# Summary: Decision Tree

|  |  |  |
| --- | --- | --- |
|  | Classification | Regression |
| Sample data |  |  |
| 1 |  | **Sort the predictors from small values to big ones, and plot it out:** |
| 2 | **Grow trees for each predictor individually** | **Grow trees for each threshold of every predictor:**  *the threshold is the “mean” value for neighbouring points* |
| 3 | **Obtain the “Gini” index for each tree**  *The “Gini” index can be calculated as*  Where:   * is the probability of “yes” in a leaf * is the probability of “no” in a leaf | **Obtain the error for each tree (or threshold)**  *Using the tree (threshold based) to produce the prediction and calculate the error as* |
| 4 | **Based on the Gini index, determine the “root” to split the tree**  *Usually, the predictor showing the smallest Gini is used as “root”* | **Based on the Error, determine the “root” to split the tree**  *Usually, the threshold showing the smallest error is used as “root”* |
| 5 | **Further split the tree using the similar method, until we are not able to split**  *- Grow tree for every remained predictor*  *- Calculate “Gini” for each tree*  *- Get the tree with the smallest Gini*  *- Repeat the above process …* | **Further split the tree using the similar method, until we are not able to split**  *- Calculate the threshold from neighbouring points, and grow trees based on thresholds*  *- Error for each tree*  *- Get the tree with the smallest error*  *- Repeat the above process …* |
| Note:   * In order to avoid overfitting, we may limit the number of levels of the tree to grow | | |