Decision Trees

Given the following data, build a decision tree with **three** leaves.

x	y
0	4
1	5
2	6
4	100

Use MSE as the measure of quality in the nodes. That means, we have an impurity

$$H(R) = \frac{1}{N}\sum_i (y^{(i)} - y_*)^2.$$

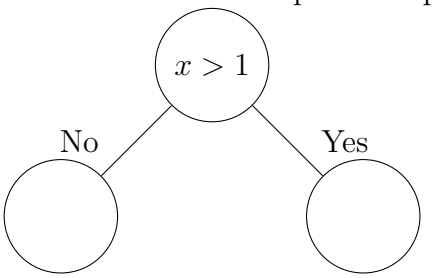
We proved that the minimum is given by $y_* = \bar{y}$. The quality of the split is given by

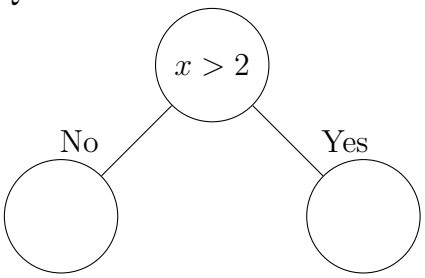
$$Q = H(R) - \frac{|R_l|}{|R|}H(R_l) - \frac{|R_r|}{|R|}H(R_r) \to max.$$

or

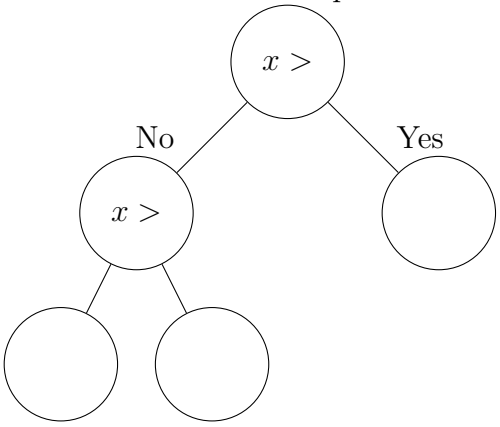
$$\tilde{Q} = \frac{|R_l|}{|R|}H(R_l) + \frac{|R_r|}{|R|}H(R_r) \to \min.$$

Problem 1. Given two possible splits calculate the impurity $\tilde{Q}$ in each case and choose the best split.





Problem 2. For the best split in Problem 1 build the tree with three leaves based on the impurity.



Problem 3. Make a sketch of the data and regression tree. Predict the value y for $x = 1.6$.

Problem 4. Construct a decision stamp for the following data

x	y
0.7	1
0.6	-1
0.7	-1
0.2	1
0.4	-1

Use Gini criterion for the impurity (instead of the entropy). Make a prediction y for $x = 0.5$.