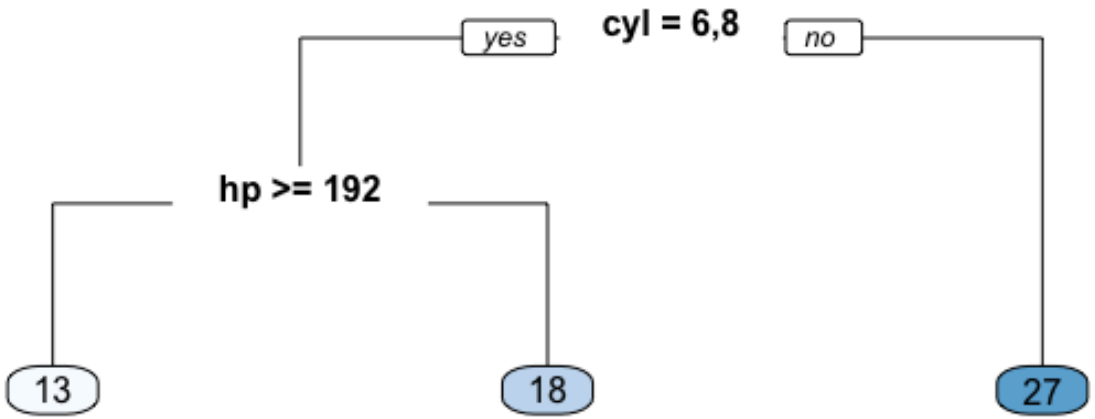
Decision trees for regression

- Decision trees are a powerful way to classify problems. On the other hand, they can also be adapted into regression problems.
- Decision trees built for a data set where the target column could be a real number, are called **regression trees**.
- Basic regression trees partition a data set into smaller groups and then fit a simple model (constant) for each subgroup.

Decision trees for regression

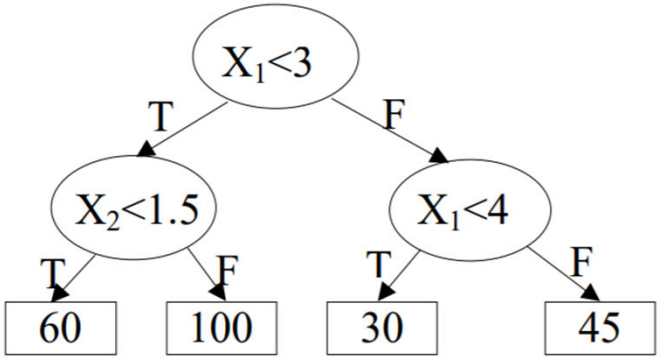
Example: we want to predict the miles per gallon a car will average based on cylinders (cyl) and horsepower (hp).



First, all observations that have 6 or 8 cylinders go to the left branch, all other observations proceed to the right branch. Next, the left branch is further partitioned by horsepower. Those 6 or 8 cylinder observations with horsepower equal to or greater than 192 proceed to the left branch; those with less than 192 hp proceed to the right. These branches lead to terminal nodes or leafs which contain our predicted response value.
In summary, all observations (cars in this example) that do not have 6 or 8 cylinders (far right branch) average 27 mpg. All observations that have 6 or 8 cylinders and have more than 192 hp (far left branch) average 13 mpg.

Decision trees for regression

Example:

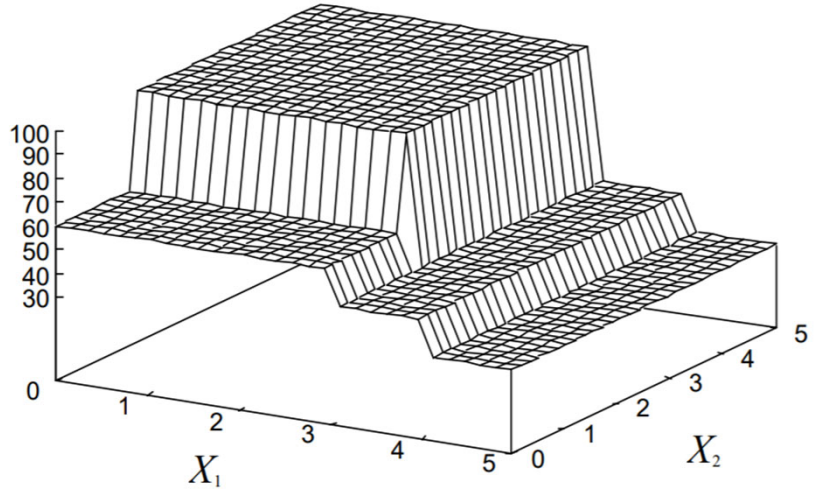


$P_1 \equiv X_1 < 3 \wedge X_2 < 1.5$, with $k_1 = 60$

$P_2 \equiv X_1 < 3 \wedge X_2 \geq 1.5$, with $k_2 = 100$

$P_3 \equiv X_1 \geq 3 \wedge X_1 < 4$, with $k_3 = 30$

$P_4 \equiv X_1 \geq 3 \wedge X_1 \geq 4$, with $k_4 = 45$



Decision trees for regression

Deciding on splits

The model begins with the entire data set, S, and searches every distinct value of every input variable to find the predictor and split value that partitions the data into two regions (R1 and R2) such that the overall sums of squares error are minimized:

$$\texttt{minimize}\left\{ SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2 \right\}$$

Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. This process is continued until some stopping criterion is reached (number of samples per node, depth, etc.).

Decision trees for regression

Example:

Outlook	Samples	Average	SSE
Rainy	5	35,6	283,2
Overcast	4	49,25	230,75
Sunny	5	41,8	956,8
			1470,75

Temp.	Samples	Average	SSE
Hot	4	37	340
Mild	6	41,5	777
Cool	4	45	614
			1731

Predictors | Target

Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	26
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

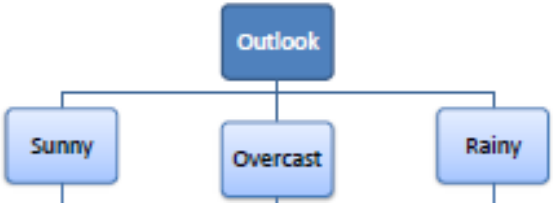
Humidity	Samples	Average	SSE
High	7	39,71	995,42
Normal	7	43,71	883,42
			1828,85

Windy	Samples	Average	SSE
False	8	43,5	782
True	6	39,333	1043,33
			1825,33

Decision trees for regression

Example:

Predictors				Target
Outlook	Temp	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



Exercise: compute the remaining splits to build the tree completely (stop criteria: samples per node < 5)

Decision trees for regression

Advantages :

- They are very interpretable.
- Making predictions is fast (no complicated calculations, just looking up constants in the tree).
- It is easy to understand what variables are important in making the prediction. The internal nodes (splits) are those variables that most largely reduced the SSE.
- There are fast, reliable algorithms to learn these trees.

Weaknesses:

- Single regression trees have high variance, resulting in unstable predictions (an alternative subsample of training data can significantly change the terminal nodes).