Decision Tree

Properties

1. Hierarchical Model
2. Employs on Divide and Conquer strategy
3. Non-parametic
4. Greedy Algorithm

Impurity Measures

1. Gini Index
   $$G = 1 - \sum_{i=1}^{n} p_i^2$$
2. Entropy
   $$H = -\sum_{i=1}^{n} p_i log_2 \left(p_i\right)$$

Split

1. Univariate: Splits the node using a single attribute.
   - Binary Split: Splits the node into two nodes, uses a single threshold
   - Mult Split: splits the node into multiple nodes, requires multiple thresholds
2. Multivariate: Uses multiple attributes

Classification

The attributes and threshold are selected based on the infogain, once a node is pure it is not split further. Usually, an attribute is used on once on a given subtree to split a node, there can however be cases where you would want to use an attribute more than once within the same subtree. While classifying, we compare the attributess and threshold and then move to subtrees iteratively until we encounter a leaf node which is assigned a label, the sample is classified as that label.

Ovefitting

If the classifier overfits, you can try prunning. Either pre-prunning or post-prunning. 1. Pre-prunning: call a node pure (leaf) if it exceeds a certain threshold of purity. 2. Post-prunning: Grow the tree and remove the leaf nodes that cause overfitting.

Pre-prunning is easier to implement, its fast. On the other hand, post-prunning is much more accurate but it's harder, slower.

Advantages of Decision Tree

Explanability: Often we want to know why the classifier has classified a given sample as something. Decision trees have a high explanability as they are just a set of rules.

Questions

1. Can decision trees be used for regression?
2. Should you or should you not use an attributes more than once within a subtree?
3. Should you or should you not use PCA for pre-processing data for Decision Tree?
4. Decision trees have boundaries along given axes, can they have boundaries as some angle from those axes?
5. How do you find the best tree? (generic)