Метод дерева рішень

Дерева рішень (decision trees) – метод для вирішення завдань класифікації та прогнозування.

Якщо *залежна (цільова) змінна* приймає *дискретні значення*, то за допомоги методу дерева рішень вирішується ***задача класифікації***.

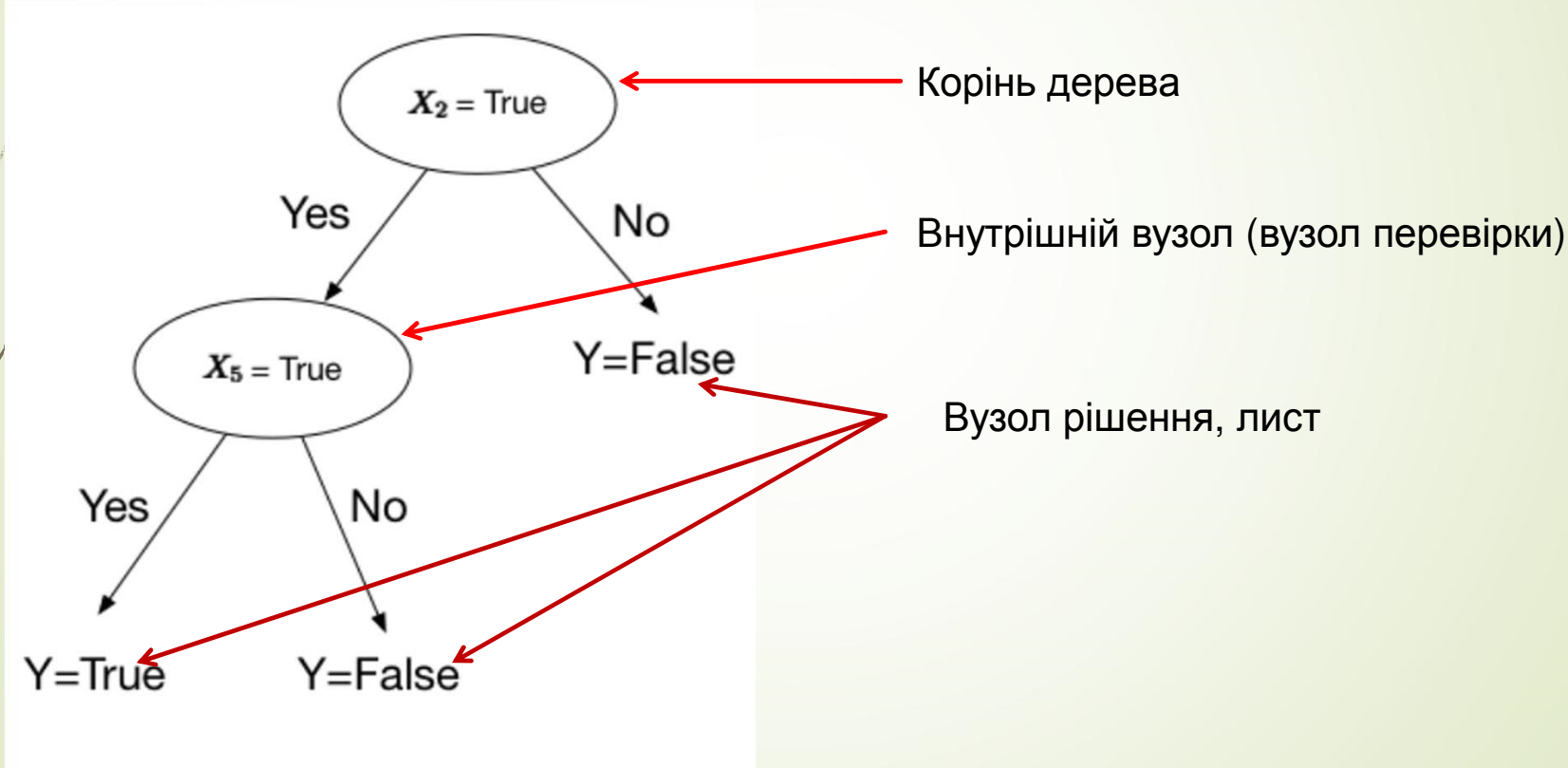
Якщо *залежна (цільова) змінна* приймає *неперервні значення*, то дерево рішень встановлює залежність цієї змінної від незалежних змінних, тобто вирішує ***задачу чисельного програмування***.

Підходи до кластеризації:

- ***алгоритми, засновані на розділенні даних (Partitioning algorithms)***, в тому числі ітеративні:
  - розділення об’єктів на k кластерів;
  - ітеративний перерозподіл об’єктів для покращення кластеризації
- ***ієрархічні алгоритми (Hierarchy algorithms)***:
  - агломерація
- ***методи, засновані на концентрації об’єктів (Density-based methods)***:
  - засновані на можливості з’єднання об’єктів
  - ігнорують шуми, знаходження кластерів довільної форми
- ***грід-методи (Grid-based methods)***:
  - квантування об’єктів у грід-структури
- ***модельні методи (Model-based)***:
  - використання моделі для знаходження кластерів, що найбільше відповідають даним

Дерево рішень:

Дерево рішень в найпростішому вигляді – це спосіб представлення правил у ієрархічній, послідовній структурі.



Переваги дерева рішень:

- інтуїтивність дерев рішень
- можливість здобувати правила з бази даних природньою мовою
- алгоритм конструювання дерева рішень не вимагає від користувача вибору вхідних атрибутів
- точність моделей порівняння з іншими методами побудови класифікаційних моделей
- наявність розроблених масштабуючих алгоритмів
- швидкий процес навчання
- наявність спеціальних алгоритмів обробки пропущених значень
- можливість працювати як з числовими, так і з категоріальними даними

Алгоритми конструювання дерев:

- *“побудова” або “створення” дерева (tree building)*
  - вибір критерію розщеплення (розбиття)
  - Вибір критерію зупинки навчання
- *“скорочення” дерева (tree pruning)*
  - вирішення питання відсікання деяких гілок дерева

Критерій розщеплення (розбиття)

1. В ході процесу створення дерева алгоритм повинен знайти такий критерій розщеплення (розбиття), щоб розбити множину на підмножини, які б асоціювалися з даним вузлом перевірки;
2. Кожен вузол перевірки повинен бути помічений певним атрибутом.

Правило вибору атрибута: він повинен розбивати вихідну множину даних так, щоб об'єкти підмножин, які отримано в результаті розбиття, були представниками одного класу або були максимально наближені до такого розбиття.

Критерії розщеплення: міра ентропії, індекс Gini.

Індекс Gini:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2,$$

де T - поточний вузол;

p_j - ймовірність класу j у вузлі T ;

n - кількість класів.

Велике дерево не означає, що воно є “придатним”

В процесі побудови дерева використовуються спеціальні процедури, що дозволяють створювати оптимальні дерева, так звані дерева “придатних розмірів”.

Дерево повинно використовувати інформацію, яка покращує якість моделі, та ігнорує ту інформацію, яка її не покращує.

Можливі стратегії:

1. нарощування дерева до певного розміру у відповідності з параметрами, заданими користувачем;
2. використання набору процедур: скорочення дерева шляхом відсікання гілок, використання правил зупинки навчання.

Зупинка побудови дерева

Зупинка – це такий момент у процесі побудови дерева, коли слід припинити подальші розгалуження.

Правила зупинки:

1. “рання зупинка” (prepruning) – визначає доцільність розбиття вузла;
2. обмеження глибини дерева;
3. завдання мінімальної кількості прикладів, що будуть міститись у кінцевих вузлах дерева.

Скорочення дерева або відсікання гілок

Точність розпізнавання – відношення об’єктів, правильно класифікованих в процесі навчання, до загальної кількості об’єктів набору даних, що брали участь у навчанні.

Помилка - відношення об’єктів, неправильно класифікованих в процесі навчання, до загальної кількості об’єктів набору даних, що брали участь у навчанні.

Відсікання гілок або заміну деяких гілок піддеревом слід здійснювати там, де ця процедура не призводить до зростання помилки. Процес відбувається знизу вгору, тобто процес є висхідним.

Алгоритми, що реалізують дерева рішень:

- CART (Classification and Regression Tree), 1974-1984

Розробники: Leo Breiman (Berkeley), Jerry Fridman (Stanford), Charles Stone (Berkeley), Richard Olsen (Stanford);

Алгоритм для побудови бінарного дерева рішень.

- C4.5
- CHAID
- CN2
- NewId
- Itrule
- Sprint