Math 342W/642/742W

Recitation – Day #17 (4.10.25)

I. Trees

(i) What is a ***tree***?

A tree is a data structure made up of n nodes (vertices) and $n-1$ edges. The trees of interest will be rooted trees, where one node is the root/ancestor to all other nodes known as child nodes. These trees will provide a hierarchical structure for our models fitting data.

(ii) What is the tree-method are we interested in for machine learning?

The tree-method of interest is known as *decision* trees.

(iii) What tree-based algorithm will we be implementing?

CART – First introduced by Breiman in 1984.

(iv) What are the two types of trees we will be considering?

- Classification Trees: $\mathcal{Y} = \{C_1, C_2, \ldots, C_k\}$
- Regression Trees: $\mathcal{Y} = \mathbb{R}$

(v) What is being done to the predictor/feature space with this tree-based method?

We are stratifying/segmenting/splitting the predictor/feature space into a discrete number of simple region ("rectangles") based upon simple decision rules.

(vi) What are the advantages of tree-based models over the linear based models we have seen?

- simple to build, construct, and explain,
- binary tree representation mimics/mirrors human-decision making
- hierarchical structure can be visualized
- can take care of qualitative predictors without creating "*dummy*" variables

(vii) What are the disadvantages of tree-based models when compared with linear based models?

- predictive accuracy may not be as good as linear models
- suspectible to overfitting

II. Regression Trees

(i) What is the candidate set of functions $\mathcal{H}$ for regression trees? Compare that with the candidate set for the linear regression model and the logistic regression model.

Model	Candidate set $\mathcal{H}$
Regression Trees	$\left\{ \sum_{m=1}^{M} c_m \cdot \mathbb{1}_{x \in R_m} \,\middle|\, c_m \in \mathbb{R},\ R_1, \dots, R_m \text{ are partitions of feature space} \right\}$
Linear Regression	$\left\{ \sum_{i=1}^{p} w_i x_i + w_0 \,\middle|\, \boldsymbol{w} \in \mathbb{R}^{p+1} \right\}$
Logistic Regression	$\left\{ \frac{1}{1+e^{-\boldsymbol{w}\cdot\boldsymbol{x}}} \,\middle|\, \boldsymbol{w} \in \mathbb{R}^{p+1} \right\}$

(ii) How are the "*splits*" of the training data made?

The "splits" are made orthogonal with respect to the axes.

(iii) After each split is made, what is computed?

Calculate SSE for each node. Assign $\hat{y} = \bar{y}$ of the responses in the nodes.

$$\text{SSE}_{\text{node}} = \sum_{i \in \text{node}} (y_i - \bar{y}_{\text{node}})^2$$

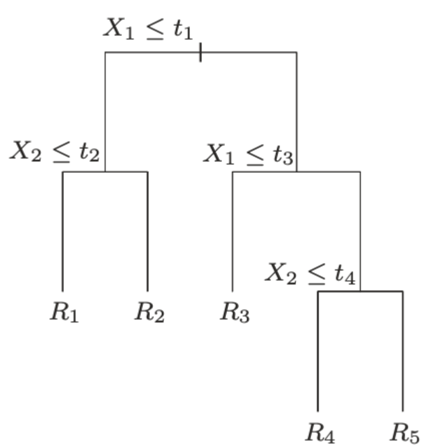
(iv) What are loss/objective function that we are trying to minimize to find the "best split"?

$$\text{SSE}_{weighted} = \frac{n_L}{n_L + n_R} \cdot \text{SSE}_L + \frac{n_R}{n_L + n_R} \cdot \text{SSE}_R$$

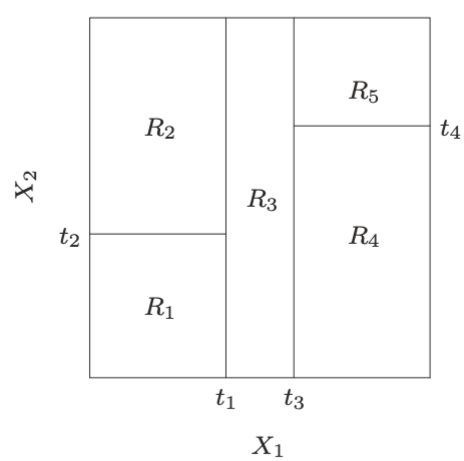
(v) What type of algorithm is CART described as?

CART is a *greedy* algorithm because it makes a locally optimal split at each iteration but may not be globally optimal.

(vi) Give a pictorial example of a regression tree and a partitioned feature space.



Regression Tree



Partitioned feature space

Figures are from *The Elements of Statistical Learning* by Hastie, Tibshirani, Friedman.