Decision tree

* classification method
* a collection of decision nodes, connected by branches
* predictor variables vs. target variable
* pure vs. diverse leaf nodes
* supervised learning – pre-classified target variables, training set
* training data – should be rich and varied
* discrete target attribute classes
* aim – leaf nodes to be as pure as possible
* algorithms for construction (measuring leaf node purity) – CART, C4.5

Metrics for measuring best split

* Gini impurity
* Information gain
* Variance reduction
* Measure of “goodness”

Classification and Regression Trees (CART)

* trees are strictly binary, containing exactly two branches for each decision node
* recursively partitions the records in the training data set into subsets of records with similar values for the target attribute
* grows the tree by conducting for each decision node, an exhaustive search of all available variables and all possible splitting values, selecting the optimal split
* classification error
* pruning nodes and branches – increases the generalizability of the results

C4.5

* recursively visits each decision node, selecting the optimal split, until no further splits are possible
* not restricted to binary splits
* produces a separate branch for each value of categorical attributes - may result in more “bushiness” than desired
* different method for measuring node homogeneity
* uses the concept of *information gain* or *entropy reduction* to select the optimal split.

C4.5 e един от алгоритмите, използвани за генерирани на т.нар. дърво на решенията. Разработен е от Ross Quinlan и е разширение на неговия по-ранен алгоритъм ID3.

Алгоритъмът рекурсивно обхожда всеки възел и избира оптималното разделяне, докато това е възможно. За разлика от други алгоритми като CART, C4.5 не е ограничен до бинарни разделяния и съответно може да създава дървета с по-голям брой разклонения. Когато разглежданата променлива е категорийна, C4.5 по-подразбиране създава разклонение за всяка нейна стойност. Това не винаги е удачно, тъй като някои от стойностите може да се срещат по-рядко или пък са естествено свързани с други такива.

C4.5 използва концепцията **за информационна печалба** или **намаляване на ентропията**, за да се избере оптималното разделяне. Първо се изчислява най-малкият брой битове средно за символ, необходими за предаването на поток от символи, представящи стойностите на дадена променлива X. За целта се използва формулата:

е вероятността на срещане на j-тата стойност на X. Резултатът се нарича ентропия на X. Формулата се базира на факта, че за събитие с вероятност p средното количество информационни битове за предаване на резултата е . Например резултатът от хвърляне на монета (честна) с вероятност 0.5, може да се предаде, като се използва (0 или 1 в зависимост от резултата). Когато възможните изходи са повече от два, се търси претеглената сума на отделните логаритми, като теглата са равни на вероятностите за появата на съответните изходи.

Концепцията за ентропия се използва по следния начин. Да предположим, че разделяне S, дели тренировъчните данни T на няколко подмножества. Средното информационно изискване може да се изчисли като претеглената сума от ентропиите на отделните подмножества:

Тук e пропорцията на записите, попаднали в подмножество . По този начин информационната печалба може да се дефинира като , тоест увеличението на информацията в резултат от разделянето на тренировъчните данни T спрямо S. Във всеки възел, при който става разклонение, C4.5 избира разделянето S, което води до най-голямата информационна печалба .