Symbolists

Symbolists are a machine learning tribe rooted in logic. **Decision trees** are their most popular algorithm. They are also known for their **ensembles** and **random forests**.

Decision Trees

- A **graph without cycles**, where a **node represents a test on the value of a single variable** - each node corresponds to a **logical proposition** and **each branch represents the outcome of the test**;
- Terminal nodes are called **leaves**, and represent the values assumed by the target variable - the **class**;
- With numeric variables, it is easy to see that a decision tree can also be seen as a **partition of the space** of the variables;
- The decision tree can also be seen as a **model of the decision process** that leads to the classification of an object.

In summary:

- **Records** represented as **conjunctions of propositions**;
- **Training** consists of **learning the decision tree**;
- **Classification** consists of **following the path** from the root to the leaf that corresponds to the record.

Training Algorithms

- Trees are **built recursively** - a tree is composed of **subtrees**;
- Algorithms recursive, **top-down** and **divide-and-conquer**.

1. If all records in the node belong to the **same class**, then the node is a **leaf** and the class is the value of the leaf;

2. Otherwise, we create a node with the best variable to discriminate between classes - the most **relevant** variable.

Over the years, several algorithms have been proposed to build decision trees. The most popular are:

- **ID3** (Quinlan, 1986) - only considers **discrete variables**;
- **C4.5** (Quinlan, 1993) - extension of ID3 and now considers numeric variables;
- **CART** (Breiman et al., 1984) - considers both discrete and numeric variables, and for each variable, it considers several tests.

Choosing the Best Variable

- Its crucial to **choose the best variable** to discriminate between classes;
- The metrics used to choose the best variable must satisfy some conditions:
  - 0 when the node is **pure** - all records belong to the same class;
  - 1 when the node is **impure** - records are evenly distributed among classes;
  - $x \in [0, 1]$ for other cases;
- There are several criteria to choose the best variable:
  - **Entropy** - a measure to quantify the uncertainty associated with the value taken by a variable, when only its distribution is known;
    - Most used criteria;
    - Quantifies the number of bits needed to encode the class of a record - lower entropy means less information;
    - **Gain Ratio** - **normalized information gain**;
  - **Gini Index** - a metric that measures the **impurity** of a node - the probability that a randomly chosen record is incorrectly classified;
  - **Chi-square** - a metric that measures the **independence** between a variable and the class.

Pruning

- **Pruning** is a technique to **reduce the size** of a decision tree by **removing** nodes that do not provide **additional information**;
- Used to avoid **overfitting** - the tree is too complex and fits the training data too well;
- **Pre-pruning** - the tree is pruned during the construction phase;
- **Post-pruning** - the tree is pruned after the construction phase.