

ДЕРЕВА РІШЕНЬ (DECISION TREES)

Надія І. Недашківська n.nedashkivska@gmail.com

ПОНЯТТЯ ДЕРЕВА РІШЕНЬ

Надія І. Недашківська n.nedashkivska@gmail.com

Приклад

Задача: чи виграє “Динамо” свій наступний матч?

Параметри (незалежні змінні):

- ☐ чи вдома грається матч;
- ☐ чи вище знаходиться суперник в турнірній таблиці;
- ☐ яка температура повітря;
- ☐ чи йде дощ.

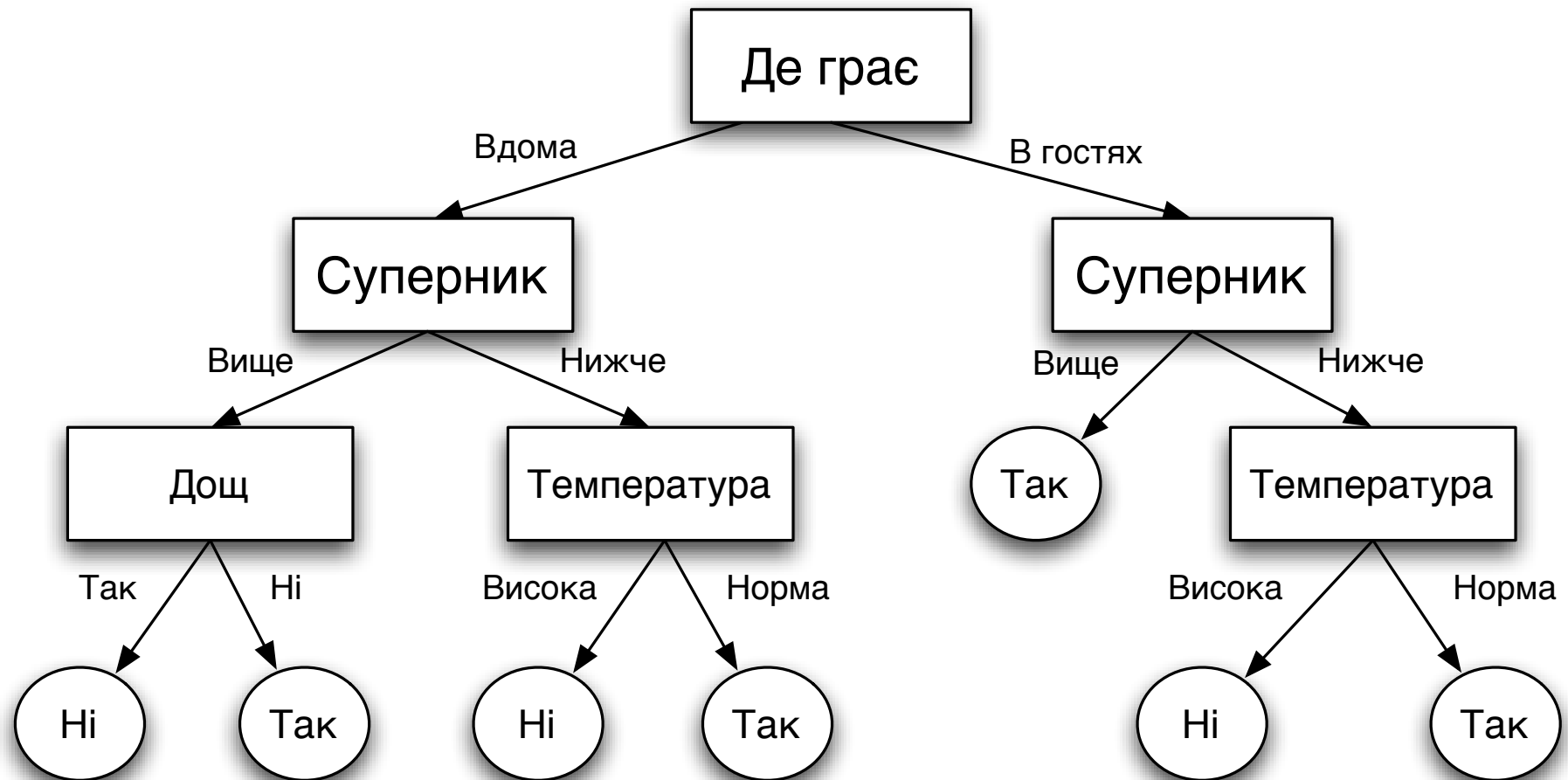
Відомі результати декількох матчів.

Спрогнозувати результат матчу при інших значеннях параметрів.

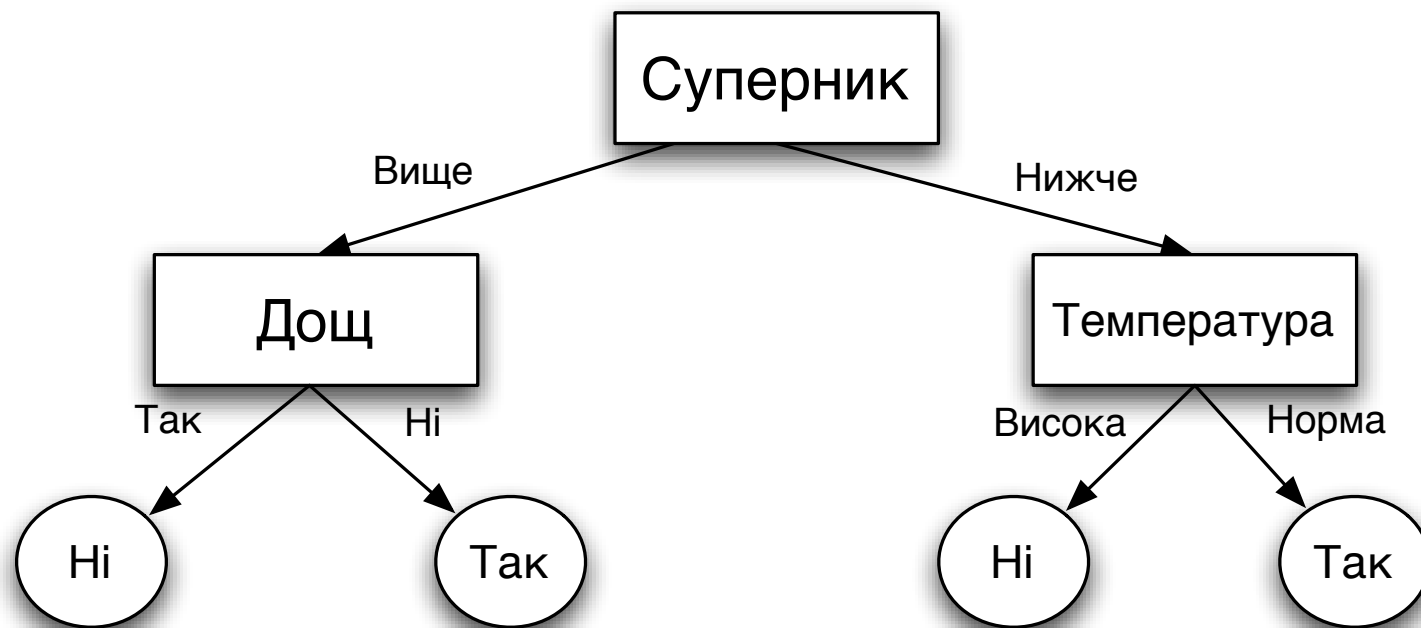
Приклад: результати попередніх ігор

Де грає	Суперник	Температура	Дощ	Перемога
Вдома	Вище	Висока	Так	Ні
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Ні
В гостях	Вище	Норма	Ні	Так

Приклад: дерево рішень



Приклад: дерево рішень: інша коренева вершина



Основні поняття

- ❑ **Дерево рішень (ДР)** - інтуїтивно зрозумілий метод класифікації шляхом задання серії уточнюючих питань.
- ❑ У бінарному ДР кожний вузол розбиває дані на підмножини за допомогою порогового значення однієї з ознак.
- ❑ У гарно спроектованому ДР кожне питання буде зменшувати кількість варіантів приблизно вдвоє, швидко звужує можливі варіанти навіть при великій кількості класів.
- ❑ Дерева рішень – це основа **випадкового лісу**, ансамблю моделей на основі дерев рішень.

Постановка задачі класифікації

Дано: Множина об'єктів $T = \{t_1, t_2, \dots, t_n\}$ навчальна вибірка
(тестові приклади)

$$t_i \rightarrow \{x_1, x_2, \dots, x_m, y\}$$

$X = \{x_1, x_2, \dots, x_m\}$ - незалежні змінні (атрибути)


$C_h = \{c_{h1}, c_{h2}, \dots, c_{hq_h}\}$ - значення, які приймає x_h

y - залежна змінна

$V = \{v_1, v_2, \dots, v_s\}$ - значення, які приймає y

Знайти: спрогнозувати значення y при нових значеннях незалежних змінних

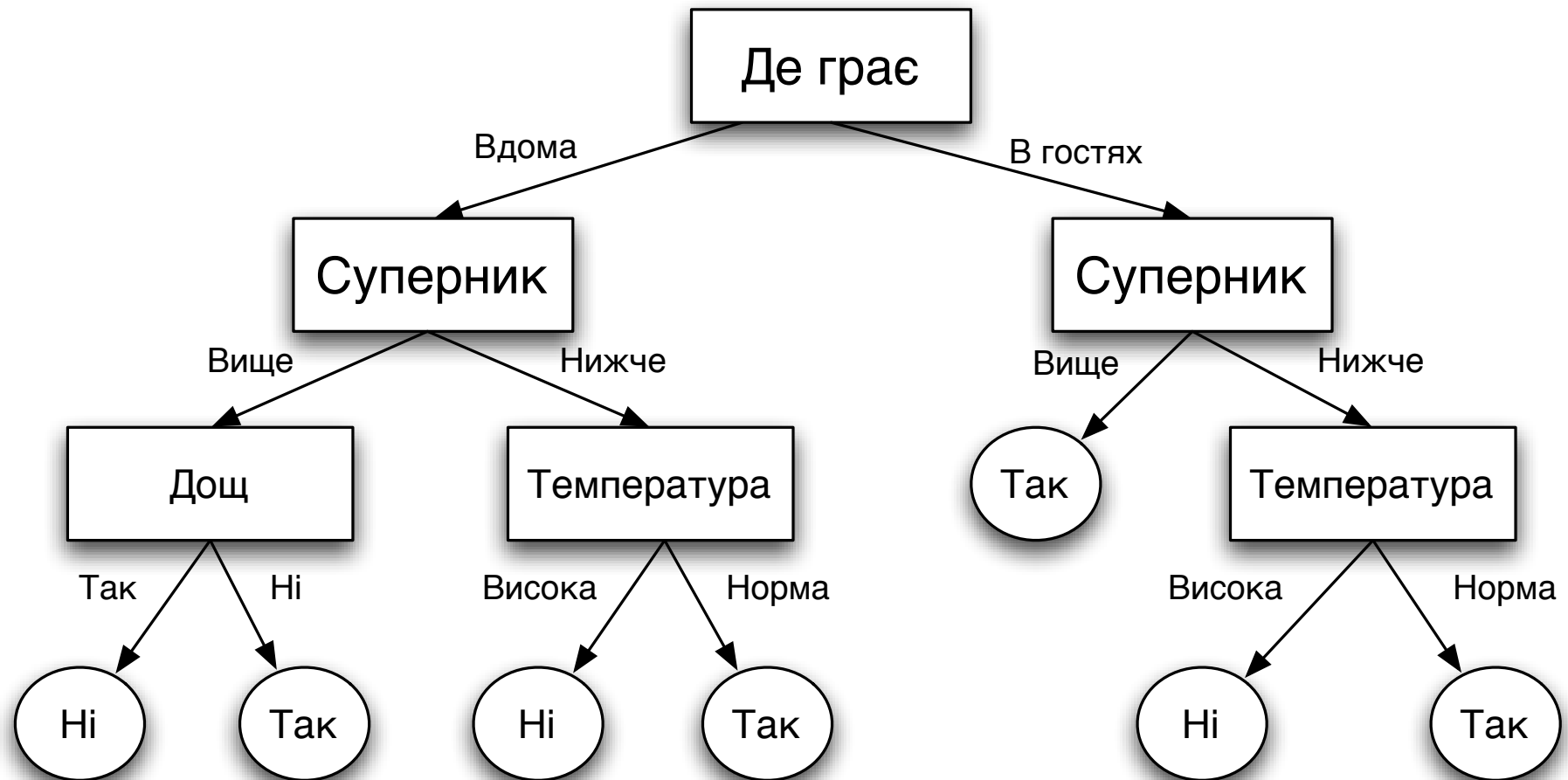
Поняття дерева рішень



Дерево рішень – зв'язний граф з множиною вершин, який не містить циклів і має окрему вершину, в яку не входить жодне ребро. Ця вершина називається коренем дерева.

Вершини двох видів: внутрішні та листи.

Приклад: дерево рішень




АЛГОРИТМ РОЗБИТТЯ

Надія І. Недашківська n.nedashkivska@gmail.com

Методи побудови дерев рішень

Алгоритм розбиття: загальний підхід



Ідея: Побудова дерева зверху-вниз, від кореня до листів

Рекурсивне розбиття навчальної вибірки на максимально більш “чисті” підмножини

Розбиття має бути значущим – класифікувати найбільшу кількість елементів навчальної вибірки

Алгоритм розбиття

1. Якщо множина T містить елементи, які відносяться до різних класів, тоді

1.1. x_h
$$T = \bigcup_{i=1}^{q_h} T_i$$

$$T_i \subseteq T : x_h = c_{hi}$$

Множини T_i більш “чисті” в порівнянні з T .

1.2. $T := T_i \quad \forall i = 1, \dots, q_h$

Методи побудови дерев рішень

Алгоритми розбиття: загальний підхід

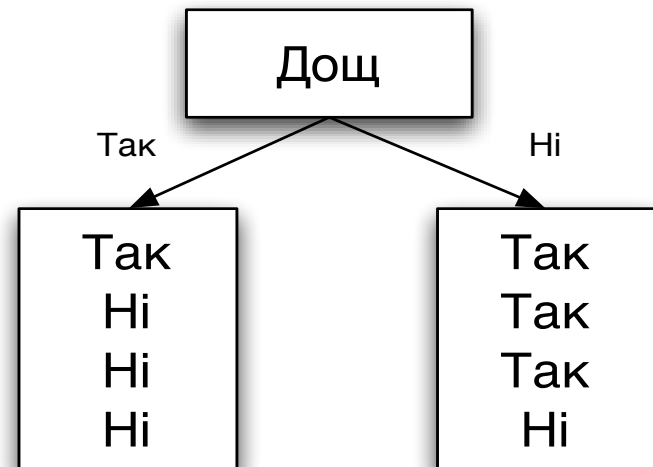
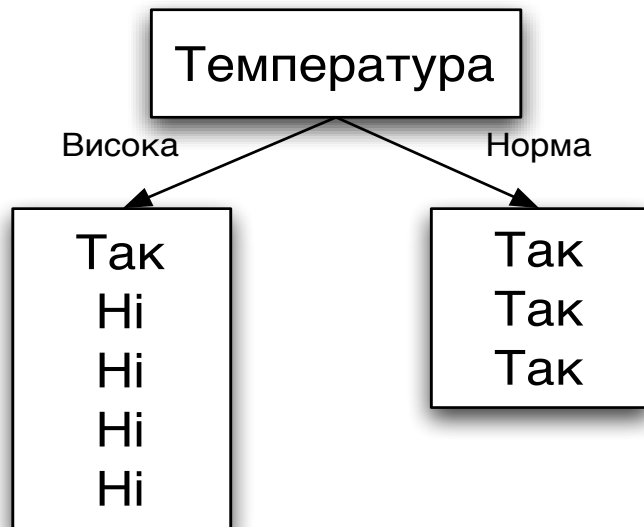
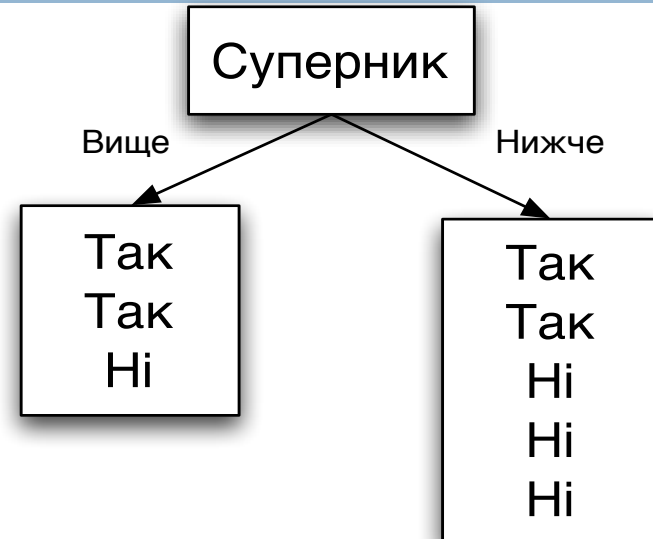
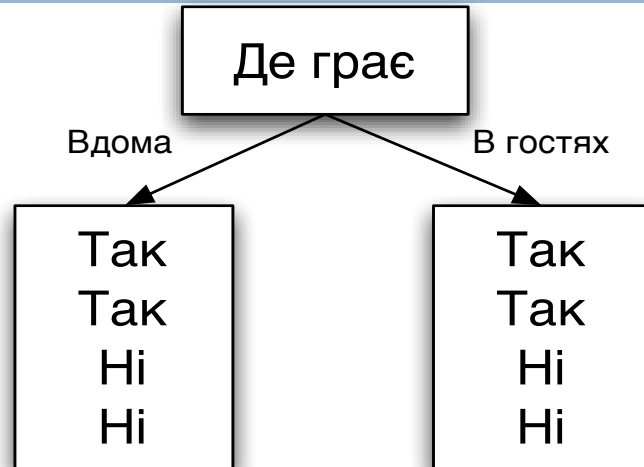
$x_h = \text{"Де грає"}$	Де грає	Суперник	Температура	Дощ	Перемога
T_1	Вдома	Вище	Висока	Так	Ні
	Вдома	Вище	Висока	Ні	Так
	Вдома	Нижче	Норма	Ні	Так
	Вдома	Нижче	Висока	Так	Ні
T_2	В гостях	Нижче	Норма	Так	Так
	В гостях	Нижче	Висока	Так	Ні
	В гостях	Нижче	Висока	Ні	Ні
	В гостях	Вище	Норма	Ні	Так
$c_{h1} = \text{"Вдома"}$					
$c_{h2} = \text{"В гостях"}$					

Вибір змінної розбиття: загальне правило

$$x_h^* : \max_{h=1,\dots,n} \begin{array}{l} \text{кількість елементів у } T_i, \\ \text{що належать одному класу} \\ i = 1, \dots, q_h \end{array}$$

Потрібний **критерій помилки**, який показує, наскільки якісно дана умова, тобто дана пара (ознака h та значення цієї ознаки q_h) розбиває вибірку T на підвибірки T_i .

Приклад: варіанти розбиття дерева



Властивості алгоритму розбиття



☐ “жадібність”

На кожному кроці робить локально оптимальний вибір, допускаючи, що результат буде глобально оптимальним

☐ ациклічність

Алгоритм не дозволяє повернутися назад і вибрати іншу змінну розбиття

Критерії вибору змінної розбиття



- ☐ Ентропійні (- приросту інформації – ID3,
- відношення приросту інформації – C4.5)
- ☐ Хі-квадрат
- ☐ Джині

Поняття ентропії

Означення: Нехай множина T складається з n об'єктів, k з яких мають властивість S .

Тоді ентропія множини T по відношенню до властивості S

$$H(T, S) = -\frac{k}{n} \log_2 \frac{k}{n} - \frac{n-k}{n} \log_2 \frac{n-k}{n}$$

$$H(T, S) = 1 \quad \text{коли} \quad k=n/2$$

$$H(T, S) = 0 \quad \text{коли} \quad k=n$$

Якщо S може приймати s різних значень, кожне з яких – в k_i випадках, тоді

$$H(T, S) = -\sum_{i=1}^s \frac{k_i}{n} \log_2 \frac{k_i}{n}$$

Приклад 1: розрахунок ентропії

Де грає	Суперник	Температура	Дощ	Перемога
Вдома	Вище	Висока	Так	Ні
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Ні
В гостях	Вище	Норма	Ні	Так

$$H(T, Victory) = -\frac{4}{8} \log_2 \frac{4}{8} - \frac{4}{8} \log_2 \frac{4}{8} = 1$$

Приклад 2: розрахунок ентропії

Де грає	Суперник	Температура	Дощ	Перемога Цільова змінна
Вдома	Вище	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Так

$$H(T, S) = - 4/7 \log_2 (4/7) - 3/7 \log_2 (3/7)$$

Критерій помилки розбиття для часткового випадку бінарного дерева

Потрібний **критерій помилки**, який показує, наскільки якісно дана умова, тобто дана пара (ознака h та значення порогу t) розбиває вибірку T на підвибірки T_L і T_R :

$$L(T, x_h, t) = \frac{|T_L|}{|T|} H(T_L) + \frac{|T_R|}{|T|} H(T_R)$$

Ентропійний критерій вибору змінної розбиття (алгоритм ID3) для загального, не бінарного дерева

Ідея: Максимізувати **приріст інформації** в результаті розбиття

Означення: Нехай множина T об'єктів, які характеризуються властивістю S , класифіковано за змінною x_h , яка приймає q_h значень.

Приростом інформації (Gain) в результаті розбиття називається

$$Gain(T, x_h) = H(T, S) - \sum_{i=1}^{q_h} \frac{|T_i|}{|T|} H(T_i, S)$$

$T_i \subseteq T : x_h = c_{hi}$ Ентропія розбиття

$$x_h^* = \arg \max_h Gain(T, x_h) \quad \text{- корінь дерева (піддерева)}$$

Приклад 1: розрахунок приросту інформації

$x_h = \text{“Де грає”}$

$$Gain(T, x_h) = H(T, S) - \sum_{i=1}^{q_h} \frac{|T_i|}{|T|} H(T_i, S)$$

$$H(T, S) = - \sum_{i=1}^s \frac{k_i}{n} \log_2 \frac{k_i}{n}$$

$$= 1 - \frac{4}{8} \cdot 1 - \frac{4}{8} \cdot 1 = 0$$

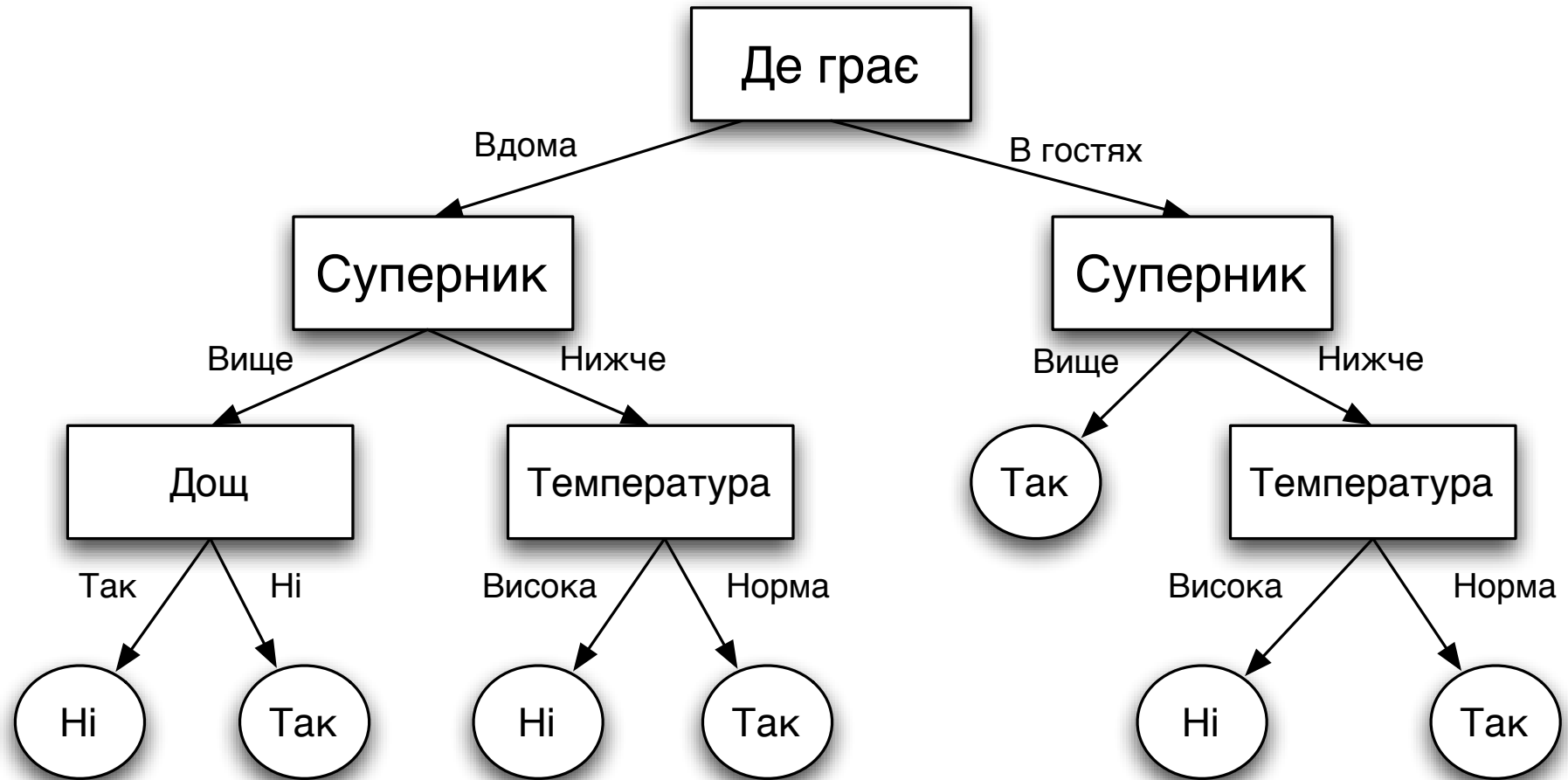
Де грає	Суперник	Температура	Дощ	Перемога
Вдома	Вище	Висока	Так	Ні
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Ні
В гостях	Вище	Норма	Ні	Так

$$Gain(T, x_h) =$$

$$= H(T, Victory) - \frac{4}{8} H(T_{\text{at_home}}, Victory) - \frac{4}{8} H(T_{\text{in_guest}}, Victory) =$$

$$- \frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}$$

Приклад 1: дерево рішень



Приклад 1: розрахунок приросту інформації

$x_h = \text{“Суперник”}$

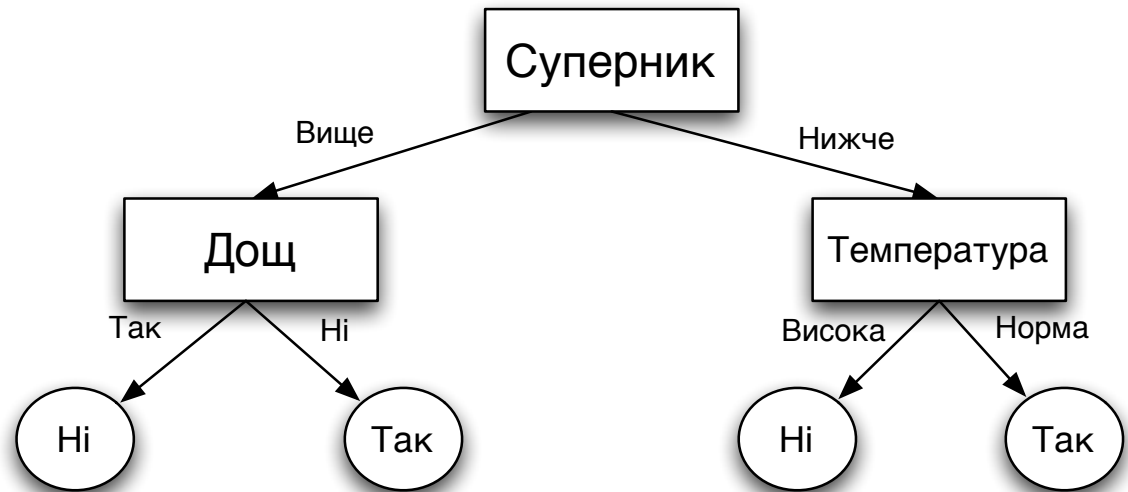
$$\text{Gain}(T, x_h) = 0.049$$

$x_h = \text{“Температура”}$

$$\text{Gain}(T, x_h) = 0.549$$

$x_h = \text{“Дощ”}$

$$\text{Gain}(T, x_h) = 0.189$$



Приклад 2: розрахунок ентропії

Де грає	Суперник	Температура	Дощ	Перемога Цільова змінна
Вдома	Вище	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Так

$$H(T, S) = - 4/7 \log_2 (4/7) - 3/7 \log_2 (3/7)$$

Приклад 2: розрахунок приросту інформації

$x_h = \text{“Суперник”}$

$$\text{Gain}(T, x_h) = H(T, S) - \sum_{i=1}^{q_h} \frac{|T_i|}{|T|} H(T_i, S)$$

Де грає	Суперник	Температура	Дощ	Перемога Цільова змінна
Вдома	Вище	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Так

$$\begin{aligned} \text{Gain}(T, x_h) = & H(T, S) - \\ & - \underbrace{2/7 H(T_1, S)} - \underbrace{5/7 H(T_2, S)} \end{aligned}$$

$$H(T_1, S) = -1/2 \log_2(1/2) - 1/2 \log_2(1/2)$$

$$H(T_2, S) = -3/5 \log_2(3/5) - 2/5 \log_2(2/5)$$

Переваги алгоритму ID3



- 1) простота та інтерпретовність класифікації.
Алгоритм може видати пояснення класифікації в термінах предметної області
- 2) трудомісткість алгоритма лінійна по довжині вибірки
- 3) не буває відмови від класифікації
- 4) простий в реалізації і легко піддається покращенням.
Можна вводити різні критерії розбиття, зупинки тощо.

Недоліки алгоритму ID3

- 1) жадібність
- 2) висока чутливість до складу вибірки
- 3) фрагментація
- 4) переускладнює структуру дерева і тому схильний до перенавчання. Класифікаційна здатність є невисокою.

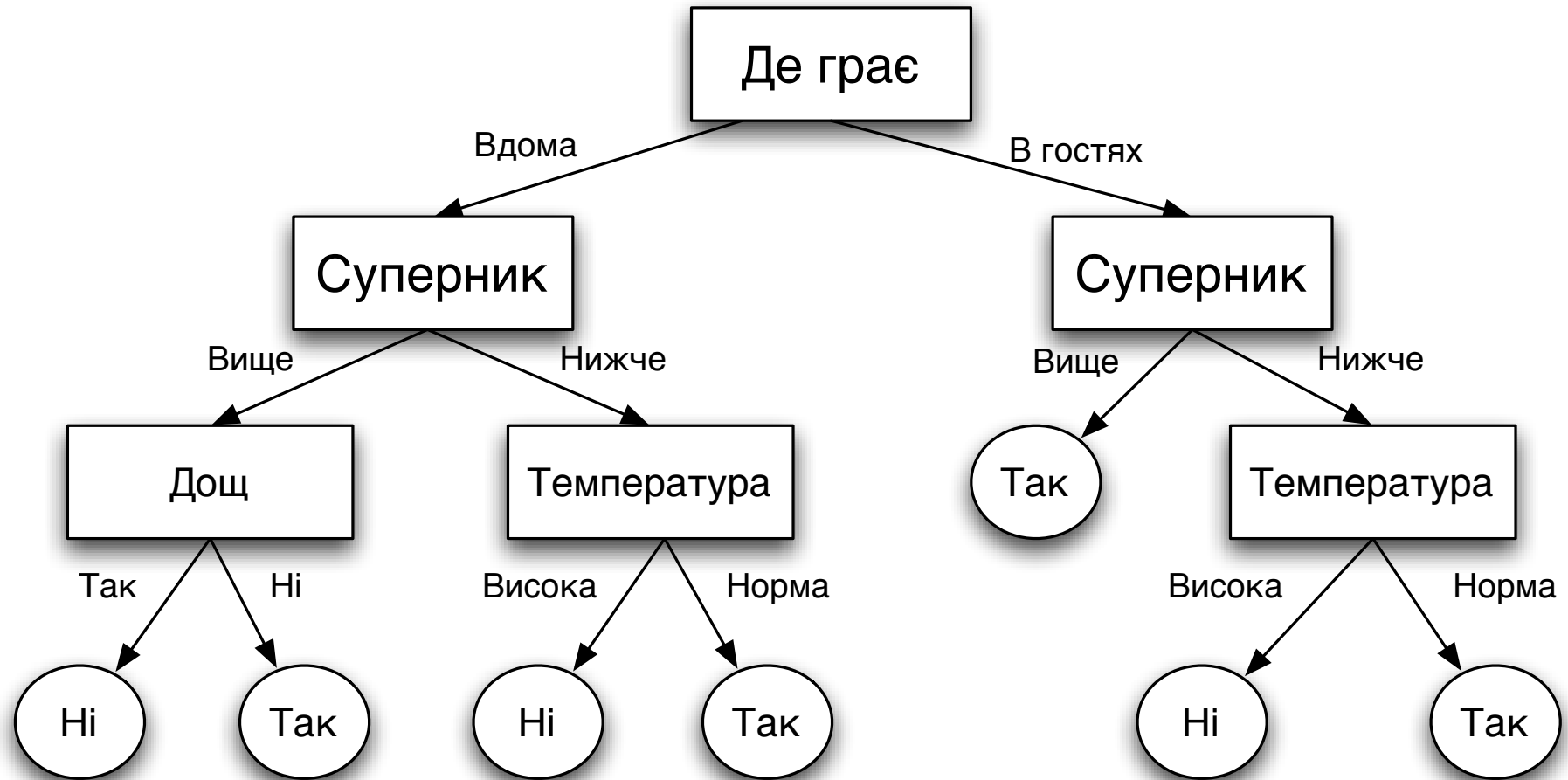
Способи подолання недоліків – застосування евристичних прийомів:

- редукція,
- погляд вперед (look ahead),
- побудова сукупності дерев – класифікаційного лісу (random forest)

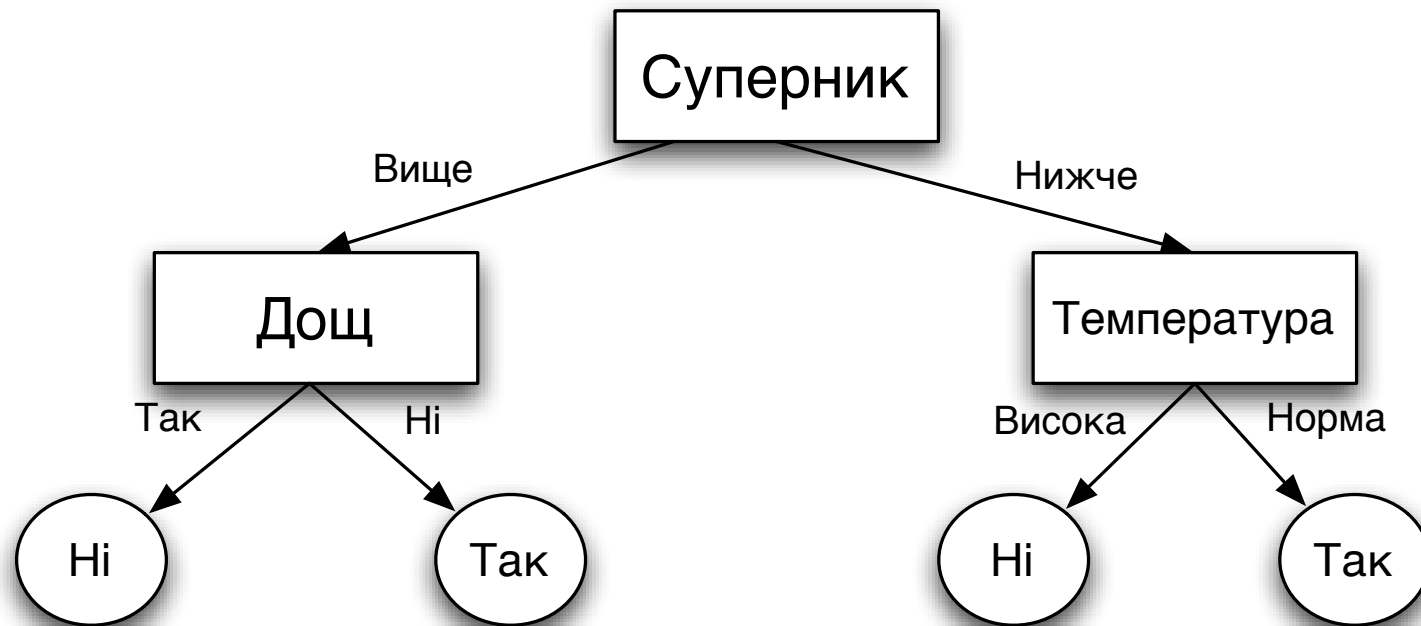
Приклад 1: результати попередніх ігор

Де грає	Суперник	Температура	Дощ	Перемога
Вдома	Вище	Висока	Так	Ні
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Ні
В гостях	Вище	Норма	Ні	Так

Приклад 1: дерево рішень



Приклад 1: дерево рішень: інша вершина



Ефект “перенавчання” (overfitting) моделі

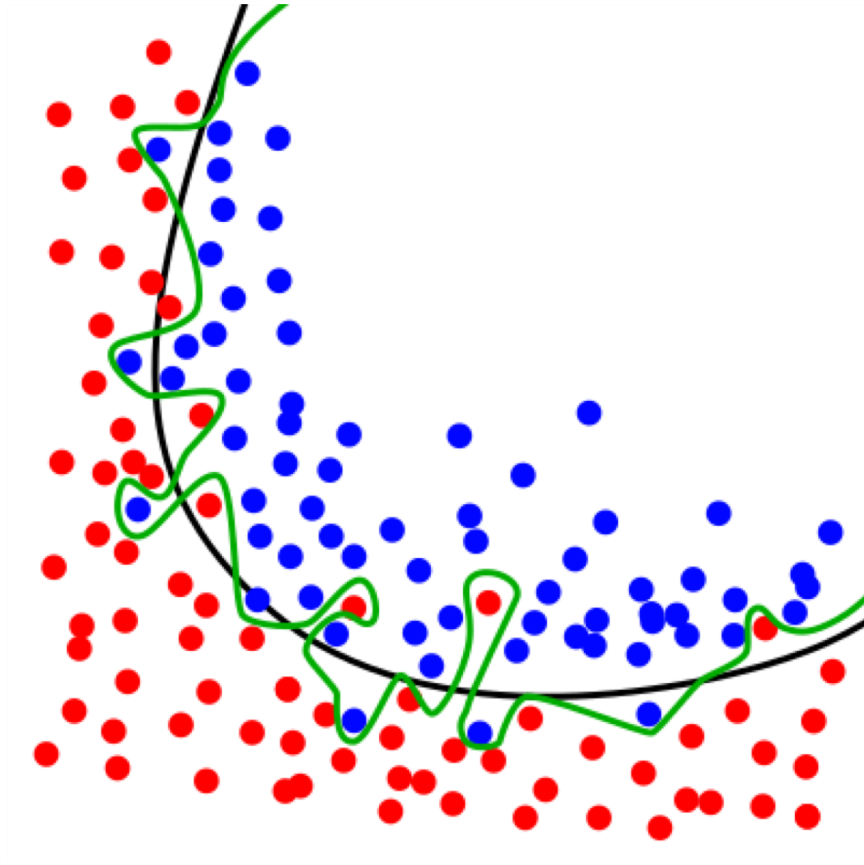


Перенавчання моделі – явище, коли побудована модель добре пояснює приклади з навчальної вибірки, але відносно погано працює на прикладах, які не брали участі в навчанні, наприклад на прикладах з тестової вибірки.

Це пов'язано з тим, що при побудові моделі («в процесі навчання») в навчальній вибірці виявляються деякі випадкові закономірності, які відсутні в генеральній сукупності.

Модель запам'ятовує величезну кількість всіх можливих прикладів замість того, щоб навчитися помічати особливості.

Ефект “перенавчання” (overfitting) моделі



Зелена крива – перенавчена модель.

Чорна крива має кращу узагальнюючу здатність, надасть кращий прогноз на нових даних.

Алгоритм ID3 схильний до перенавчання

x_h — ключ об'єкта

$$Gain(T, x_h) = H(T, S) - \sum_{i=1}^q \frac{|T_i|}{|T|} H(T_i, S)$$

$$T_i \subseteq T : x_h = c_{hi} \quad |T_i| = 1$$

$$Gain(T, x_h) = H(T, S) - \sum_{i=1}^q \frac{1}{q} H(T_{x_h=c_{hi}}, S) = H(T, S)$$

$$H(T, S) = - \sum_{i=1}^s \frac{k_i}{n} \log_2 \frac{k_i}{n}$$

Алгоритм C4.5 вибору змінної розбиття

Відношенням приросту інформації (**Gain_ratio**) в результаті розбиття називається:

$$Gain_ratio(T, x_h) = \frac{Gain(T, x_h)}{Split_info(T, x_h)}$$

- Приріст інформації в результаті розбиття

- Оцінка потенційної інформації в результаті розбиття

$$Split_info(T, x_h) = - \sum_{i=1}^q \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right)$$

$$x_h^* = \arg \max_h Gain_ratio(T, x_h) \quad \text{- корінь дерева (піддерева)}$$

Приклад 2: ілюстрація алгоритму C4.5

x_h = “Суперник”

$$Gain_ratio(T, x_h) = \frac{Gain(T, x_h)}{Split_info(T, x_h)}$$

Де грає	Суперник	Температура	Дощ	Перемога Цільова змінна
Вдома	Вище	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Так

$$Gain(T, x_h) = H(T, S) -$$

$$- \underbrace{2/7 H(T_1, S)} - \underbrace{5/7 H(T_2, S)}$$

$$Split_info(T, x_h) = - \sum_{i=1}^q \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right)$$

$$Split_info(T, x_h) = - 2/7 \log_2 (2/7) - 5/7 \log_2 (5/7)$$

Модифіковані алгоритми розбиття C4.5 (ID3)

Робота з неперервними ознаками – встановлення порогів для розбиття:

- 1) приклади сортуються у порядку зростання значень вибраної ознаки $\{c_1, c_2, \dots, c_q\}$
- 2) розглядаються значення порогів $\{c_1, c_2, \dots, c_{q-1}\}$
- 3) вибирається оптимальне значення порогу – що забезпечує максимальне значення Gain_ratio (у випадку C4.5) або Gain (у випадку ID3)

Приклад оцінювання кредитного ризику

№ клієнта	Збереження	Інші активи (нерухомість, автомобіль тощо)	Річний дохід	Кредитний ризик
1	Середні	Високі	75	Низький
2	Низькі	Низькі	50	Високий
3	Високі	Середні	25	Високий
4	Середні	Середні	50	Низький
5	Низькі	Середні	100	Низький
6	Високі	Високі	25	Низький
7	Низькі	Низькі	25	Високий
8	Середні	Середні	75	Низький

Приклад оцінювання кредитного ризику за модифікованим алгоритмом ID3

	№ розбиття	Дочірні вузли	Приріст інформації
“Активи” – корінь дерева	1	Збереження=Низькі	0,360
		Збереження=Середні	
		Збереження=Високі	
	2	Активи=Низькі	0,548
		Активи=Середні	
		Активи=Високі	
	3	Дохід ≤25	0,159
		Дохід >25	
	4	Дохід ≤50	0,347
		Дохід >50	
	5	Дохід ≤75	0,092
		Дохід >75	

Для ознаки «Дохід» потрібно вибрати оптимальне значення порогу рівне 50.

Приклад оцінювання кредитного ризику

Результуюче дерево

