Decision Tree API: A summary

Shehzeen S. Khan, J028
B.Tech Data Science, Semester 5

Default Parameters:

*,criterion='gini',splitter='best',max_depth=None,min_samples_split=2,min_samples_leaf=1,min_weight_fraction_leaf=0.0,max_features=None,random_state=None,max_leaf_nodes=None,min_impurity_decrease=0.0,min_impurity_split=None,class_weight=None,ccp_alpha=0.0

Parameters:

- criterion: measures the quality of the fit, gini for Gini impurity and “entropy for information gain.
- splitter: choses the split at each node, can be random or best fit
- max_depth: sets maximum depth of the tree
- min_samples: minimum number of required to split at each internal node
- min_samples_leaf: minimum number of samples required at each leaf node, mayaffect smoothing the model, especially regression
- min_weight_fraction_leaf: the minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when sample_weight is not provided
- max_features: number of features to consider when looking for best fit
- random_state: controls randomness of the estimator
- max_leaf_nodes: best nodes for relative reduction of impurity
- min_impurity_decrease: will be split if this split induces a decrease of the impurity greater than or equal to this value
- min_impurity_split: threshold to stop tree growth
- class_weight: assigns weight to classes if none is set then it will consider that the data is balanced
- ccp_apha: complexity parameter used for Minimal Cost-Complexity Pruning. The Subtree with the largest cost complexity that is smaller than ccp_alpha will be chosen

Attributes:

- classes: the classes labels (single output problem), or a list of arrays of class labels (multi-output problem)
- feature_importances_: return the feature importance
- max_features_int: returns the inferred value of max_features.
- n_classes: the number of classes (for single output problems), or a list containing the number of classes for each output (for multi-output problems).
- n_features: the number of features when fit is performed.
- n_outputs: the number of outputs when fit performed.
- tree: underlying Tree object