Decision Trees

Preface

Decision Trees are graphical machine learning models that are used to predict a record's target feature value (represented by leaves) from the other values (aka non-target feature values) of that record (represented by branches). A node in a decision tree can either be an internal (i.e. non-leaf) node or an external (i.e. leaf) node. Internal nodes are labelled with a non-target feature, whereas external nodes are labelled with a target feature value (e.g. survived/died).

There are two main types of decision trees:

- Class-based – where the target feature values are classes (i.e. discrete).
- Regressive – where the target feature values are real numbers (i.e. continuous).

Algorithms for constructing decision trees usually work top-down, by choosing the decision rule at each step that best splits the distinct values of the target feature. Decision Trees are considered white-box models, meaning that there is an identifiable structure, because each split is clearly explainable. Different algorithms use different metrics for measuring the best split.

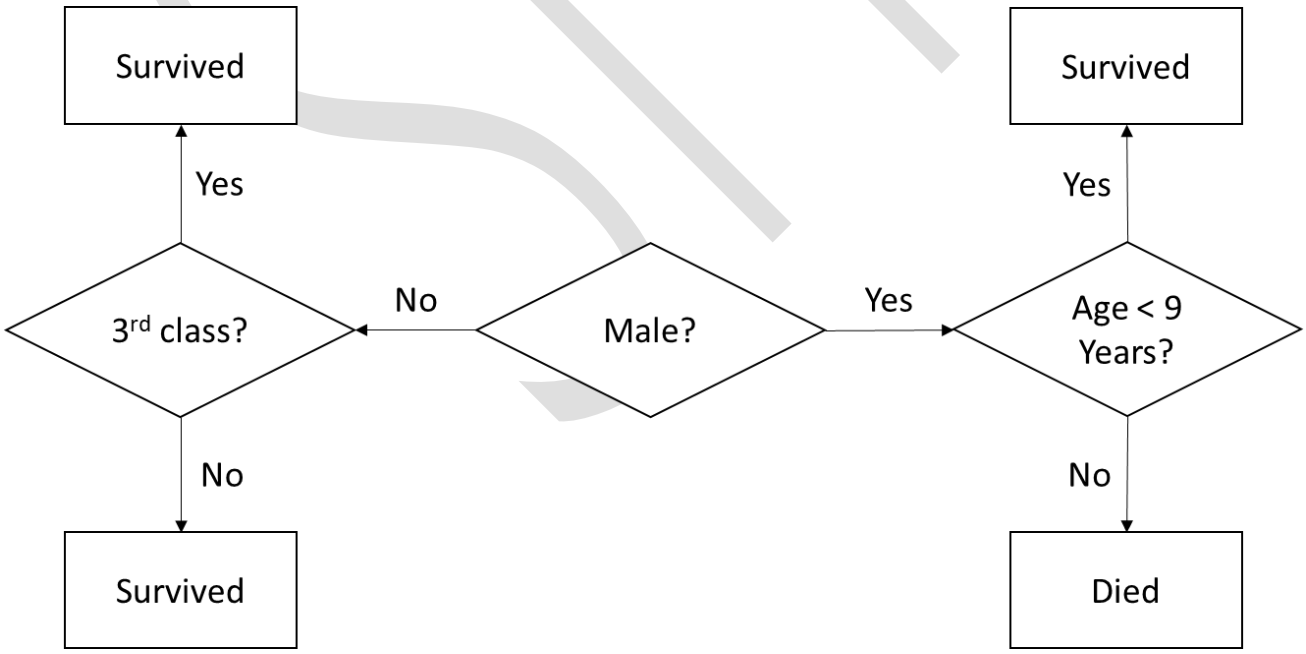

Figure 1: Decision Tree Representing Titanic Survivors.

Gini Impurity Gain

Gini Impurity is a measure of the impurity of a dataset's target feature values. A high Gini Impurity is an indicator that the target feature values of records in a dataset are highly impure. More literally, Gini Impurity represents the chance of a target feature being chosen at random multiplied by the chance that it is miscategorised by the Decision Tree. Gini Impurity for a dataset can be computed as follows:

$$I = \sum_{i=1}^{J} P(i) \times \sum_{k \neq i} P(k) \quad \text{(Eq.1)}$$

$$\sum_{k \neq i} P(k) = 1 - P(i) \quad \text{(Eq.2)}$$

Where:

- I – Gini Impurity metric.
- J – number of distinct target feature values.
- $P(i)$ – probability of the correct target feature value being chosen.
- $\sum_{k \neq i} P(k)$ – probability that the correct target feature is miscategorised.

Gini Impurity gain (i.e. impurity reduction) for a split can be computed as follows:

$$G = I_b - \left(\sum_{i=1}^{J} I_{a_i} \times \frac{n_i}{n_t} \right) \quad \text{(Eq.3)}$$

Where:

- G – Gini Impurity gain.
- I_b – Gini Impurity before the split.
- J – number of nodes created by the split.
- I_{a_i} – Gini Impurity of a node created by the split.
- n_i – number of records in a node created by the split.
- n_t – total number of records.

Information Gain

The concept of Information goes along with the concept of Entropy, which is a measure of uncertainty; in this case, uncertainty about a dataset's target feature values. A

high Entropy is an indicator that the target feature values of records in a dataset are highly uncertain.

Information is inversely proportional to entropy, so a high level of Information is an indicator that the target feature values of records in a dataset are highly certain. Entropy for a dataset can be computed as follows:

$$H = -\sum_{i=1}^{J} P(i) \times \log_b P(i) \quad \text{(Eq.4)}$$

Where:

- H – Entropy metric.
- J – number of distinct target feature values.
- $P(i)$ – probability of the correct target feature value being chosen.
- b – base constant (typically 2).

Information Gain (i.e. Entropy reduction) for a split can be computed as follows:

$$I = H_b - \left(\sum_{i=1}^{J} H_{a_i} \times \frac{n_i}{n_t} \right) \quad \text{(Eq.5)}$$

Where:

- I – Information gain.
- H_b – Entropy before the split.
- J – number of nodes created by the split.
- H_{a_i} – Entropy of a node created by the split.
- n_i – number of records in a node created by the split.
- n_t – total number of records.

Variance Reduction

Variance Reduction is often employed in cases where the target feature is, meaning that use of many other metrics would first require discretisation before being applied. The Variance Reduction of a node is defined as the total reduction of the variance of the target feature due to the split at this node. Variance Reduction for a dataset can be computed as follows:

$$I_V(N) = \frac{1}{|S|^2} \sum_{i=S} \sum_{j=S} \frac{1}{2} \times (x_i - x_j)^2 \quad \text{(Eq.6)}$$
$$- \left(\frac{1}{|S_t|^2} \sum_{i=S_t} \sum_{j=S_t} \frac{1}{2} \times (x_i - x_j)^2 + \frac{1}{|S_f|^2} \sum_{i=S_f} \sum_{j=S_f} \frac{1}{2} \times (x_i - x_j)^2 \right)$$

Where:

- I_V– Variance Reduction metric.
- N– Variance Reduction metric.
- S – set of pre-split sample indices.
- S_t– set of sample indices for which the split test is true.
- S_f – set of sample indices for which the split test is false.
- x– target feature.