ML Problems 14 03/06/2025

Decision Trees and Random Forest

Problem 1. Given the following data, build two decision stamps and use them for bagging. Use the following bootstrap sets: $x^{(1)}$, $x^{(1)}$, $x^{(2)}$, $x^{(2)}$, $x^{(3)}$, $x^{(6)}$, $x^{(7)}$, $x^{(8)}$ and $x^{(2)}$, $x^{(3)}$, $x^{(4)}$, $x^{(4)}$, $x^{(4)}$, $x^{(5)}$, $x^{(6)}$, $x^{(7)}$.

x_1	x_2	y	y_{pred}
1	2	1	
3	3	2	
2	1	1	
3	1	2	
6	3	4	
5	4	3	
7	2	6	
5	5	3	

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

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

The quality of the split is given by $\tilde{Q} = \frac{|R_l|}{|R|}H(R_l) + \frac{|R_r|}{|R|}H(R_r) \to \min$.

$$\tilde{Q} = \frac{|R_l|}{|R|}\frac{1}{|R_l|}\sum_{y^{(i)}\in R_l} |y^{(i)} - y_L| + \frac{|R_r|}{|R|}\frac{1}{|R_r|}\sum_{y^{(i)}\in R_r} |y^{(i)} - y_R| = \sum_{y^{(i)}\in R_l} |y^{(i)} - y_L| + \sum_{y^{(i)}\in R_r} |y^{(i)} - y_R|$$

Problem 2. Use the data above for the random forest regression. In addition to bootstrap, use feature 1 in the first tree and feature 2 for the second.

x_1	x_2	y	y_{pred}
1	2	1	
3	3	2	
2	1	1	
3	1	2	
6	3	4	
5	4	3	
7	2	6	
5	5	3	