Regression trees

Model representation and loss

Herman Kamper

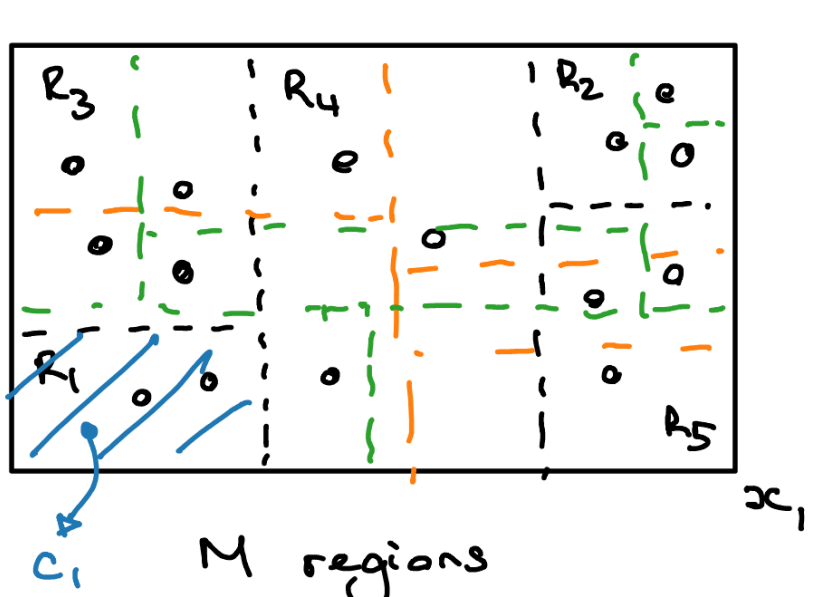
http://www.kamperh.com/

Regression tree model

Model: $f(\underline{x}; \underline{\theta}) = \sum_{m=1}^{M} c_m I\{\underline{x} \in R_m\}$

Loss: $J(\underline{\theta}) = \sum_{n=1}^{N} (y^{(n)} - f(\underline{x}^{(n)}))^2$

$= \sum_{n=1}^{N} (y^{(n)} - \sum_{m=1}^{M} c_m I\{\underline{x}^{(n)} \in R_m\})^2$



Parameters: $R_1, \ldots, R_M$, $c_1, \ldots, c_M$ $\Big\} \underline{\theta}$

How do we learn $\underline{\theta}$?

If we knew regions (but not the c's):

$\Rightarrow c_m = \frac{1}{N_m} \sum_{i:\, \underline{x}^{(i)} \in R_m} y^{(i)}$

N_m: # training items in R_m

Regression trees

Tree building algorithm

Herman Kamper

http://www.kamperh.com/

Tree growing algorithm:

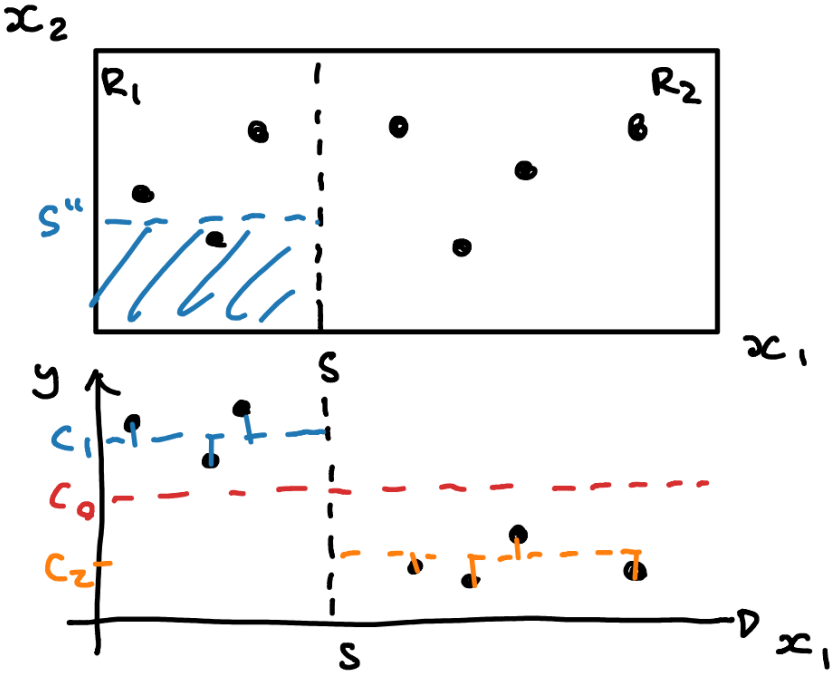
top-down

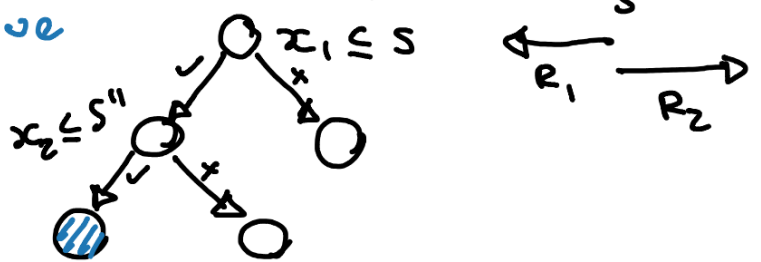
1. Start at top of tree (everything in one region)
2. for each leaf node (region):
   for each feature x_j and split point s:
   Calculate reduction in loss if we split there
3. Choose best (j, s) combination to split; create new child nodes (regions)

greedy

4. Repeat from (2) until stop condition is met

recursive





Example:

Before split: $J = \sum_{n=1}^{7} (y^{(n)} - c_0)^2$

Consider splitting x_1 at s:

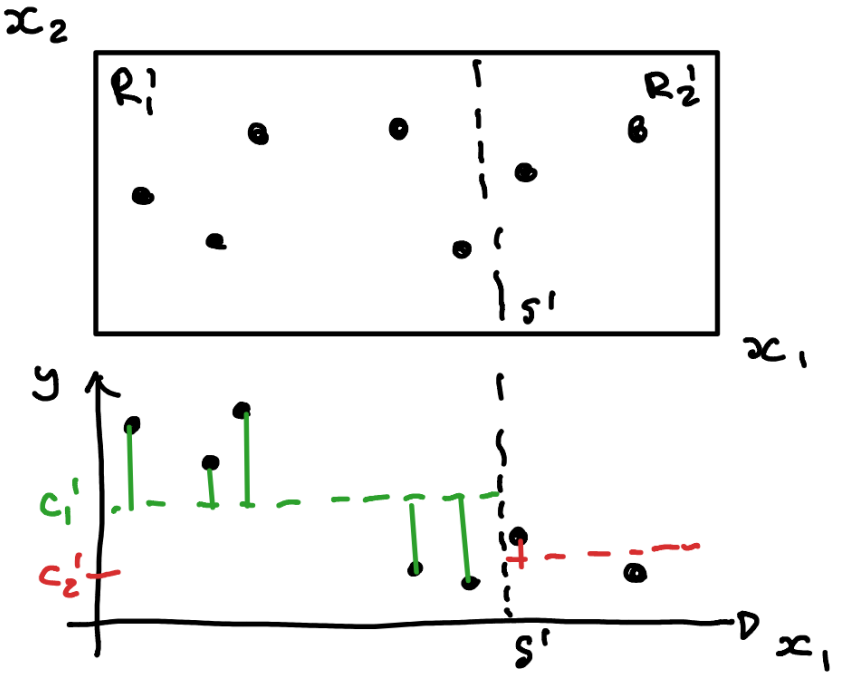
$R_1 = \{\underline{x} \mid x_1 \leq s\}$

$R_2 = \{\underline{x} \mid x_1 > s\}$

If we split: $J = \sum_{i:\underline{x}^{(i)} \in R_1} (y^{(i)} - c_1)^2 + \sum_{i:\underline{x}^{(i)} \in R_2} (y^{(i)} - c_2)^2$

Tree growing algorithm:

1. Start at top of tree (one region = everything in)
2. for each leaf node (region):
   for each feature x_j and split point s:
   Calculate reduction in loss if we split there
3. Choose best (j, s) combination to split; create new child nodes (regions)
4. Repeat from (2) until stop condition is met



$$\min_{j,s} \left\{ \sum_{i:\, \underline{x}^{(i)} \in R_1} (y^{(i)} - c_1)^2 + \sum_{i:\, \underline{x}^{(i)} \in R_2} (y^{(i)} - c_2)^2 \right\}$$

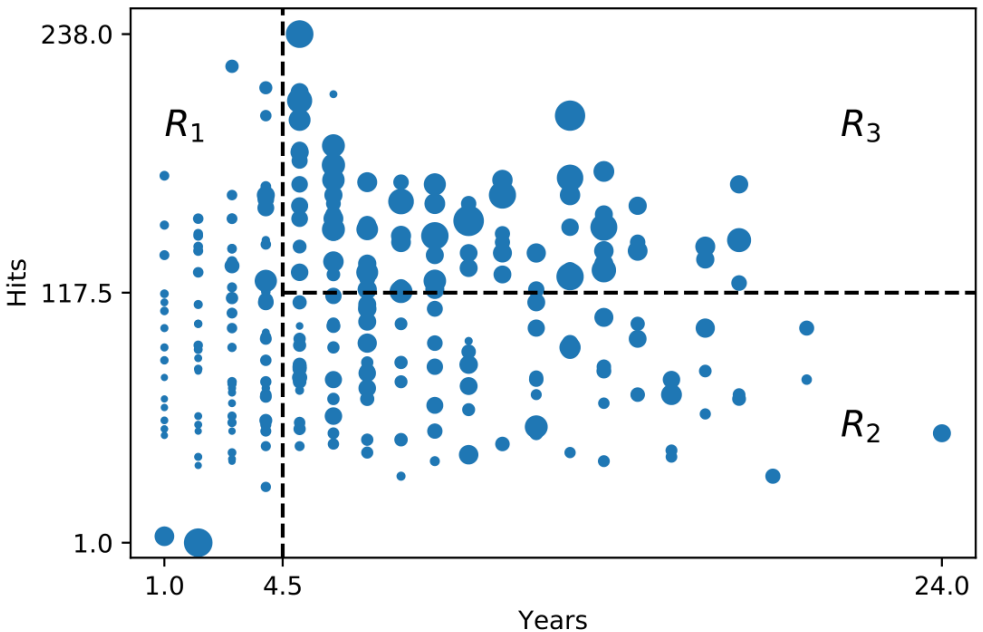
Regression trees

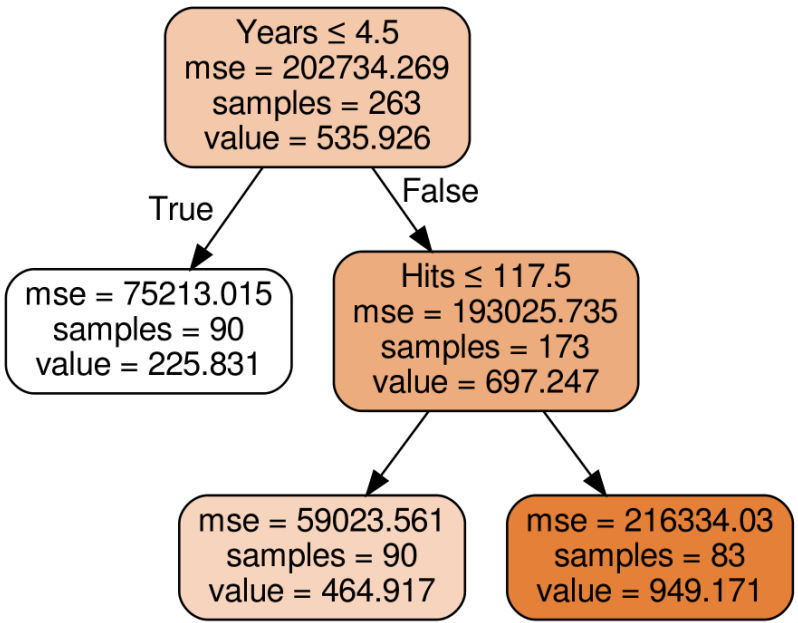
Regression trees in practice and tree pruning

Herman Kamper

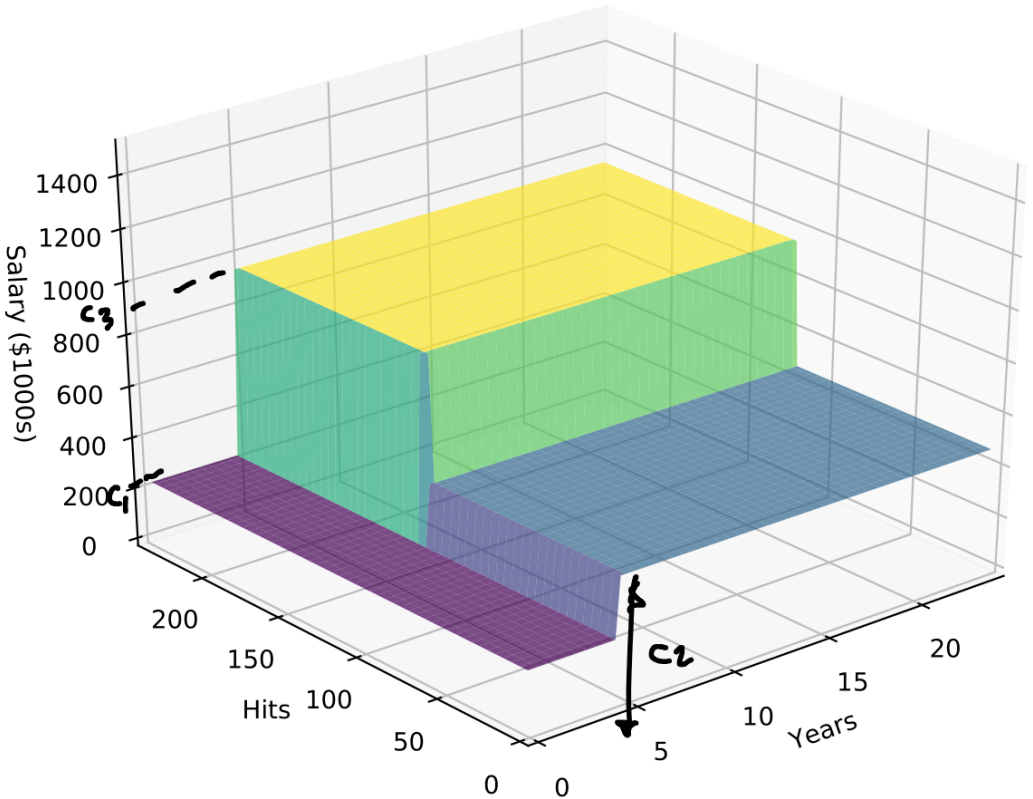
http://www.kamperh.com/

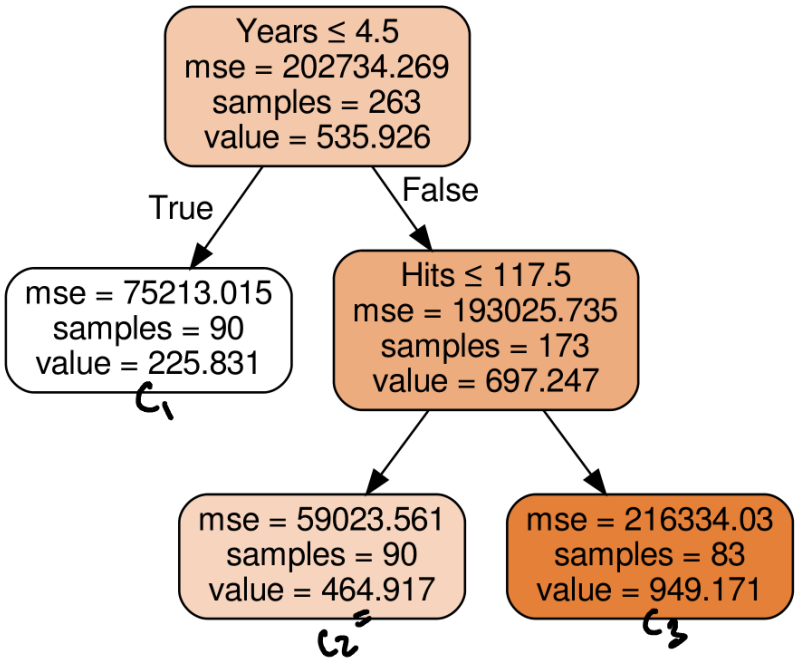
Regression tree on hitters data





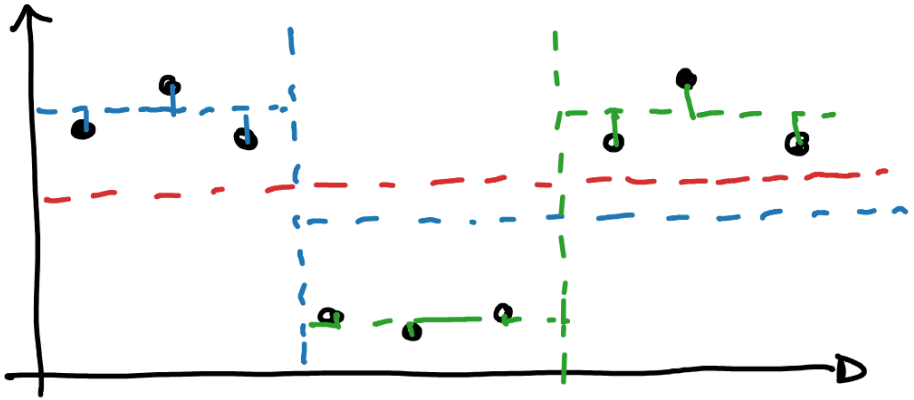
Regression tree on hitters data

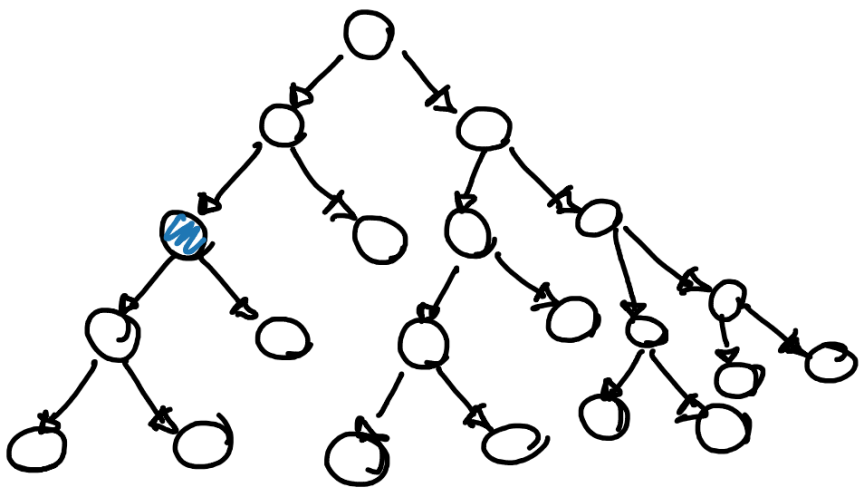




Regression trees in practice:

- Can easily overfit
- How do we regularise?
  - Can stop if gain is small
  - But this can be short-sighted
- Can use tree pruning





$$J = \sum_{m=1}^{M} \sum_{i:\, x^{(i)} \in R_m} \left(y^{(i)} - c_m\right)^2$$