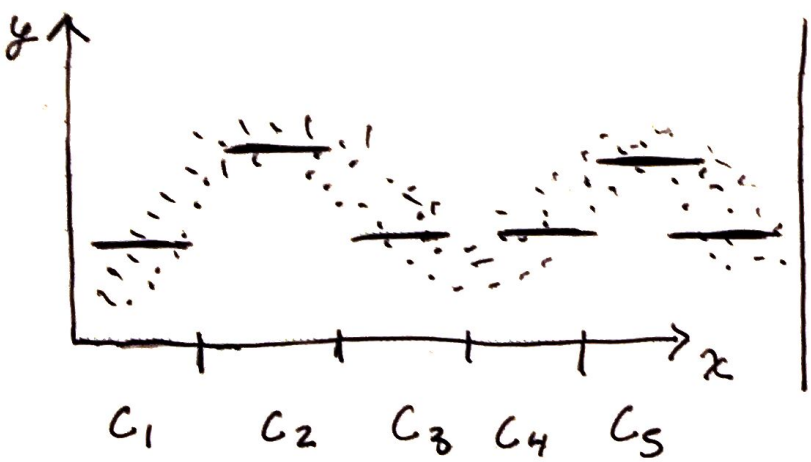
Math 342W Lecture 18

Classification and Regression Trees

Classification Trees: $y = \{C_1, C_2, \ldots, C_n\}$

Regression Trees: $y \subseteq \mathbb{R}$

Lets consider p_raw = 1 and build up Regression Trees...



$f: \mathbb{R}^p \to \mathbb{R}$

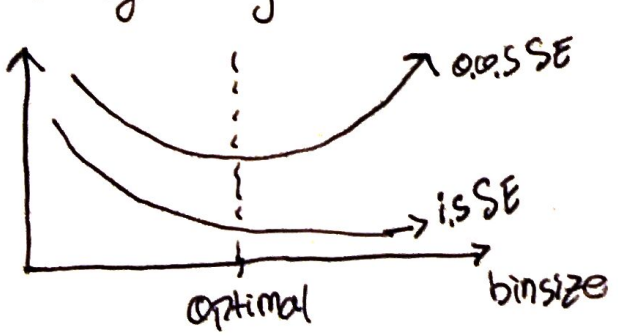
we start by splitting the domain and categorize it, choosing $\hat{y}$ within each category.

$$\mathcal{H} = \{ w_1 \mathbb{1}_{x \leq 1} + w_2 \mathbb{1}_{x \in (1,2]} + \ldots + w_p \mathbb{1}_{x > p} : \vec{w} \in \mathbb{R}^p \}$$

What choice of bin size would be best? → Hyperparameter

Large binsize → Underfit and Small binsize → Overfit

We can use the model selection procedure to optimize this by using oos Metrics.



With equally sized bins the total bins are defined by B^p, which is uncontrollably large.

↳ Change bins based on $\frac{d}{dx}$

We can restrict our bin splitting to be orthogonal to axis. In this way we can impose limits on the algorithm

Regression Trees Algorithm

① Begin with $D = \langle X, \vec{y} \rangle$

② Consider all possible orthogonal to axis splits $\langle X_\ell, \vec{y}_\ell \rangle$, $\langle X_r, \vec{y}_r \rangle$ with rules

$$\begin{array}{l} x_1 \le x_{11},\ x_1 \le x_{12},\ \dots,\ x_1 \le x_{1(n-1)} \\ x_2 \le x_{21},\ x_2 \le x_{22},\ \dots,\ x_2 \le x_{2(n-1)} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ x_p \le x_{p1},\ x_p \le x_{p2},\ \dots,\ x_p \le x_{p(n-1)} \end{array}$$

we only consider to the $(n-1)^{th}$ bin or else it would be 1 bin.

For each split there will be two daughter nodes. Compute SSE_ℓ and SSE_r

③ Locate split with lowest weighted average $\frac{n_L SSE_L + n_R SSE_R}{n_L + n_R}$

④ Assign $\hat{y}$'s to be $\bar{y}$'s of two bins

⑤ Recursively repeat ①→④ for daughter bins until bin has $n_0 \le$ observations, where n_0 is a hyperparameter

Classification Trees Algorithm

① " "

② " " $Gini_\ell$ & $Gini_r$

③ " " $\frac{n_L Gini_L + n_R Gini_R}{n_L + n_R}$

④ " " mode[y's]

$$Gini_L = \sum_{L=1}^{k} \hat{P}_L (1 - \hat{P}_L)$$

&

$$\hat{P}_L = \frac{\#\, y_i = L}{n_L}$$