DS 3001: Trees (Q6)

Q6. Show that any decision tree is equivalent to a linear regression on a set of dummy variables that represent the optimal splits in the tree.

Hint: You can think of CART as partitioning the feature space into a set of sets $\{S_1, S_2, \dots, S_K\}$, and then predicting the average for all of the observations in each set S_k, m_k:

$$\hat{y}(x) = \sum_{k=1}^{K} \mathbb{I}\{x \text{ is in } S_k\} m_k$$

where $\mathbb{I}\{P(x,k)\}$ takes the value 1 if the proposition $P(x,k)$ is true and 0 otherwise. Now, doesn't that look like least-squares regression on a set of dummy/one-hot-encoded variables?

Conversely, can any linear regression be represented by a tree?

given a point x.

$$\hat{y}(x) = m_k \quad \text{if} \quad x \in S_k$$

AKA: $\hat{y}(x) = \sum_{k=1}^{K} m_k \mathbb{I}_k(x)$

Design matrix. $D \in \mathbb{R}^{n \times k}$ each column = region in S_k
each row = data point in $x^{(i)}$

$$D_{ik} = \mathbb{I}_k(x^{(i)}) = \begin{cases} 1 & \text{if } x^{(i)} \in S_k \\ 0 & \text{otherwise} \end{cases}$$

fit regression model $\hat{y} = D\beta$, $\beta = [\beta_1, \beta_2, \beta_3, \dots, \beta_K]^T$

... gives: $\beta_k = \arg\min_{\beta} \sum_{i: x^{(i)} \in S_k} (y^{(i)} - \beta)^2 = \frac{1}{|S_k|} \sum_{x^{(i)} \in S_k} y^{(i)} = m_k$

$\therefore \hat{y}(x) = \sum_{k=1}^{K} \beta_k \mathbb{I}_k(x)$ when $\beta_k = m_k$

A linear regression can't always be written as a tree.

for example: $\hat{y}(x) = 3x_1 + 5x_2$

This is a continuous plane, not a piecewise.