

# Algoritmos para descobrir melhor split

## Algoritmos de árvore: ID3, C4.5, C5.0 e CART ¶

Quais são todos os vários algoritmos de árvore de decisão e como eles diferem uns dos outros? Qual deles é implementado no scikit-learn?

[ID3](#) (Iterative Dichotomiser 3) foi desenvolvido em 1986 por Ross Quinlan. O algoritmo cria uma árvore multiway, encontrando para cada nó (ou seja, de maneira gananciosa) o recurso categórico que produzirá o maior ganho de informação para alvos categóricos. As árvores crescem até seu tamanho máximo e, em seguida, geralmente é aplicada uma etapa de poda para melhorar a capacidade da árvore de generalizar para dados invisíveis.

C4.5 é o sucessor do ID3 e removeu a restrição de que os recursos devem ser categóricos, definindo dinamicamente um atributo discreto (baseado em variáveis numéricas) que particiona o valor do atributo contínuo em um conjunto discreto de intervalos. C4.5 converte as árvores treinadas (ou seja, a saída do algoritmo ID3) em conjuntos de regras se-então. A precisão de cada regra é então avaliada para determinar a ordem em que devem ser aplicadas. A poda é feita removendo a pré-condição de uma regra se a precisão da regra melhorar sem ela.

C5.0 é a versão mais recente do Quinlan sob uma licença proprietária. Ele usa menos memória e cria conjuntos de regras menores que o C4.5, sendo ao mesmo tempo mais preciso.

CART (Árvores de Classificação e Regressão) é muito semelhante a C4.5, mas difere porque suporta variáveis-alvo numéricas (regressão) e não calcula conjuntos de regras. CART constrói árvores binárias usando o recurso e o limite que produzem o maior ganho de informação em cada nó.

scikit-learn usa uma versão otimizada do algoritmo CART; entretanto, a implementação do scikit-learn não suporta variáveis categóricas por enquanto.