Data Mining and Knowledge Discovery

Petra Kralj Novak

Petra.Kralj.Novak@ijs.si

Practice, 2010/12/2

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Practice plan

- 2010/11/25: Predictive data mining
  - Decision trees
  - Naïve Bayes classifier
  - Evaluating classifiers (separate test set, cross validation, confusion matrix, classification accuracy)
  - Predictive data mining in Weka
- 2010/12/2: Numeric prediction and descriptive data mining
  - Numeric prediction models
  - Association rules
  - Regression models and evaluation in Weka
  - Descriptive data mining in Weka
  - Discussion about seminars and exam

- 2010/12/16: Written exam, Seminar proposal presentations

- 2011/2/1: Deadline for data mining papers (written seminar)
- 2011/2/3: Data mining seminar presentations

Numeric prediction

Baseline,

Linear Regression,

Regression tree,

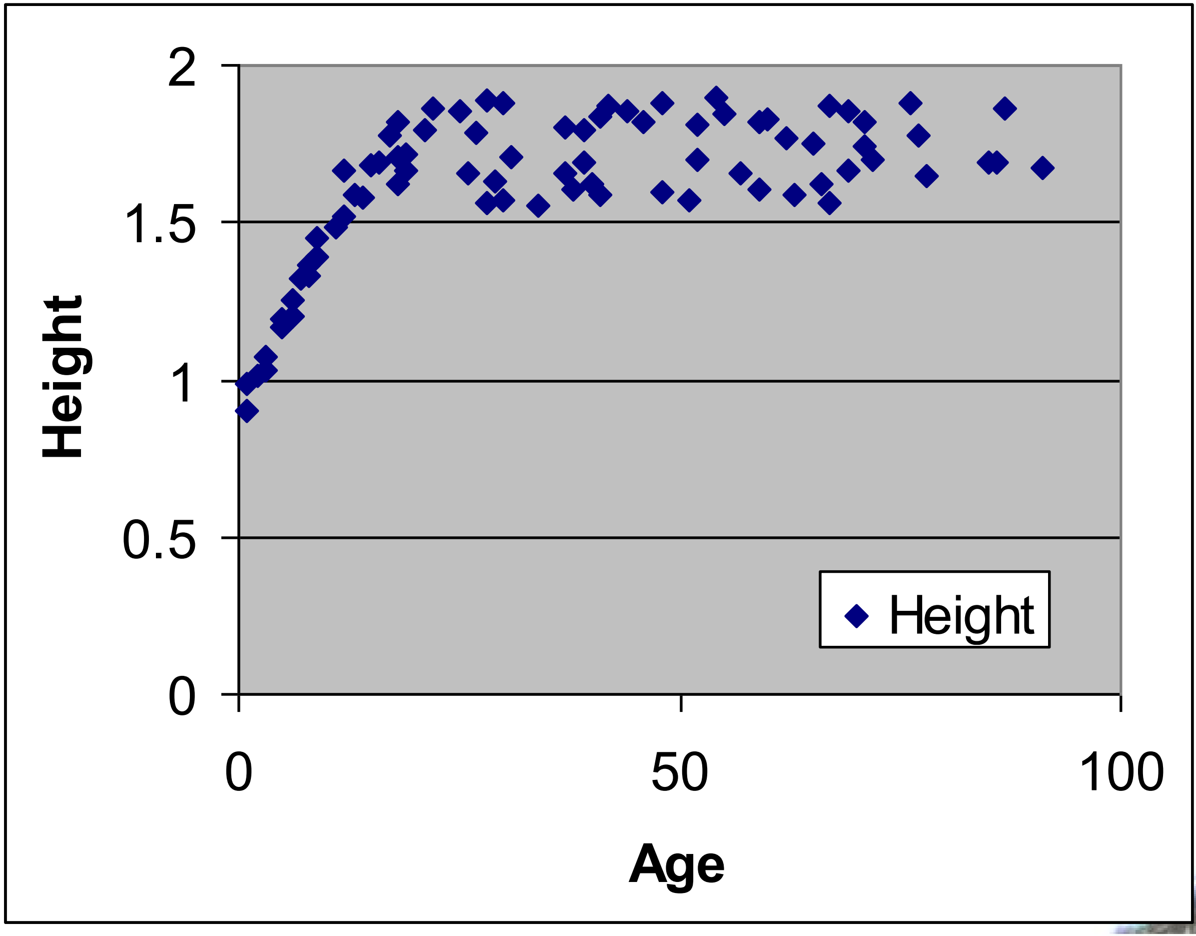
Model Tree,

KNN

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

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
...	...

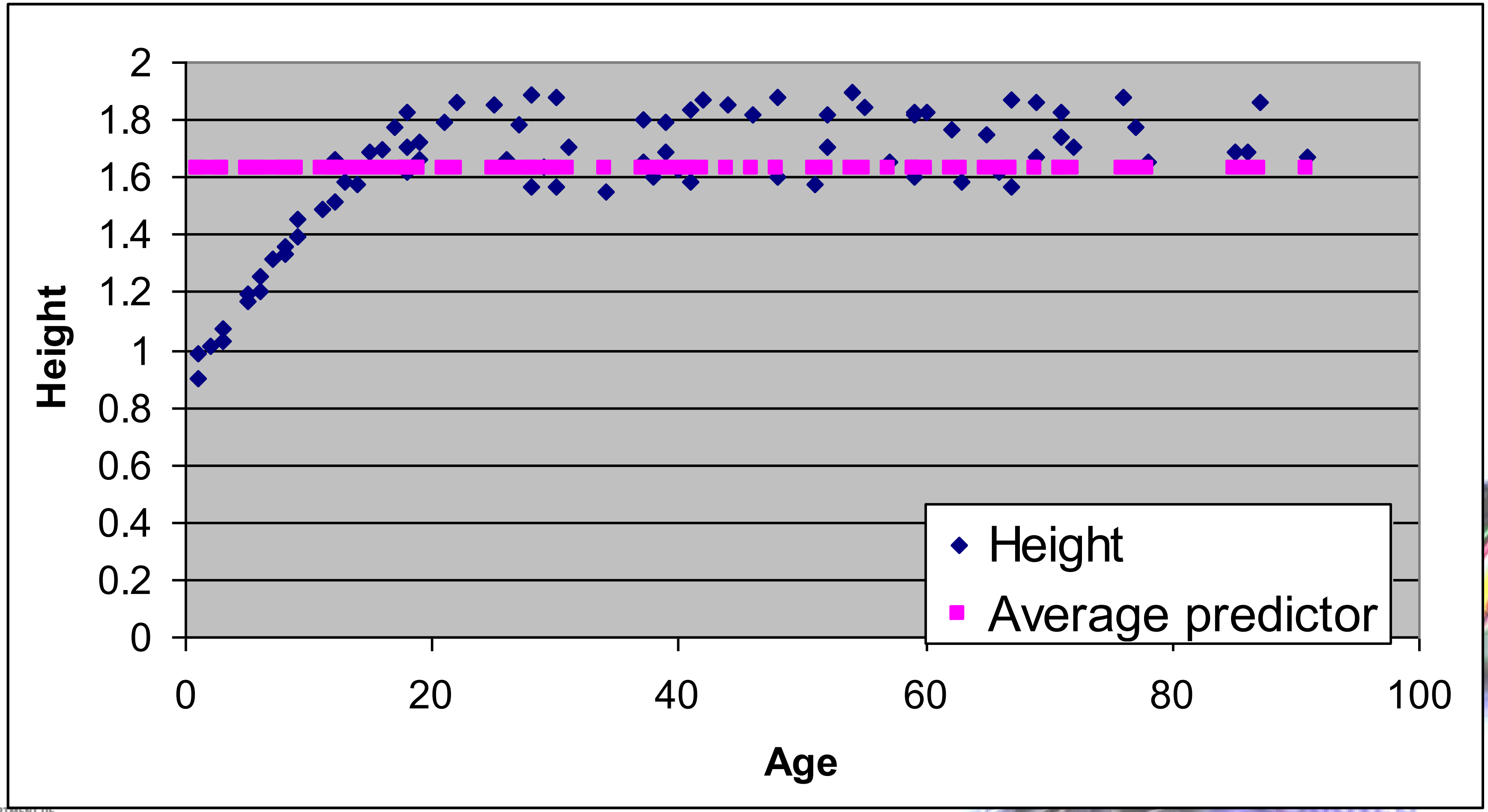
Test set

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



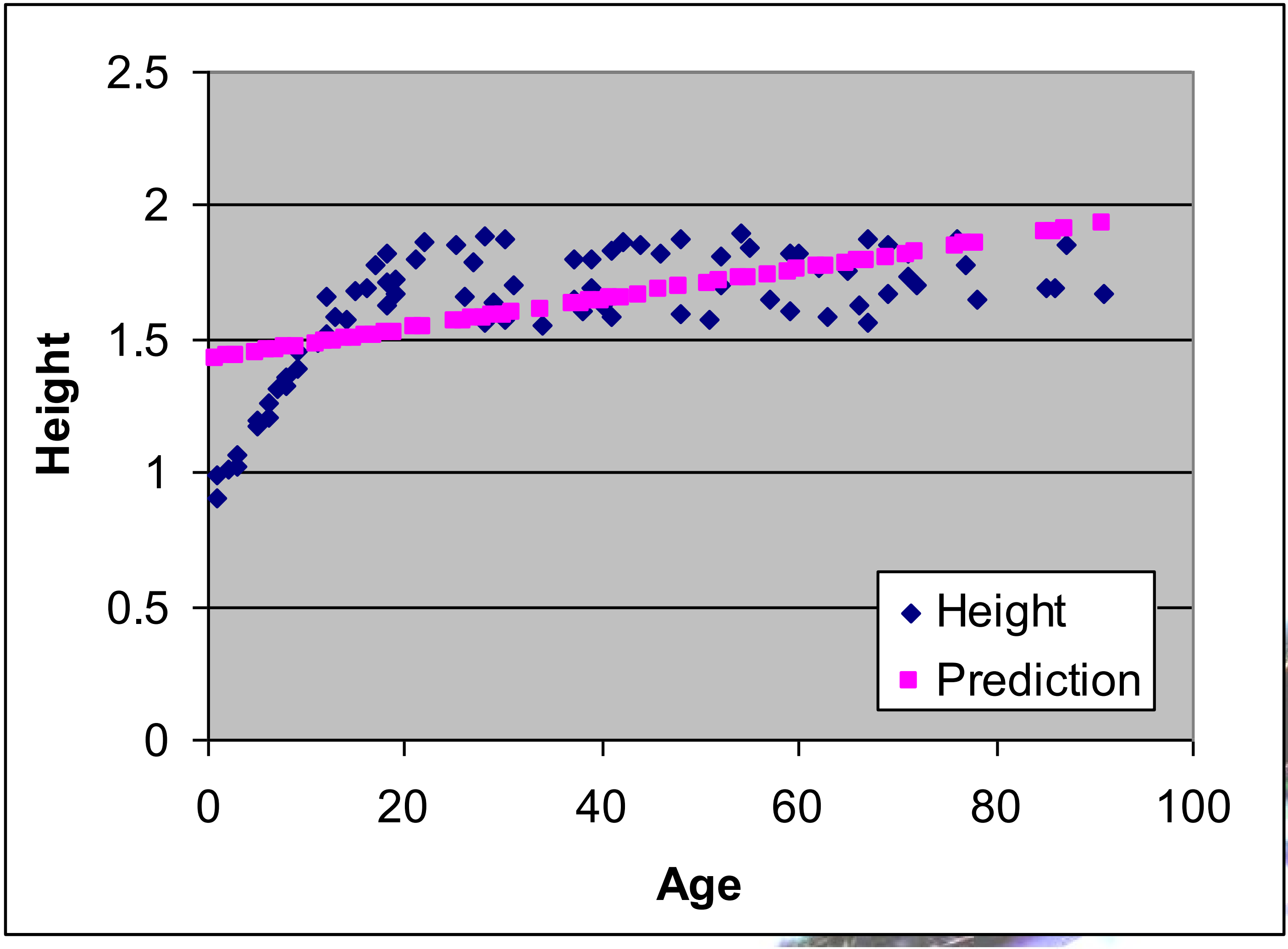
Baseline predictor: prediction

Average of the target variable is 1.63

Age	Height	Baseline
2	0.85	
10	1.4	
35	1.7	
70	1.6	

Linear Regression Model

Height = 0.0056 * Age + 1.4181

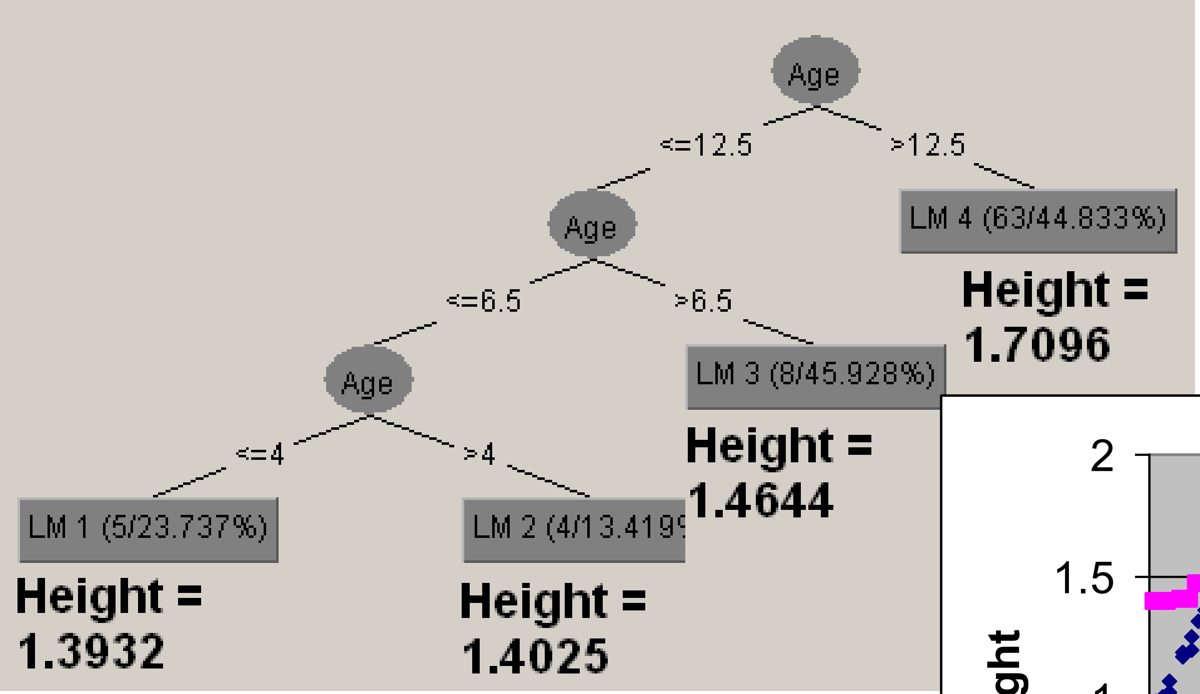


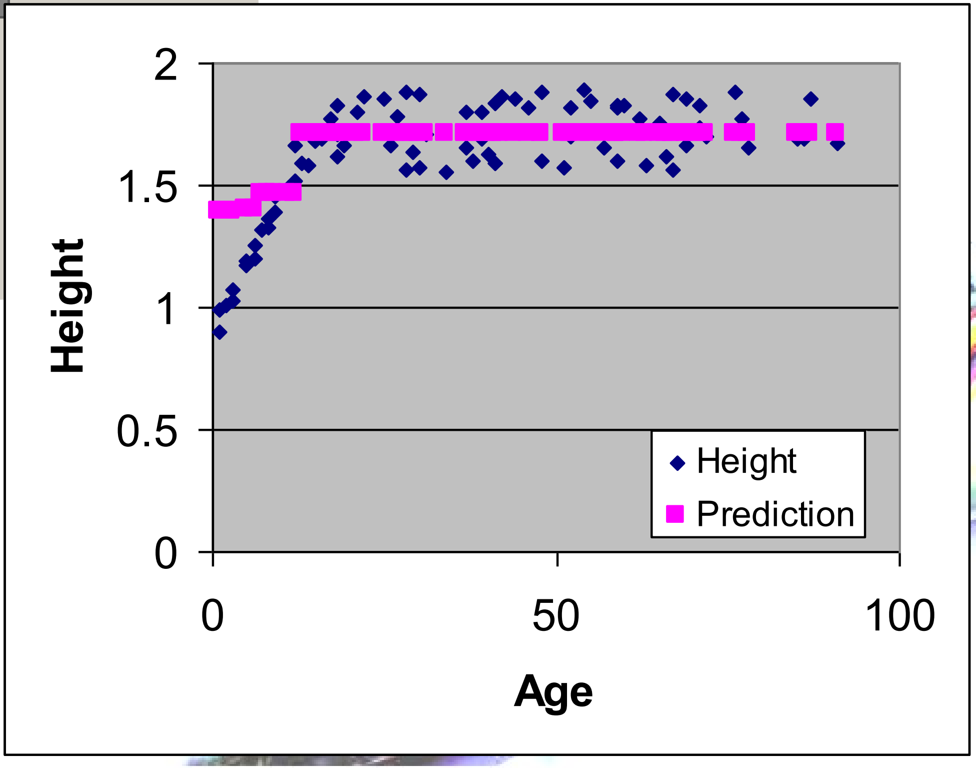
Linear Regression: prediction

Height = 0.0056 * Age + 1.4181

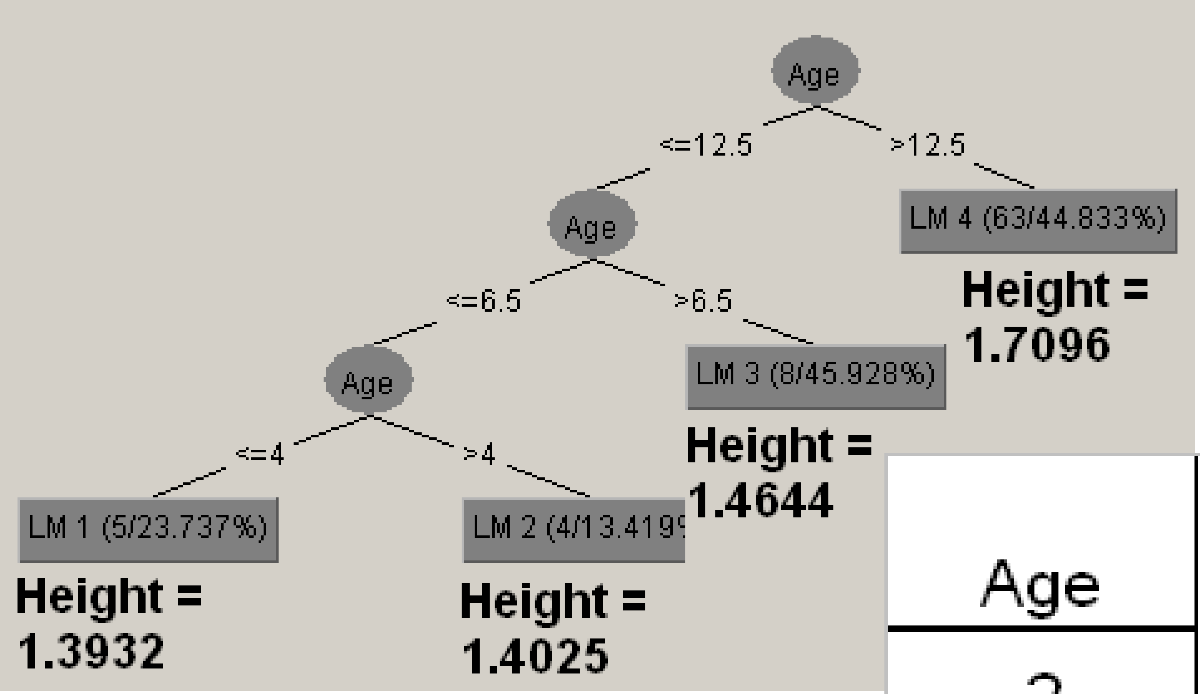
Age	Height	Linear regression
2	0.85	
10	1.4	
35	1.7	
70	1.6	

Regression tree



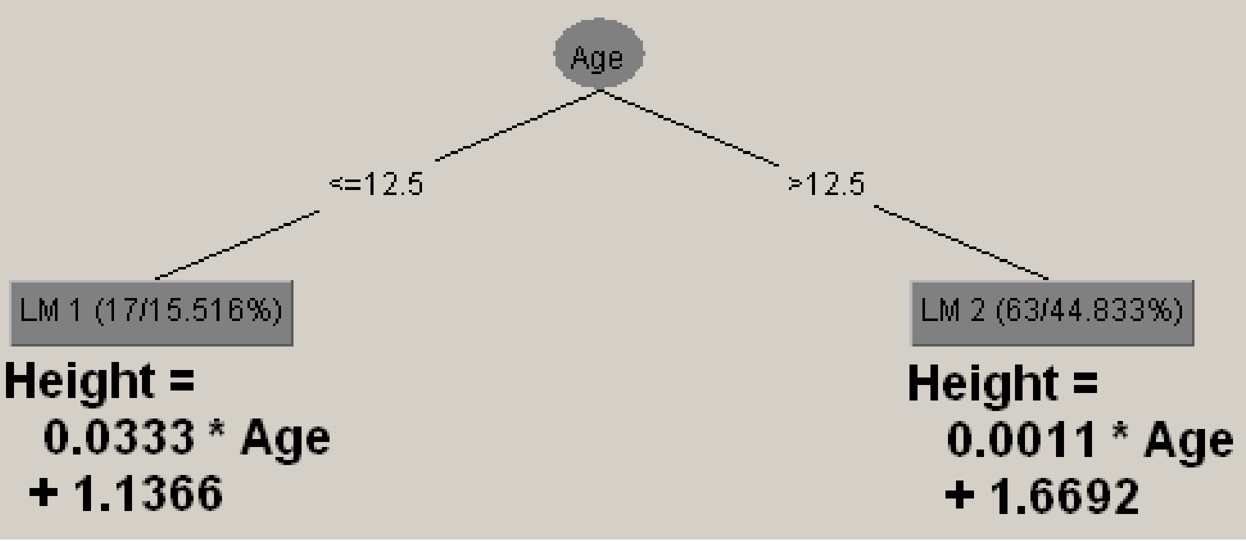


Regression tree: prediction

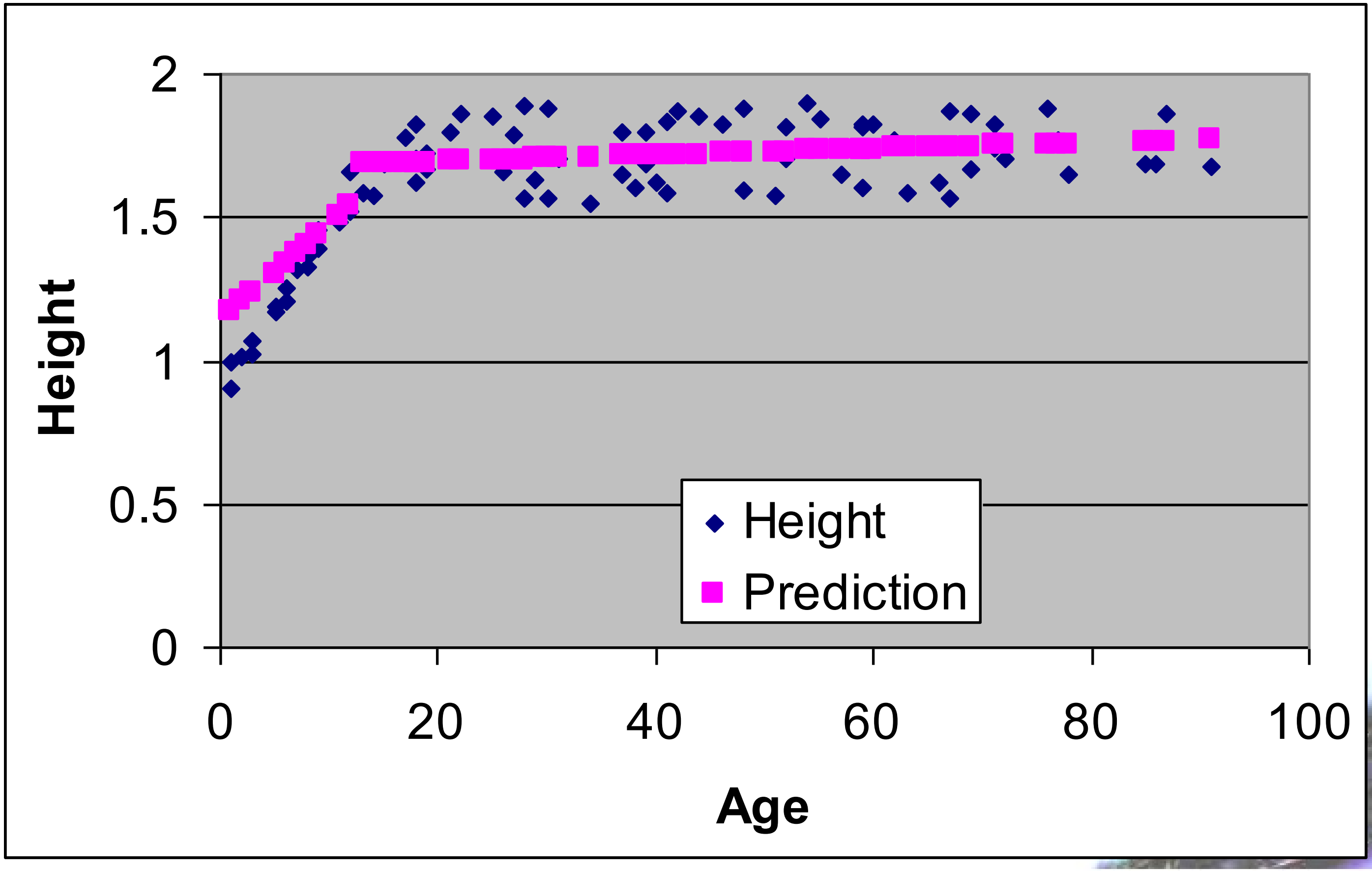


Age	Height	Regression tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	

Model tree



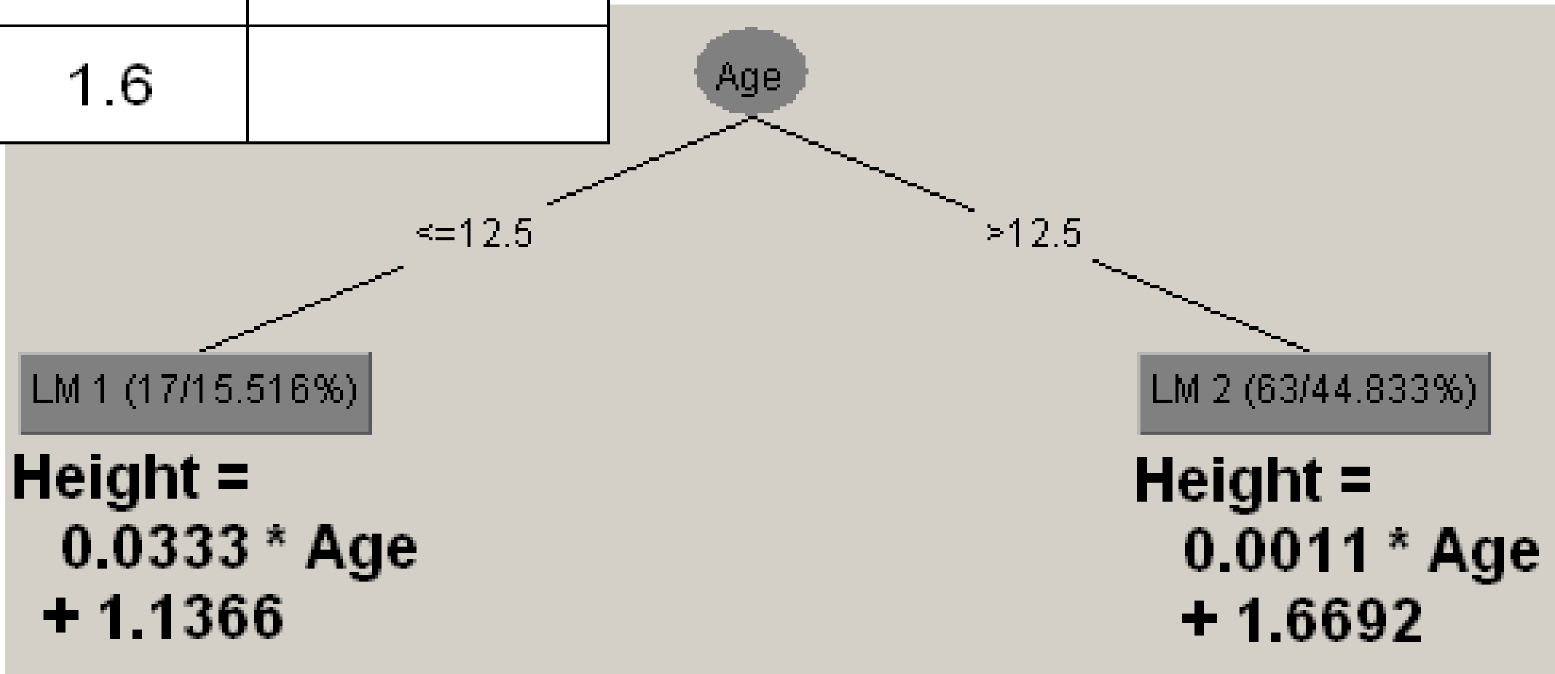
Age
<=12.5
>12.5
LM 1 (17/15.516%)
LM 2 (63/44.833%)
Height = 0.0333 * Age + 1.1366
Height = 0.0011 * Age + 1.6692



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

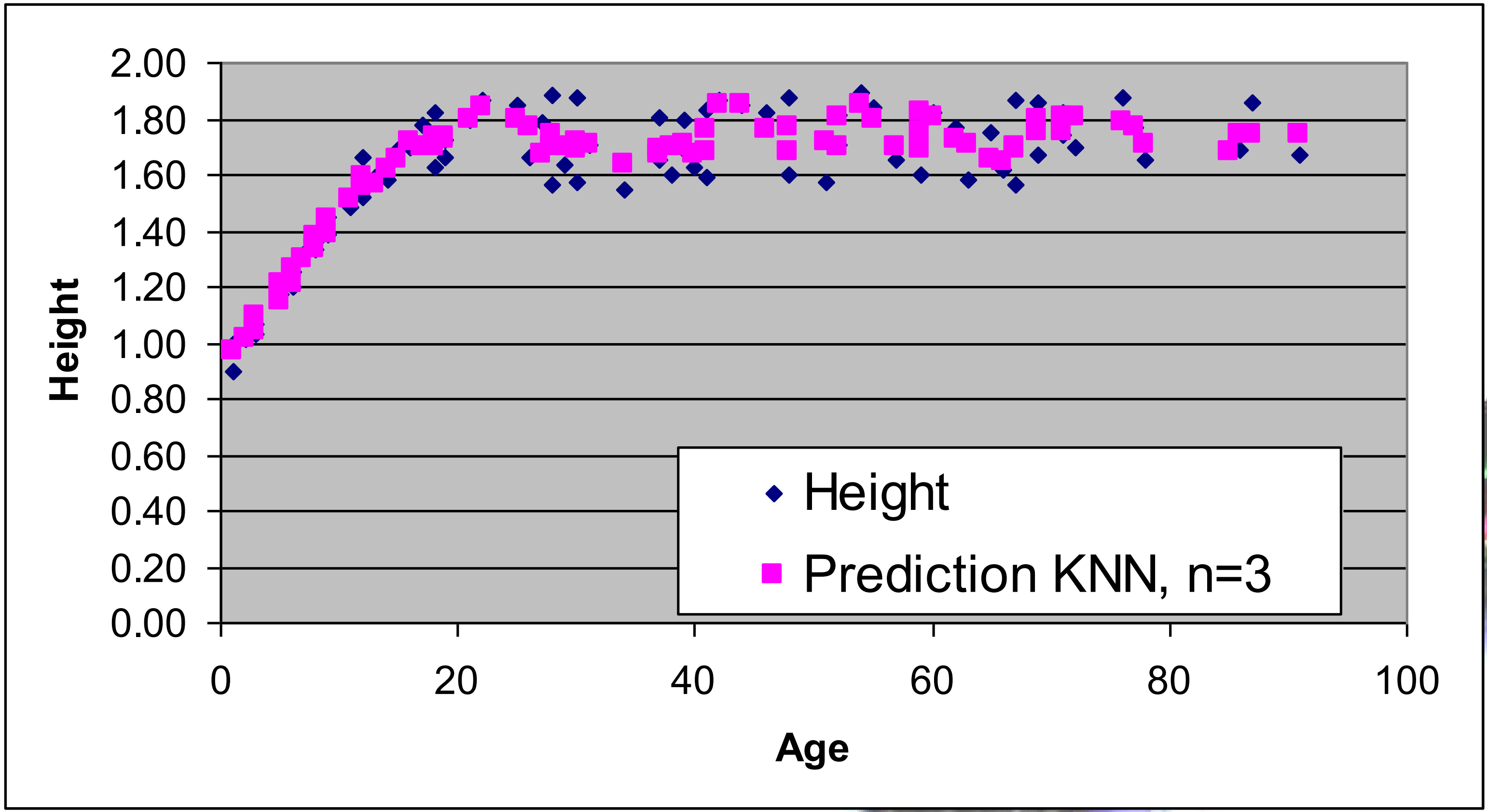
Model tree: prediction

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 non-target attributes) and predicts the average of their target variable
- In this example, K=3



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	1.00
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	$\dfrac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
root mean-squared error	$\sqrt{\dfrac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
mean absolute error	$\dfrac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
relative squared error	$\dfrac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\dfrac{1}{n}\sum_i a_i$
root relative squared error	$\sqrt{\dfrac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
relative absolute error	$\dfrac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
correlation coefficient	$\dfrac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\dfrac{\sum_i(p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_p=\dfrac{\sum_i(p_i-\bar{p})^2}{n-1}$, and $S_A=\dfrac{\sum_i(a_i-\bar{a})^2}{n-1}$

Numeric prediction discussion

- Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
  - Is this a classification or a numeric prediction problem?
  - What if such a variable is an attribute, is it nominal or numeric?
- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.

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,

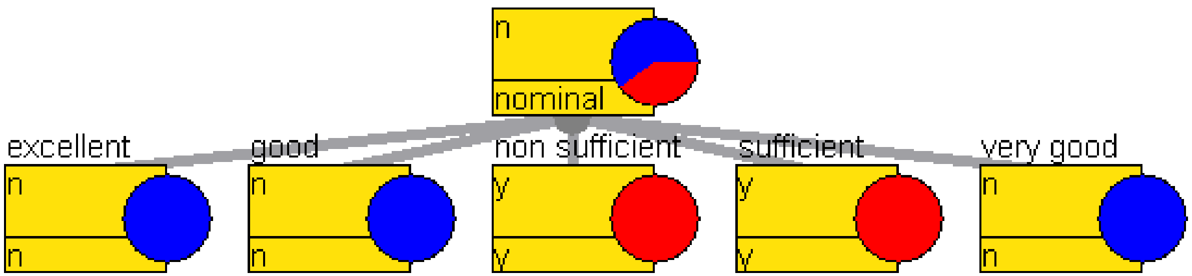
it should be considered numeric.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

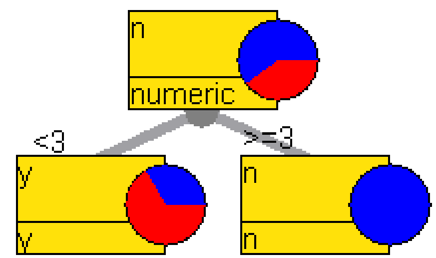
Nominal or numeric attribute?

- 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.

Numeric prediction discussion

- Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
  - Is this a classification or a numeric prediction problem?
  - What if such a variable is an attribute, is it nominal or numeric?
- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.

Can KNN be used for classification tasks?

- **YES**.
- In numeric prediction tasks, the average of the neighborhood is computed
- In classification tasks, the distribution of the classes in the neighborhood is computed

Numeric prediction discussion

- Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
  - Is this a classification or a numeric prediction problem?
  - What if such a variable is an attribute, is it nominal or numeric?
- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.

Similarities between KNN and Naïve Bayes.

- Both are “**black box**” models, which do not give the insight into the data.
- Both are “**lazy classifiers**”: they do not build a model in the training phase and use it for predicting, but they need the data when predicting the value for a new example (partially true for Naïve Bayes)

Numeric prediction discussion

- Consider a dataset with a target variable with five possible values:
    1. non sufficient
    2. sufficient
    3. good
    4. very good
    5. excellent
  - Is this a classification or a numeric prediction problem?
  - What if such a variable is an attribute, is it nominal or numeric?
- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

<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>

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**:

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

- **Confidence**:

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

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	

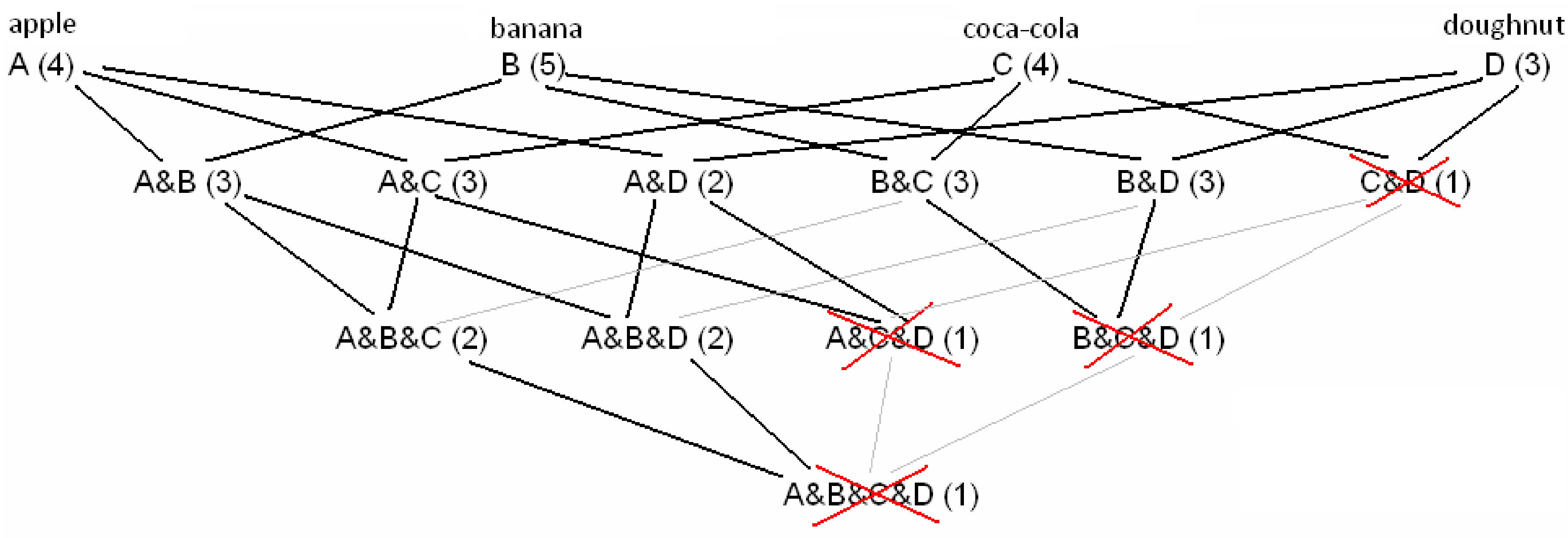
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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



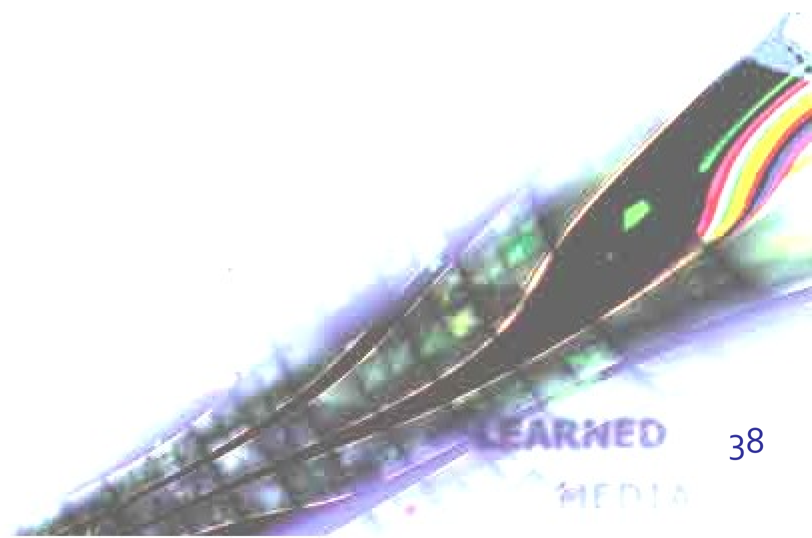
Frequent itemsets:

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

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$\rightarrow$Y) = Support (XY) = #XY / #D .... P(XY)

Confidence(X$\rightarrow$Y) = #XY / #X .... P(Y|X)

Lift(X$\rightarrow$Y) = Support(X$\rightarrow$Y) / (Support (X)*Support(Y))

Leverage(X$\rightarrow$Y) = Support(X$\rightarrow$Y) – Support(X)*Support(Y)

Conviction(X $\rightarrow$ Y) = 1-Support(Y)/(1-Confidence(X$\rightarrow$Y))

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.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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)