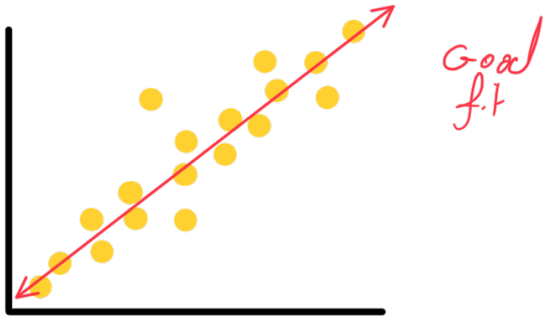
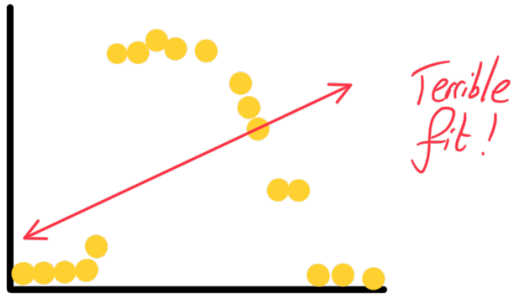
Regression Trees

Example

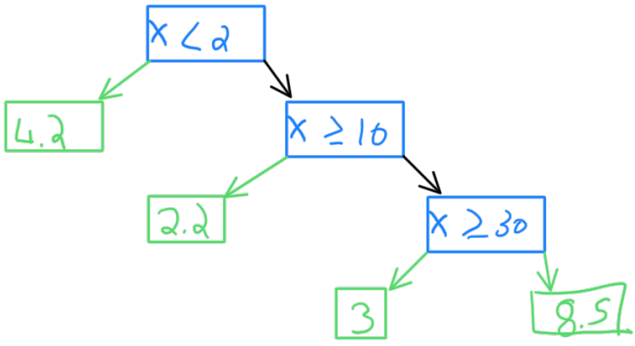




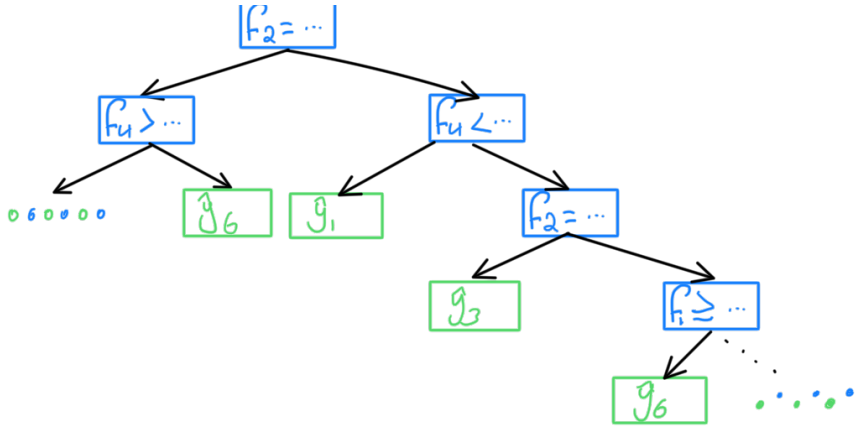
we want to fit this

Can't use Straight Line

Option: Regression Tree



In Practice

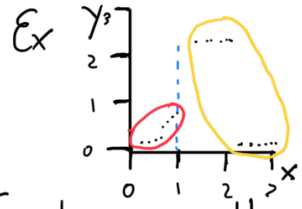


To build the Tree: (1 variable)

find a Split in feat. $f_1 \leq S_p$

Average Response on Both sides of the Split

Ex



$T(x) \Rightarrow$ $x < 1$

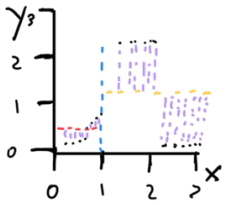
0.5 | 1.3

Avg $\approx$ 0.5

Avg $\approx$ 1.3

To choose a split $(x_i + x_{i+1})/2$

Then



$$SSR = \sum_i^N (y_i - T(x_i))^2$$

Repeat Choosing Splits for all different X's and store the SSR for each. Then the Split we choose for the Tree is the one with the lowest SSR

Now we repeat this process for both sides of T(x) using our SSR as the Max for the LHS and min for the RHS

We keep Making Nodes until we Reach our defined stopping.

* To prevent overfitting we generally limit the Minimum Size required to Split.

What if we have Multiple variables?

We calculate the Best SSR foreach feature then Choose the feature with the lowest SSR as our Splitting variable