Tree-Classifier for Linear Regression (TCLR)

TCLR is a novel tree model proposed by Tong-yi Zhang and Bin Cao et al. (2021) to capture the functional relationships between features and target, through partitions the feature space into a set of rectangles, and embodies a specific function in each one. It is conceptually simple yet powerful for distinguishing mechanisms. The entire feature space is divided into disjointed unit intervals by hyperplanes parallel to the coordinate axes. In each partition, we model target y as the function of a feature $x_{\hat{j}}$ $(\hat{j} = 1, \cdots, \hat{m})\, \hat{m} \leq m$, linear function is used in our studied problem. It is worth noting that TCLR has the function of data screening by discarding which cannot be modeled functionally on some resulting leaves, since it’s a domain knowledge driven model. TCLR chooses features and split-points to attain the best fit and recursively binary partitions the space, until some stopping rules are applied.

How to grow a TCLR tree

TCLR can automatically decide on splitting features and points with mature criteria for splitting and pruning the tree. In each leaf, we model the response as a function of one feature $x_{\hat{j}}$ $(\hat{j} = 1, \cdots, \hat{m})\, \hat{m} \leq m$ every time, viz., $y = f\left(x_{\hat{j}}\right)$,where f is called “form prior” determined by domain knowledge.

If we implement our criterion of maximizing the Linearity Goodness (LG, but not linearity exclusive), then, finding the best partition in terms of maximum LG is generally computationally feasible. Different LG metrics in TCLR include the following:

The Pearson Correlation Coefficient(ρ):

$$\rho = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(x_{1\mathrm{i}} - \overline{x_1})(x_{2\mathrm{i}} - \overline{x_2})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_{1\mathrm{i}} - \overline{x_1})^2 \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_{2\mathrm{i}} - \overline{x_2})^2}} \tag{1}$$

The coefficient of determination, R-Square(R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_{1\mathrm{i}} - x_{2\mathrm{i}})^2}{\sum_{i=1}^{n}(x_{1\mathrm{i}} - \overline{x_1})^2} \tag{2}$$

Maximal Information Coefficient(MIC) [6]:

$$I(x_1, x_2) = \sum_{i=1}^{m}\sum_{j=1}^{k} P\big(x_1 = a_i, x_2 = b_j\big) log \frac{P(x_1=a_i, x_2=b_j)}{P(x_1=a_i)P(x_2=b_j)} \tag{3}$$

$$I^*(x_1, x_2) = \max(I(x_1, x_2)|S, m, k) \tag{4}$$

$$MIC(S) = max_{m \times k < B(n)} \left\{\frac{I^*(x_1, x_2)}{log(min\{m,k\})}\right\} \tag{5}$$

Where, x_1 and x_2 are variables, $a \in \{a_1, a_2, \cdots, a_m\}(m \leq n)$, denote m bins of variable x_1 and $b \in \{b_1, b_2, \cdots, b_k\}(k \leq n)$, denote k bins of variable x_2. The maximum is over all grids G with a columns and b rows, for given dataset. $B(n) = n^{0.6}$ is taken in TCLR.

Starting with all the data, consider a splitting variable j and splitting point i, and a child node pair of $R^l(x|x_j \leq i)$ and $R^r(x|x_j > i)$. The optimal splitting variable j and splitting point i of a given dataset with positive linear correlation are solved as: (if ρ<0, the smaller the better).

$$\max_{j,i}[\tfrac{1}{2}\Big(\underset{R^l}{\mathrm{LG}}\big(x_{\hat{j}}, f\big) + \underset{R^r}{\mathrm{LG}}\big(x_{\hat{j}}, f\big)\Big) - \underset{R^l+R^r}{\mathrm{LG}}\big(x_{\hat{j}}, f\big)] \tag{6}$$

For each splitting variable, the determination of splitting point i can be done very quickly and hence by scanning through all the inputs, determination of the best pair (j, i) is feasible. Having found the best split, we partition the data into the two resulting regions and repeat this process cyclistically. We give a pseudo-program as follows,

```
Algorithm: TCLR (takes $\rho > 0$ as an example)

fit Dataset $S = (x_1, x_2, \cdots, x_j, f)$
produce CHOOSE split point set $\{x_{ij}\}$ ($i$ for 1~n, $j$ for 1~m)
    best feature =$-1$
    best value =$-1$
    $\Delta LG$=0
        For $x_{IJ}$ in $\{x_{ij}\}$
           produce Sub-dataset $\{S_l, S_r\}$
              If $\left|\frac{\rho(x_j^l, f^l) + \rho(x_j^r, f^r)}{2}\right| - |\rho(x_j, f)| \geq \max(\Delta LG, \min inc = 0)$
                  return $\Delta LG$=$\left|\frac{\rho(x_j^l, f^l) + \rho(x_j^r, f^r)}{2}\right| - |\rho(x_j, f)|$
                         best feature = $j$
                         best value =$i$
              end if
        end for
        return ($i$, $j$)
end produce
```

The TCLR tree must be pruned, obviously a very large tree might overfit the data, while a small one might not capture important information. Thus, tree size is a tuning parameter governing the model’s complexity. Let $|T|$ denote the number of terminal nodes and LG_m denote the LG on the m-th leaf. We define the utility criteria as:

$$U(TCLR) = \sum_{m=1}^{|T|} LG_m - \beta|T| \tag{7}$$

The tuning parameter $\beta \geq 0$ governs the tradeoff between tree size and its goodness of fit to the data. For each β, there is a unique tree. β is regulated by three parameters in TCLR, viz., [|min-inc|] (default=0.01) is the minimum increment of Gains Function in Eq.(S3.4), [|Threshold|] (default=0.95), when LG in a node is equivalent to or higher than the present threshold, the node will become a leaf. And [min-size] (default=3), the amount of data in a leaf shouldn’t be smaller than the minimum size.