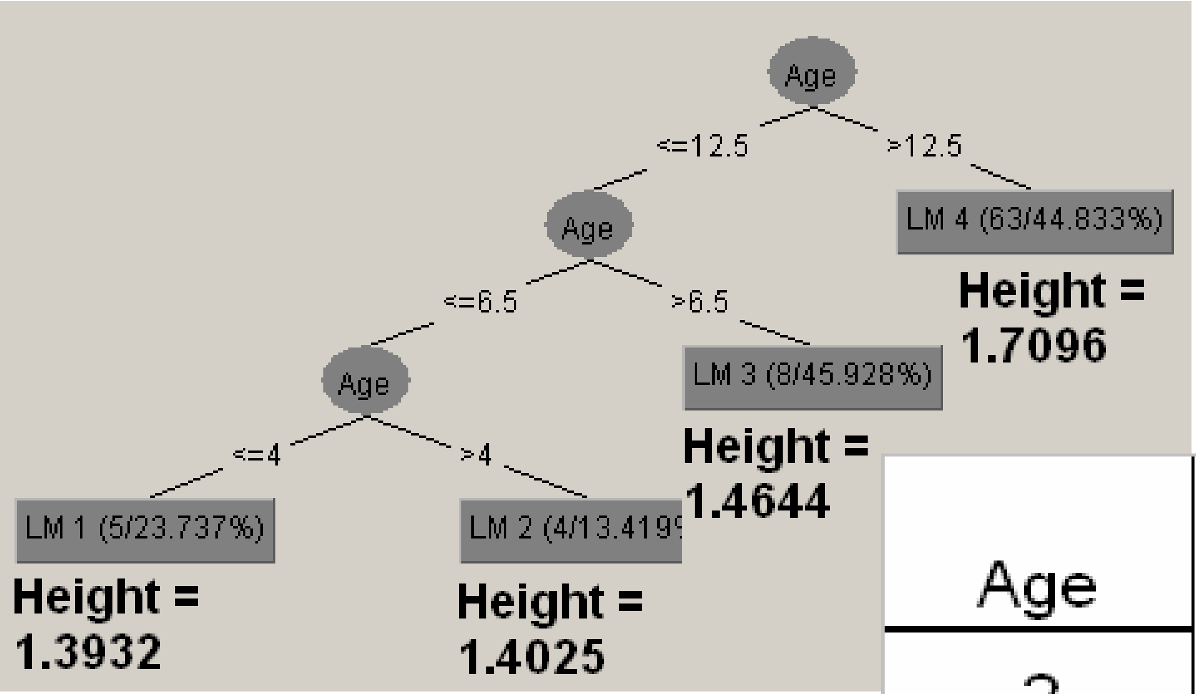
LEARNED MEDIA

Regression tree





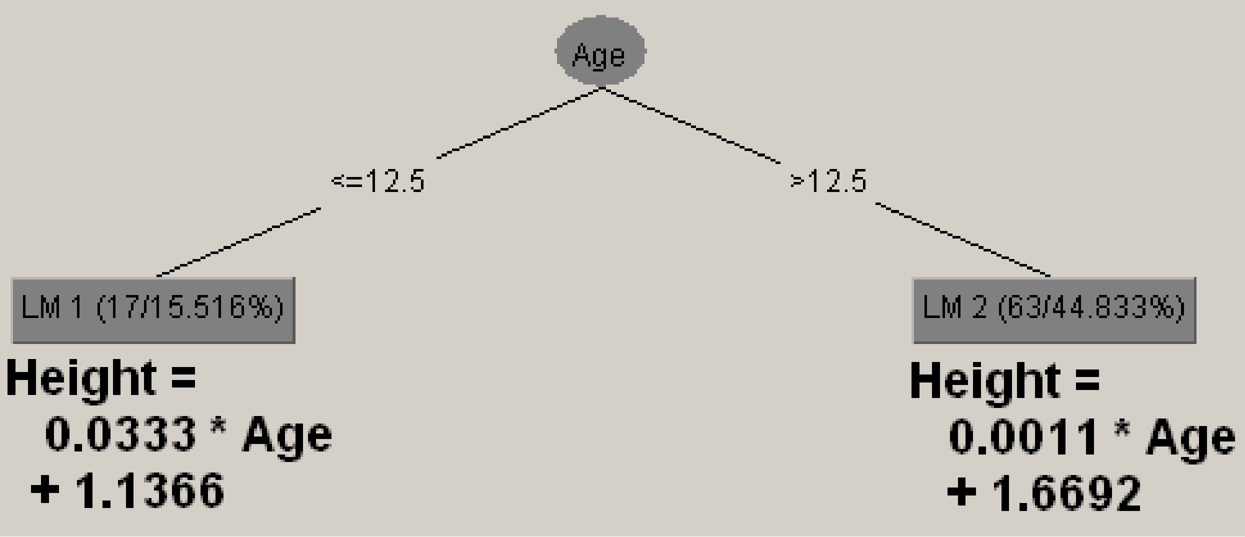
Regression tree: prediction

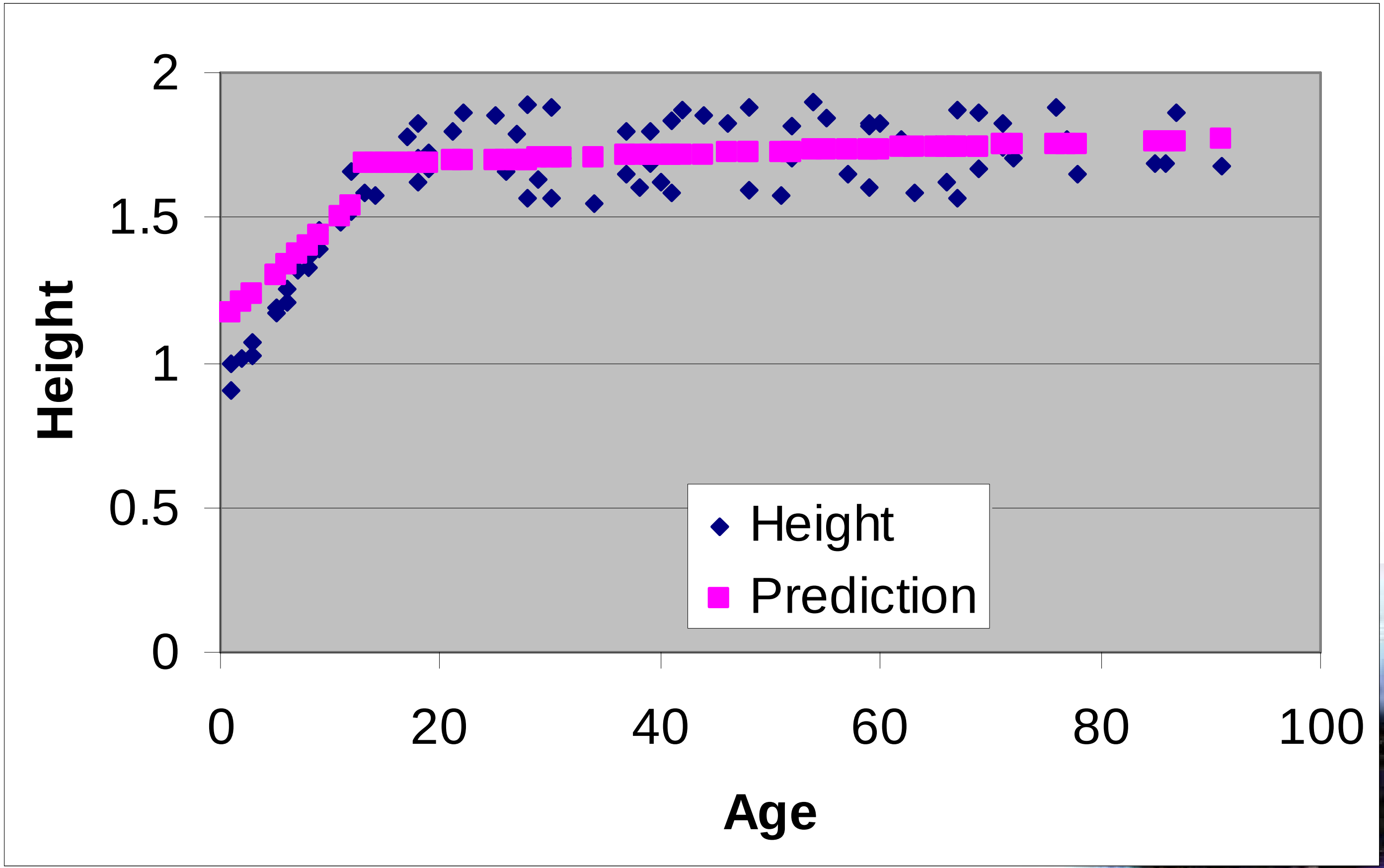


Age	Height	Regression tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

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

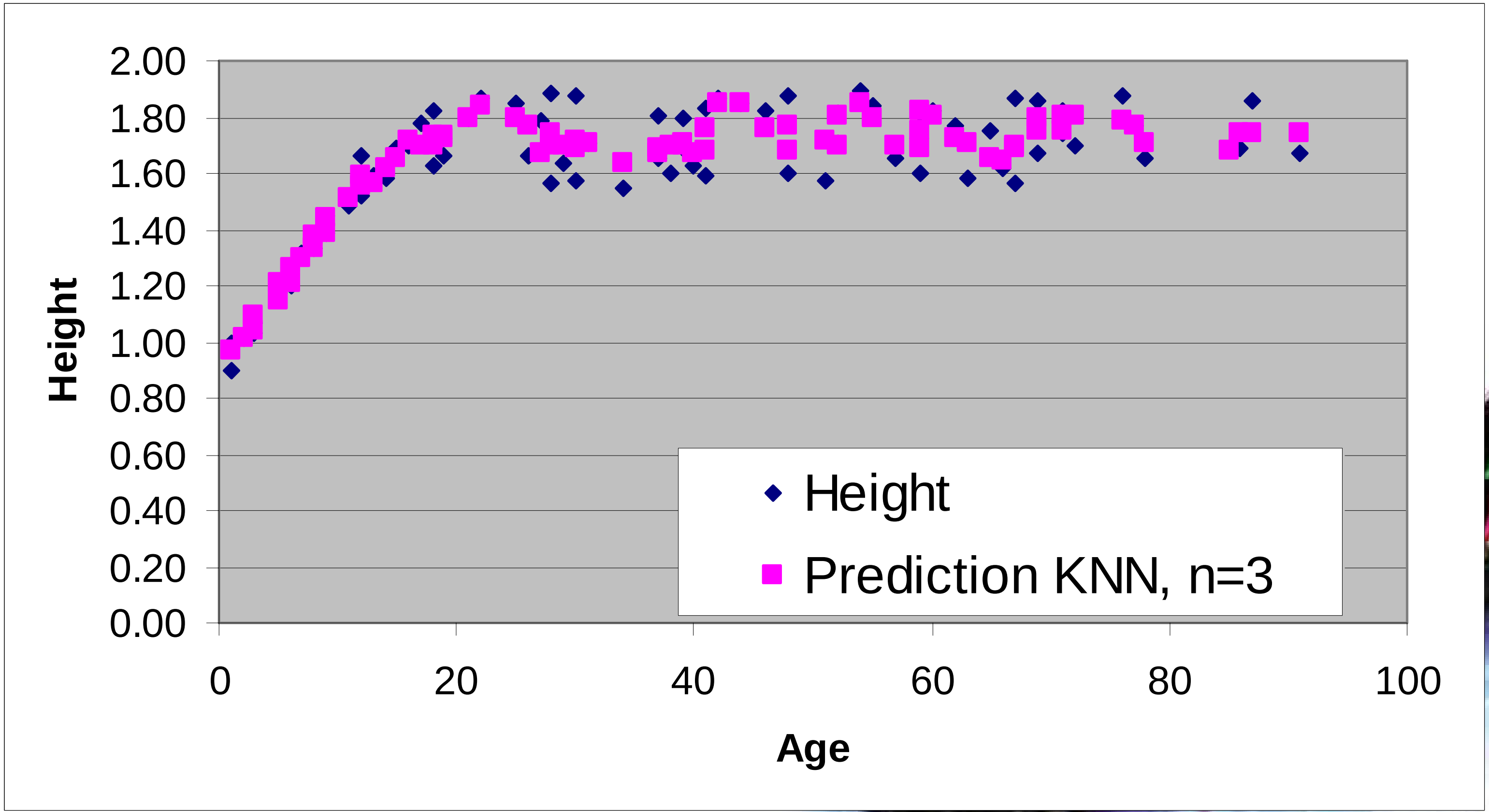

Height
Age
0
0.5
1
1.5
2
0
20
40
60
80
100
Height
Prediction

Model tree: prediction

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

KNN – K nearest neighbors

- Looks at K closest examples (by age) and predicts the average of their target variable
- 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	

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	

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	1.01
10	1.4	1.63	1.47	1.46	1.47	1.51
35	1.7	1.63	1.61	1.71	1.71	1.67
70	1.6	1.63	1.81	1.71	1.75	1.81

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 $S_A=\frac{\sum_i(a_i-\bar{a})^2}{n-1}$

Age	Height	Baseline	pi-ai	Linear regression	pi-ai
2	0.85	1.63	0.78	1.43	0.58
10	1.4	1.63	0.23	1.47	0.07
35	1.7	1.63	-0.07	1.61	-0.09
70	1.6	1.63	0.03	1.81	0.21
mean-squared error					
root mean-squared error					
mean absolute error					
relative squared error					
root relative squared error					
relative absolute error					
correlation coefficient					

Age	Height	Regression tree	pi-ai	Model tree	pi-ai	kNN	pi-ai
2	0.85	1.39	0.54	1.20	0.35	1.01	0.16
10	1.4	1.46	0.06	1.47	0.07	1.51	0.11
35	1.7	1.71	0.01	1.71	0.01	1.67	-0.03
70	1.6	1.71	0.11	1.75	0.15	1.81	0.21
mean-squared error							
root mean-squared error							
mean absolute error							
relative squared error							
root relative squared error							
relative absolute error							
correlation coefficient							

Discussion

- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.

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

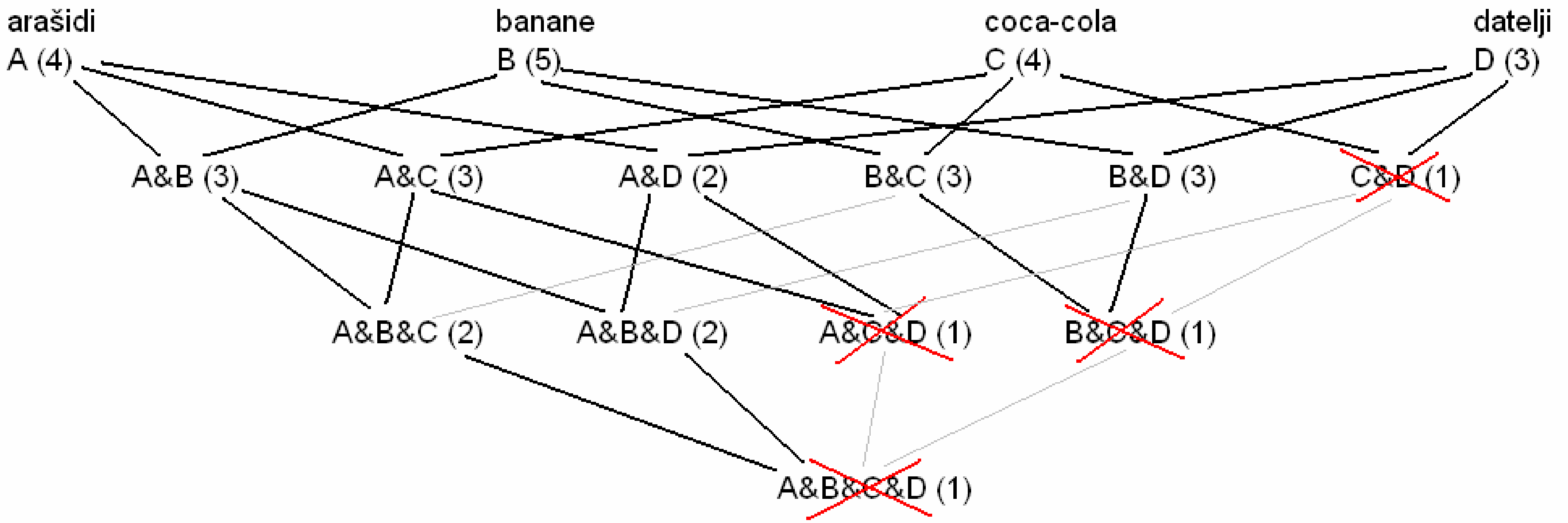
A	B	C	D
1	1	1	1
	1	1	
	1		1
1		1	
1	1		1
1	1	1	

Frequent itemsets algorithm

Items in an itemset should be sorted alphabetically.

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

Frequent itemsets lattice



Frequent itemsets:

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

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!

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