Robust Regression---what we're doing

What LOESS robust regression does:

Assume y_k was originally generated using the following recipe:

With probability p:

$$y_k = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + N(0,\sigma^2)$$

But otherwise

$$y_k \sim N(\mu, \sigma_{huge}^2)$$

Computational task is to find the Maximum Likelihood β_0 , β_1 , β_2, p, μ and σ_{huge}

Mysteriously, the reweighting procedure does this computation for us.

Your first glimpse of two spectacular letters:

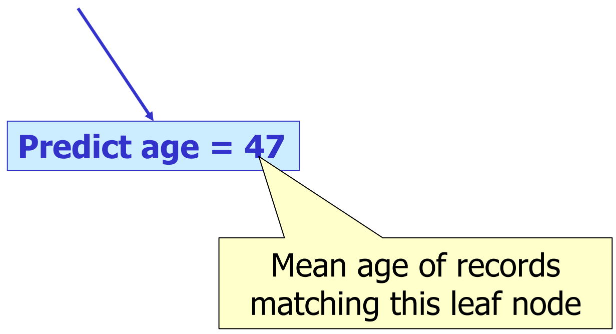
E.M.

Regression Trees

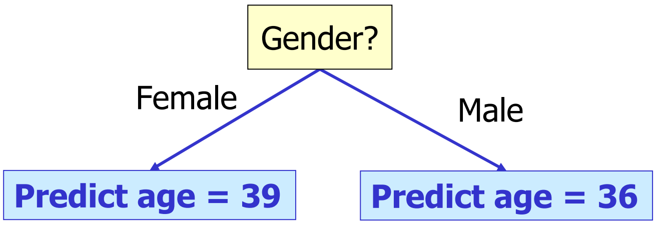
Regression Trees

- "Decision trees for regression"

A regression tree leaf



A one-split regression tree



Choosing the attribute to split on

Gender	Rich?	Num. Children	Num. Beany Babies	Age
Female	No	2	1	38
Male	No	0	0	24
Male	Yes	0	5+	72
:	:	:	:	:

- We can't use information gain.
- What should we use?

Choosing the attribute to split on

Gender	Rich?	Num. Children	Num. Beany Babies	Age
Female	No	2	1	38
Male	No	0	0	24
Male	Yes	0	5+	72
:	:	:	:	:

MSE(Y|X) = The expected squared error if we must predict a record's Y value given only knowledge of the record's X value

If we're told $x=j$, the smallest expected error comes from predicting the mean of the Y-values among those records in which $x=j$. Call this mean quantity $\mu_y^{x=j}$

Then…

$$MSE(Y \mid X) = \frac{1}{R} \sum_{j=1}^{N_X} \sum_{(k \text{ such that } x_k = j)} (y_k - \mu_y^{x=j})^2$$

Choosing the attribute to split on

Gender	Rich?	Num. Children	Num. Beany Babies	Age
Female	N			
Male	N			
Male	Y			
:	:			

Regression tree attribute selection: greedily choose the attribute that minimizes MSE(Y|X)

Guess what we do about real-valued inputs?

Guess how we prevent overfitting

MSE(Y|X) = The expected squared error if we must predict a record's Y value given only knowledge of the record's X value

If we're told $x=j$, the smallest expected error comes from predicting the mean of the Y-values among those records in which $x=j$. Call this mean quantity $\mu_y^{x=j}$

Then…

$$MSE(Y \mid X) = \frac{1}{R} \sum_{j=1}^{N_X} \sum_{(k \text{ such that } x_k = j)} (y_k - \mu_y^{x=j})^2$$

Pruning Decision

...property-owner = Yes

Gender?

Do I deserve to live?

Female

Predict age = 39

property-owning females = 56712
Mean age among POFs = 39
Age std dev among POFs = 12

Male

Predict age = 36

property-owning males = 55800
Mean age among POMs = 36
Age std dev among POMs = 11.5

Use a standard Chi-squared test of the null-hypothesis "these two populations have the same mean" and Bob's your uncle.

Linear Regression Trees

...property-owner = Yes

Also known as "Model Trees"

Gender?

Female

Predict age =

26 + 6 * NumChildren - 2 * YearsEducation

Male

Predict age =

24 + 7 * NumChildren - 2.5 * YearsEducation

Leaves contain linear functions (trained using linear regression on all records matching that leaf)

Split attribute chosen to minimize MSE of regressed children.

Pruning with a different Chi-squared

Linear Regression Trees

…property-owner = Yes

Gender?

Female

Also known as "M… l Trees"

Predict age =

26 + 6 * N…
2 * Year…

Detail: You typically ignore any categorical attribute that has been tested on higher up in the tree during the regression. But use all untested attributes, and use real-valued attributes even if they've been tested above

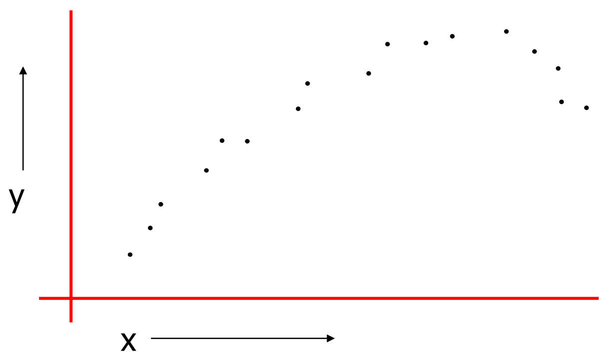
Leaves contain … functions (traine… linear regression … records matching t…)

… chosen to minimize … regressed children.

Pruning with a different Chi-squared

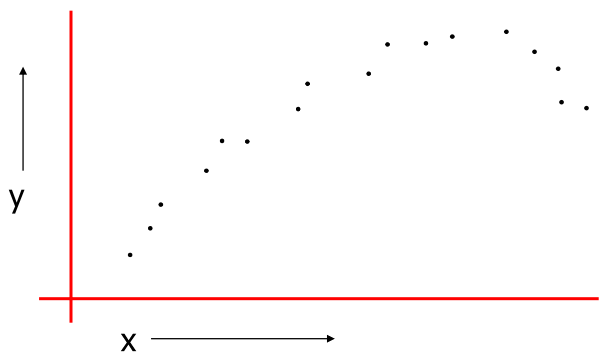
Test your understanding

Assuming regular regression trees, can you sketch a graph of the fitted function $y^{est}(x)$ over this diagram?



Test your understanding

Assuming linear regression trees, can you sketch a graph of the fitted function $y^{est}(x)$ over this diagram?



Multilinear Interpolation

