Decision Trees

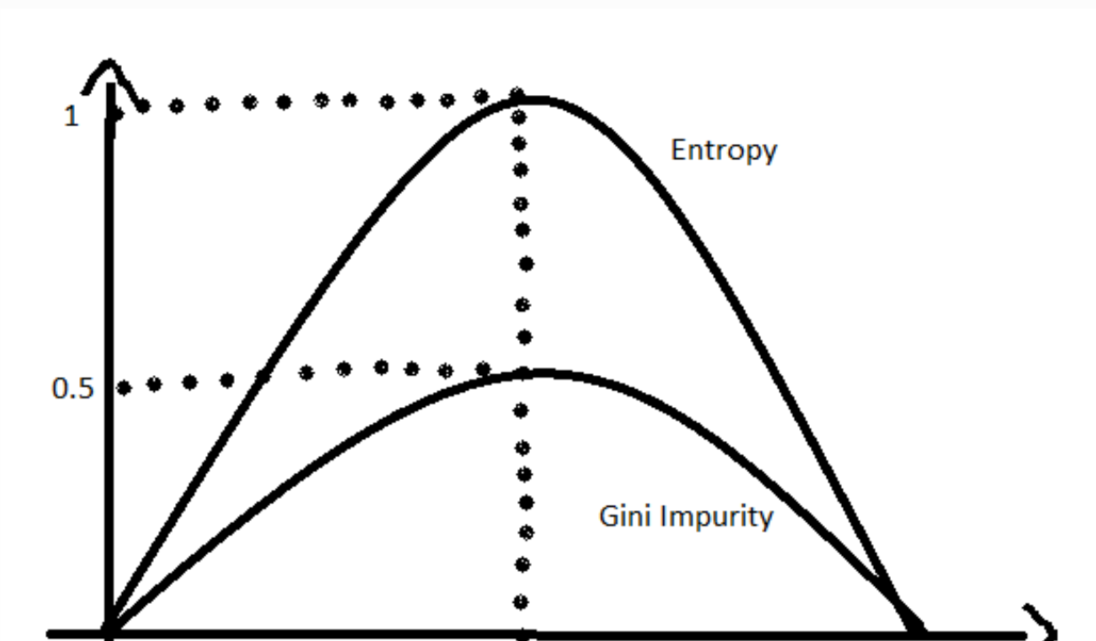
⇒ used to predict Categorial variables

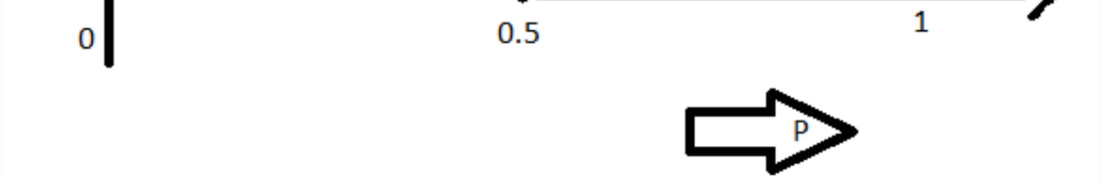
Measures of impurity

Gini Impurity

Entropy

Variance



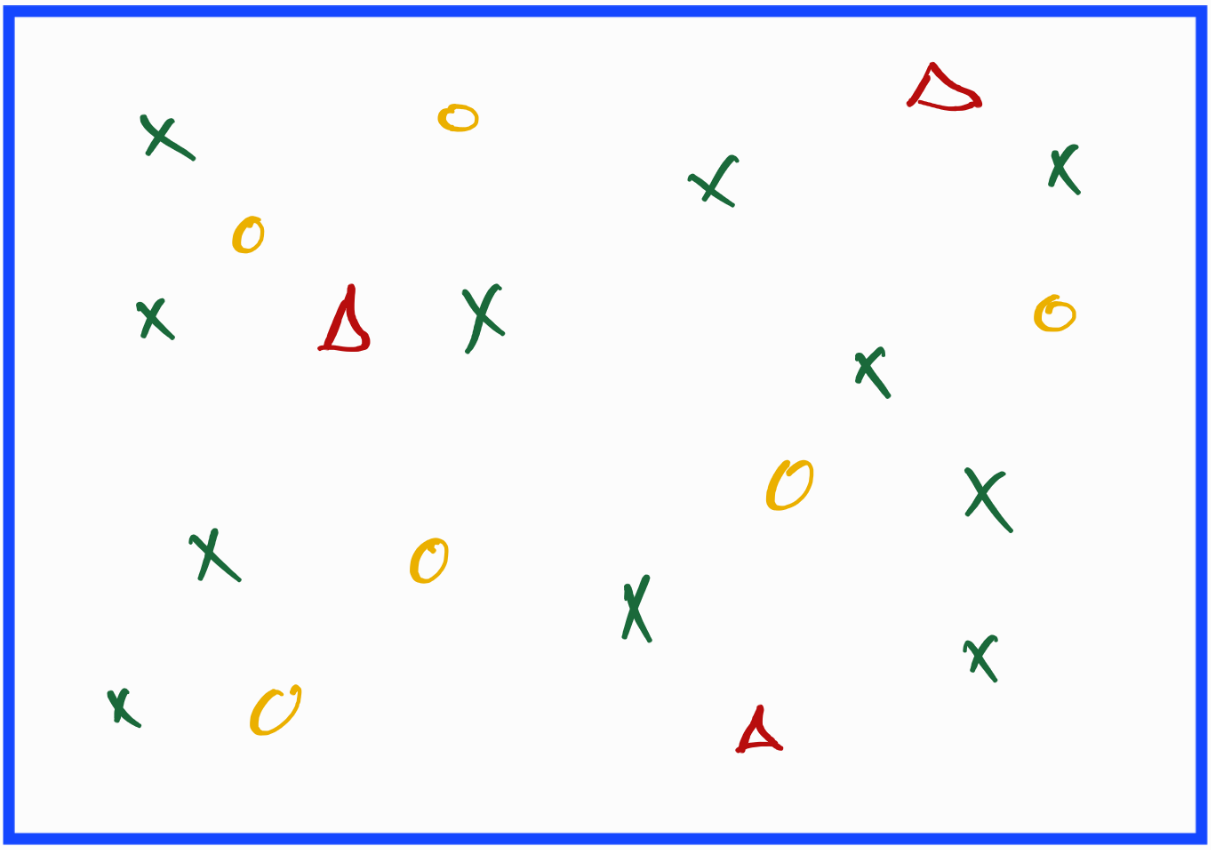


$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

⇒ Gini Impurity

Measures the probability of incorrectly classifying an element in a dataset.



Pruning

=> Pre-Pruning

* Prevents a tree from building below or about certain threshold)

Hyper Parameters

Parameters used to control the learning process.

1. Max_depth: the higher the value the more complex the tree.

2. Min_Sample_Split: Minimum number of samples required to split a node

3. Minimum_Samples Leaf: Minimum samples required at a leaf node.

Hyper Parameter Tuning

- **Grid Search (PG)** Searching for the best combination of hyper-parameters from a pre-defined set of values.

- PG (Parameter Grid)

- Grid Search CV( )

Cross Validation: Pre-Pruning

Training	Test

Model 1

Model 2

Model 3

Used model on test data and

see how well it does.

Split data into N groups and run N experiments. On each experiement, N-1 groups are used to train the model and the model is tested on the group left out.

⇒ Cost Complexity Pruning : Post Pruning

* Starting from a full-tree, create a sequence of trees that are sequentially smaller.

* At each step of the Algorithm:

① try removing possible sub tree

② Compute the Complexity Parameter: (α)

③ Remove subtree with minimum(α)

* with the list of subtrees one usually reverts back to using cross-validation errors to find the best final pruned tree

$$\alpha = \frac{Error\begin{pmatrix}\text{Pruned}\\ \text{tree}\end{pmatrix} - Error\begin{pmatrix}\text{Original}\\ \text{tree}\end{pmatrix}}{\text{no. of nodes reduced}}$$

⇒ as α increases

impurity increases

2) $\boxed{\text{Complexity decreases}}$

3) Accuracy decreases

Impurity Measures

⇒ Gini index : Classification

easy to compute

$$1 - \sum p_i^2$$

0 ← | → 0.5

⇒ Entropy : Classification

$$-\sum p_i \log(p_i)$$

Expensive

0 ← → 1

$\Rightarrow$ Information Gain: Classification

$$E(Y) - E(Y|X)$$

expensive

$0 \longleftarrow\longrightarrow 1$

$\Rightarrow$ Variance : Regression

$$\frac{\sum (x - \bar{x})^2}{N}$$

most common measure of spread.

$$\geq 0$$