# Regel-gebaseerde regressie

## Capita Selecta computerwetenschappen
## Artificiële intelligentie (|H05N0a|)

Philippe Tanghe Li Quan

27 april 2011

# Overzicht

- optimalisatiecriterium Gecode: mean squared error
- regressiemodel: beslissingsboom + pruning

# Beslissingsboom

**Algorithm 1**: buildRegressionModel(data, itemsets, items)

**if** *nb instances of data* $\leq$ *MAX_NB_LEAF* **then**
    prediction $\leftarrow$ average (OR mean) of class values of instances
    add items to itemsets (as last rule)
**else**
    items2 $\leftarrow$ find best pattern with rimcp on data
    items2 $\leftarrow$ items2 $\cup$ items
    dataC $\leftarrow$ instances of data that cover items in items2
    dataNC $\leftarrow$ instances of data that don't cover items in items2
    buildRegressionModel(dataC, itemsets, items2)
    buildRegressionModel(dataNC, itemsets, items)
**end**

# Pruning

**Algorithm 2**: pruneRules(data, predOr, itemsets)

newPredOr ← predOr
**repeat**
    itemsets ← pruneRule(data, newPredOr, items)
    apply rules of itemsets on newPredOr
**until** *size new ruleset* == *size old ruleset*
return itemsets

**Algorithm 3**: pruneRule(data, predOr, itemsets)

bestRulesSoFar ← itemsets
tempData ← predOr
bestError ← error of predOr on data
**for** *each rule in itemsets* **do**
    tempRules ← itemsets \ rule
    apply rules of tempRules on tempData
    error ← error of tempData on data
    **if** *(error - RULECOST) < bestError* **then**
        bestError = (error - *RULECOST*)
        bestRulesSoFar ← tempRules
    **end**
**end**
return bestRulesSoFar

| prune | *RULECOST* | *MAX_NB_LEAF* | nb of rules | error trainingset | error testset |
|---|---|---|---|---|---|
| false | / | 3 | 33 | 67.33 | 158.77 |
| true | 5 | 3 | 13 | 87.90 | 158.47 |
| false | / | 5 | 20 | 85.92 | 143.73 |
| true | 5 | 5 | 12 | 99.04 | 142.38 |
| false | / | 7 | 15 | 92.35 | 141.78 |
| true | 5 | 7 | 10 | 106.33 | 139.90 |
| false | / | 10 | 11 | 102.05 | 139.84 |
| true | 5 | 10 | 9 | 109.06 | 137.96 |
| true | 0 | 3 | 33 | 67.33 | 158.77 |
| true | 0 | 5 | 20 | 85.92 | 143.73 |
| true | 0 | 7 | 15 | 92.35 | 141.78 |
| true | 0 | 10 | 11 | 102.05 | 139.84 |
| true | 3 | 3 | 15 | 80.13 | 158.47 |
| true | 3 | 5 | 14 | 92.03 | 143.73 |
| true | 3 | 7 | 12 | 99.35 | 141.78 |
| true | 3 | 10 | 11 | 102.05 | 139.84 |
| true | 10 | 3 | 12 | 94.9 | 158.47 |
| true | 10 | 5 | 8 | 124.32 | 142.38 |
| true | 10 | 7 | 8 | 118.53 | 143.39 |
| true | 10 | 10 | 8 | 114.29 | 137.96 |

Tabel: Voorspelling met gemiddelde. Invloed van de parameters op de training en test set error.

| prune | *RULECOST* | *MAX_NB_LEAF* | nb of rules | error trainingset | error testset |
|---|---|---|---|---|---|
| false | / | 3 | 33 | 54.20 | 127.00 |
| true | 5 | 3 | 11 | 89.30 | 128.20 |
| false | / | 5 | 20 | 67.60 | 125.55 |
| true | 5 | 5 | 11 | 89.65 | 126.90 |
| false | / | 10 | 11 | 91.10 | 118.15 |
| true | 5 | 10 | 10 | 99.04 | 126.40 |

Tabel: Voorspelling met mediaan. Invloed van de parameters op de training en test set error.