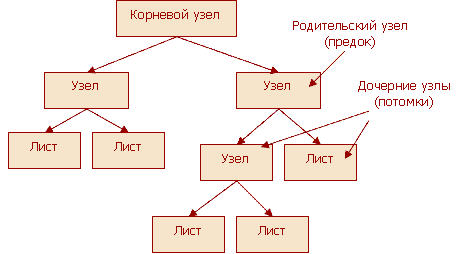
Дерево прийняття рішень - це засіб підтримки прийняття рішень при прогнозуванні, широко застосовується в статистиці і аналізі даних. Отже, дерево рішень, подібно до його «прототипу» з живої природи, складається з «гілок» і «листя». Гілки (ребра графа) зберігають в собі значення атрибутів, від яких залежить цільова функція; на листках ж записується значення цільової функції. Існують також і інші вузли - батьківські і нащадки - за якими відбувається розгалуження, і можна розрізнити випадки.



**Загальний алгоритм побудови дерева прийняття рішення**

Пропонуємо розглянути на прикладі загальний алгоритм побудови дерева, щоб потім перейти до його окремих випадків:

Спочатку необхідно вибрати атрибут Q (в нашому випадку, припустимо: рівень доходу> 500 $ в місяць) і помістити його в кореневий вузол.

Потім, з наших тестових прикладів (або набору даних) для кожного значення атрибута i (в нашому випадку їх два - «так» і «ні») вибираємо тільки ті, для яких Q = i.

Далі, рекурсивно будуємо дерево прийняття рішень.

Основна проблема, очевидно, криється в першому кроці - на якій підставі вибирається кожен наступний атрибут Q? На це питання існує кілька відповідей у ​​вигляді приватних алгоритмів прийняття рішень - головними з яких є алгоритми ID3, C4.5 і CART

**Основні алгоритми**

ID3. В основі цього алгоритму лежить поняття інформаційної ентропії - тобто, заходи невизначеності інформації (зворотного міру інформаційної корисності величини). Для того щоб визначити наступний атрибут, необхідно підрахувати ентропію всіх невикористаних ознак щодо тестових зразків і вибрати той, для якого ентропія мінімальна. Цей атрибут і буде вважатися найбільш доцільним ознакою класифікації.

C5. Цей алгоритм - удосконалення попереднього методу, що дозволяє, зокрема, «усікати» гілки дерева, якщо воно занадто сильно «розростається», а також працювати не тільки з атрибутами-категоріями, а й з числовими. В общем-то, сам алгоритм виконується за тим же принципом, що і його попередник; відмінність полягає в можливості розбиття області значень незалежної числової змінної на кілька інтервалів, кожен з яких буде атрибутом. Відповідно до цього вихідна безліч ділиться на підмножини. В кінцевому підсумку, якщо дерево виходить занадто великим, можлива зворотна угруповання - декількох вузлів в один лист. При цьому, оскільки перед побудовою дерева помилка класифікації вже врахована, вона не збільшується.

CART. Алгоритм розроблений з метою побудови так званих бінарних дерев рішень - тобто тих дерев, кожен вузол яких при розбитті «дає» тільки двох нащадків. Грубо кажучи, алгоритм діє шляхом поділу на кожному кроці безлічі прикладів рівно навпіл - по одній гілці йдуть ті приклади, в яких правило виконується (правий нащадок), за іншою - ті, в яких правило не виконується (лівий нащадок). Таким чином, в процесі «зростання» на кожному вузлі дерева алгоритм проводить перебір всіх атрибутів, і вибирає для наступного розбиття той, який максимізує значення показника, що обчислюється по математичній формулі і залежить від відносин числа прикладів в правому і лівому нащадку до загальної кількості прикладів.