
H2O.ai

Regularization: Observation and Feature Sampling

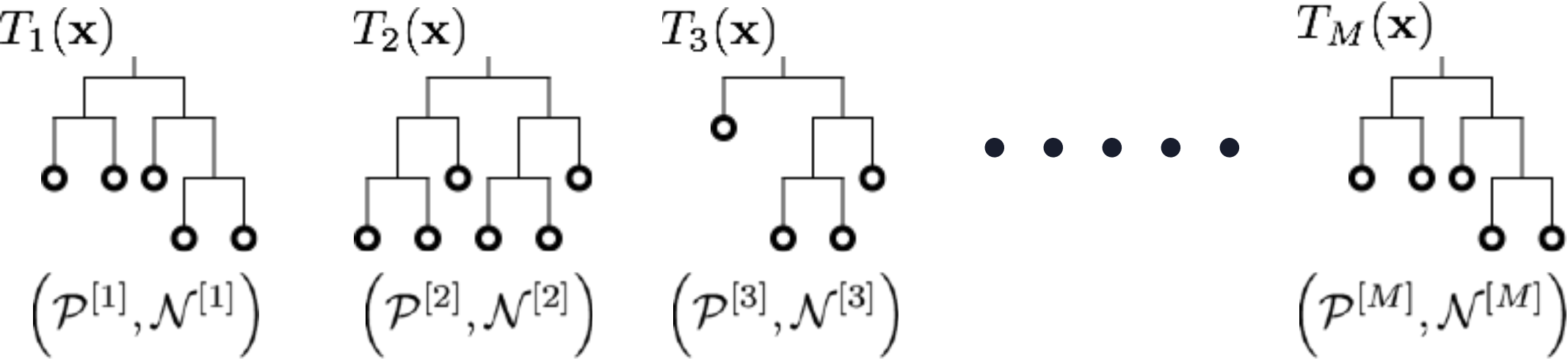
- For each tree, select from *P* features (columns) and *N* rows

$$\mathcal{P} \in \{1, 2, \ldots, P\}$$

$\mathcal{N} \in \{1, 2, \ldots, N\}$

$$\mathcal{P}^{[i]} \subseteq \mathcal{P}, \ \left|\mathcal{P}^{[i]}\right| = P^{[i]} = \gamma_{\mathrm{COL}} P$$

$$\mathcal{N}^{[i]} \subseteq \mathcal{N},\ |\mathcal{N}^{[i]}| = N^{[i]} = \gamma_{\mathrm{ROW}} N$$


$T_1(\mathbf{x})$
$T_2(\mathbf{x})$
$T_3(\mathbf{x})$
$T_M(\mathbf{x})$
$\left(\mathcal{P}^{[1]}, \mathcal{N}^{[1]}\right)$
$\left(\mathcal{P}^{[2]}, \mathcal{N}^{[2]}\right)$
$\left(\mathcal{P}^{[3]}, \mathcal{N}^{[3]}\right)$
$\left(\mathcal{P}^{[M]}, \mathcal{N}^{[M]}\right)$

Scoring and Stopping

- How often should we check our validation error? (Computation time versus generalization)
  - **`score_each_iteration`**: score model after each tree
  - **`score_tree_interval`**: score model after n trees

- Setting criteria for stopping
  - **`stopping_rounds`**: early stop if stopping metric's moving average does not improve for this many rounds
  - **`stopping_metric`**: metric for early stopping
  - **`stopping_tolerance`**: Relative tolerance for metric-based stopping criterion (stop if relative improvement is not at least this much)
  - **`max_runtime_secs:`** maximum runtime to allow for model building

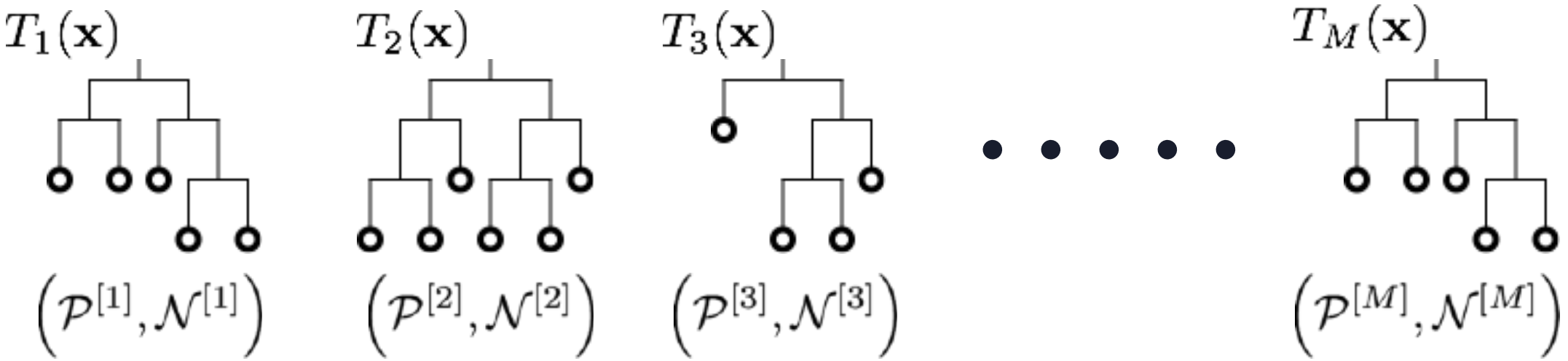
H2O.ai

Regularization: Observation and Feature Sampling

- For each tree, select from *P* features (columns) and *N* rows

$$\mathcal{P} \in \{1, 2, \ldots, P\} \qquad \mathcal{N} \in \{1, 2, \ldots, N\}$$

$$\mathcal{P}^{[i]} \subseteq \mathcal{P},\ |\mathcal{P}^{[i]}| = P^{[i]} = \gamma_{\text{COL}} P \qquad \mathcal{N}^{[i]} \subseteq \mathcal{N},\ |\mathcal{N}^{[i]}| = N^{[i]} = \gamma_{\text{ROW}} N$$



H2O.ai