

# Intelligens orvosi műszerek

## VIMIA023

A mintapéldákból tanuló számítógépes program  
(egyik lehetőség):  
**döntési fák**

2018 ősz

<http://www.mit.bme.hu/oktatas/targyak/vimia023>

dr. Pataki Béla

[pataki@mit.bme.hu](mailto:pataki@mit.bme.hu)

(463-)2679

Összefoglalva: hogyan lehet egy számítógépes döntést megvalósítani?

(1) Analitikus tervezés:

Begyűjteni az analitikus (fizikai, kémiai, biológiai stb. összefüggésekből felépített) modelleket az adott problémára. Megtervezni analitikusan a konkrét mechanizmust, és azt algoritmusként implementálni.

(2) Megtervezni a tanulás mechanizmusát, és azt algoritmusként implementálni.

Majd alkalmazásával *megtanulni a minták alapján* a döntés tényleges mechanizmusát, és azt algoritmusként implementálni.

A következőkben olyan módszereket vizsgálunk, amikor minták, **mintapéldák alapján akarjuk kialakítani a döntési rendszerünket.**

Nagyon gyakran a számítógép tanulása azt jelenti, hogy van egy csomó mintapéldánk, és ez hordozza az információt. Nincsenek előre kialakított szabályaink a feladatra, csak a mintáink. Persze ilyenkor is valamilyen struktúrában igyekszünk felhasználni a minták hordozta tudást.

Tehát a módszerünk – hogyan használjuk fel a mintákat – bizonyos mértékben problémafüggetlen.

A hétköznapi életben nagyon gyakori a minta alapján való tanulás, a klasszikus számítógépes megoldásokban a humán fejlesztő szabályokat alkotott, és ezeket programozta be.

Sokféle tanítható eszközt kitaláltak:

- neurális hálók
- Bayes-hálók
- kernelgépek
- döntési fák
- legközelebbi szomszéd osztályozók
- stb. stb.

A következőkben először a **döntési fákkal** foglalkozunk – működésük az emberi gondolkodás számára jóval jobban áttekinthető, mint pl. a mesterséges neurális hálóké. (white box  $\Leftrightarrow$  black box)

Vannak regressziós döntési fák is, de mi az osztályozással foglalkozunk.

**Példaprobléma** (Russel-Norvig: Mesterséges intelligencia c. könyv példája – idétlen példa, de jól megérthető ennek alapján a módszer):  
döntés, hogy egy adott étteremben várjunk-e asztalra.

**Cél:** előállítani a döntés logikai definícióját.

A problémát tulajdonságokkal vagy **attribútumokkal** jellemezzük.

1. **Alternatíva**: van-e a környéken más megfelelő étterem. (I/N)
2. **Bár**: van-e az étteremben kényelmes bár, ahol várhatunk. (I/N)
3. **Pén/Szom**: igaz pénteken- és szombatonként. (I/N)
4. **Éhes**: éhesek vagyunk-e. (I/N)
5. **Kuncsaft**: hányan vannak már benn (*Senki, Néhány és Tele*).
6. **Ár**: mennyire drága (*Olcsó, Közepes, Drága*)
7. **Eső**: esik-e kint (I/N)
8. **Foglalás**: foglalni kell-e előzetesen az asztalt (I/N)
9. **Típus**: az étterem típusa (*Francia, Olasz, Thai v. Burger*)
10. **BecsültVár**: a pincér által becsült várakozási idő  
(0-10 perc, 10-30, 30-60, >60 perc)

A **döntés eredményét** a döntési függvény adja meg – döntési faként ábrázoljuk.

# Az eddigi kiruccanásaink tapasztalata - tanítóminták

| Pl.      | Attribútumok  |             |             |             |               |              |             |             |                |               | Cél             |
|----------|---------------|-------------|-------------|-------------|---------------|--------------|-------------|-------------|----------------|---------------|-----------------|
|          | <i>Altern</i> | <i>Bár</i>  | <i>Pént</i> | <i>Éhes</i> | <i>Kuncs</i>  | <i>Ár</i>    | <i>Eső</i>  | <i>Fogl</i> | <i>Típus</i>   | <i>Becs</i>   | <i>VárniFog</i> |
| $X_1$    | <i>Igen</i>   | <i>Nem</i>  | <i>Nem</i>  | <i>Igen</i> | <i>Néhány</i> | <i>Drága</i> | <i>Nem</i>  | <i>Igen</i> | <i>Francia</i> | <i>0–10</i>   | <i>Igen</i>     |
| $X_2$    | <i>Igen</i>   | <i>Nem</i>  | <i>Nem</i>  | <i>Igen</i> | <i>Tele</i>   | <i>Olcsó</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Thai</i>    | <i>30–60</i>  | <i>Nem</i>      |
| $X_3$    | <i>Nem</i>    | <i>Igen</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Néhány</i> | <i>Olcsó</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Burger</i>  | <i>0–10</i>   | <i>Igen</i>     |
| $X_4$    | <i>Igen</i>   | <i>Nem</i>  | <i>Igen</i> | <i>Igen</i> | <i>Tele</i>   | <i>Olcsó</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Thai</i>    | <i>10–30</i>  | <i>Igen</i>     |
| $X_5$    | <i>Igen</i>   | <i>Nem</i>  | <i>Igen</i> | <i>Nem</i>  | <i>Tele</i>   | <i>Drága</i> | <i>Nem</i>  | <i>Igen</i> | <i>Francia</i> | <i>&gt;60</i> | <i>Nem</i>      |
| $X_6$    | <i>Nem</i>    | <i>Igen</i> | <i>Nem</i>  | <i>Igen</i> | <i>Néhány</i> | <i>Közep</i> | <i>Igen</i> | <i>Igen</i> | <i>Olasz</i>   | <i>0–10</i>   | <i>Igen</i>     |
| $X_7$    | <i>Nem</i>    | <i>Igen</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Senki</i>  | <i>Olcsó</i> | <i>Igen</i> | <i>Nem</i>  | <i>Burger</i>  | <i>0–10</i>   | <i>Nem</i>      |
| $X_8$    | <i>Nem</i>    | <i>Nem</i>  | <i>Nem</i>  | <i>Igen</i> | <i>Néhány</i> | <i>Közep</i> | <i>Igen</i> | <i>Igen</i> | <i>Thai</i>    | <i>0–10</i>   | <i>Igen</i>     |
| $X_9$    | <i>Nem</i>    | <i>Igen</i> | <i>Igen</i> | <i>Nem</i>  | <i>Tele</i>   | <i>Olcsó</i> | <i>Igen</i> | <i>Nem</i>  | <i>Burger</i>  | <i>&gt;60</i> | <i>Nem</i>      |
| $X_{10}$ | <i>Igen</i>   | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Tele</i>   | <i>Drága</i> | <i>Nem</i>  | <i>Igen</i> | <i>Olasz</i>   | <i>10–30</i>  | <i>Nem</i>      |
| $X_{11}$ | <i>Nem</i>    | <i>Nem</i>  | <i>Nem</i>  | <i>Nem</i>  | <i>Senki</i>  | <i>Olcsó</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Thai</i>    | <i>0–10</i>   | <i>Nem</i>      |
| $X_{12}$ | <i>Igen</i>   | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Tele</i>   | <i>Olcsó</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Burger</i>  | <i>30–60</i>  | <i>Igen</i>     |

**Mi a viselkedésünk jellemző mintája? Van-e egyáltalán egy konzisztens - minden tanítópéldának megfelelő – viselkedési mintánk? (9216 lehetséges példa van, 12 alapján tanulunk)**

# Döntési fák tanulása

**döntési fa** = egy logikai függvény

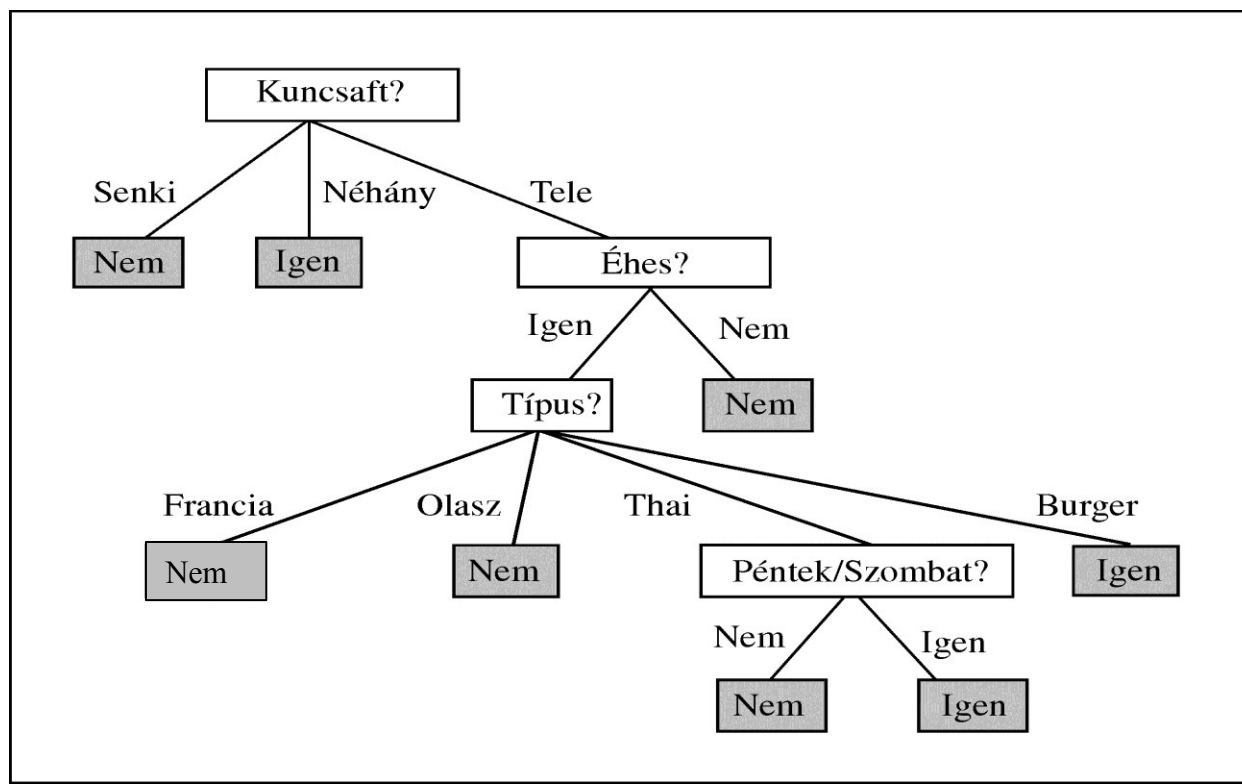
**bemenete:** egy tulajdonsághalmazzal leírt objektum vagy szituáció

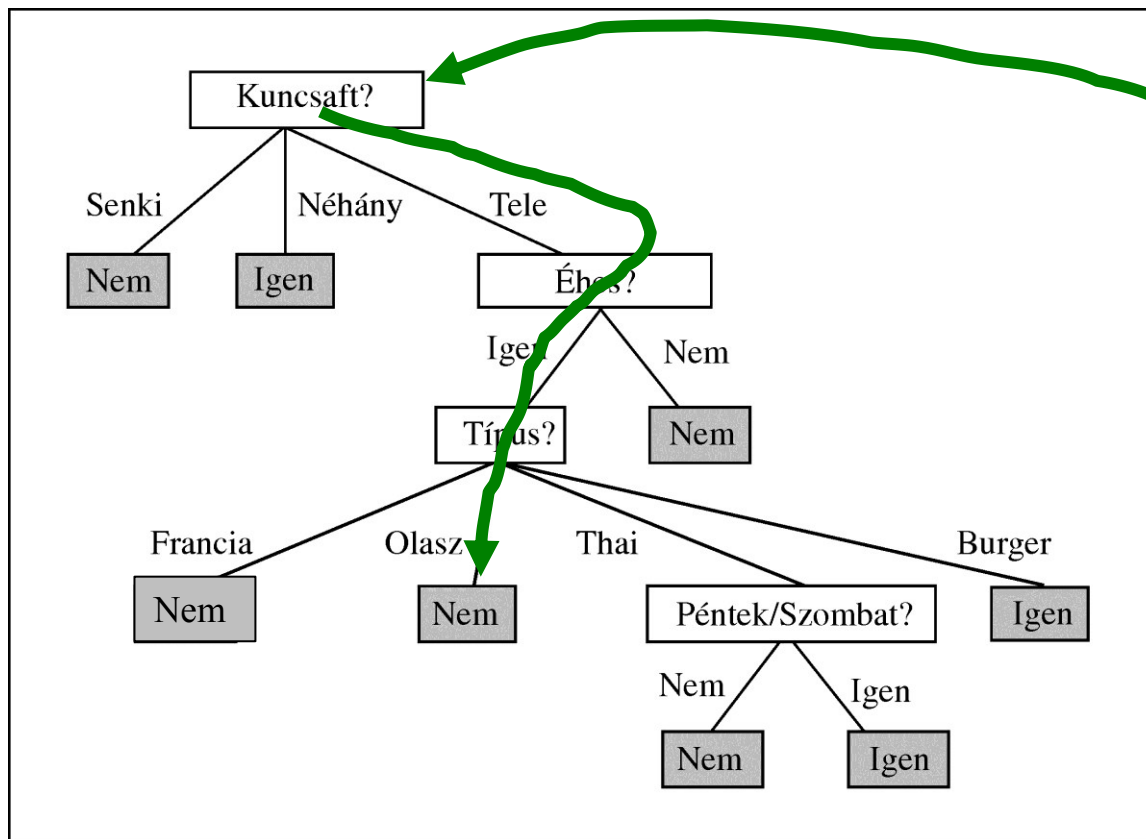
**kimenete:** egy igen/nem „döntés/cselekvés”

**belső csomópont** = valamelyik tulajdonság tesztje

**él** = a teszt lehetséges értéke

**levél** = logikai érték, amelyet akkor kell kiadni, ha ezt a levelet elértük.





Egy konkrét eset:

Kuncsaft = Tele, és  
Éhes, és  
Étterem = Olasz

Tehát a döntési fa válasza erre az esetre: nem fog várni az illető arra, felszabaduljon asztal. (Ez volt az egyik tanítómintánk – egy a tizenkettőből.)



## **Döntési fák kialakítása példák alapján**

**példa:** (attribútumok értékei, célváltozó értéke)

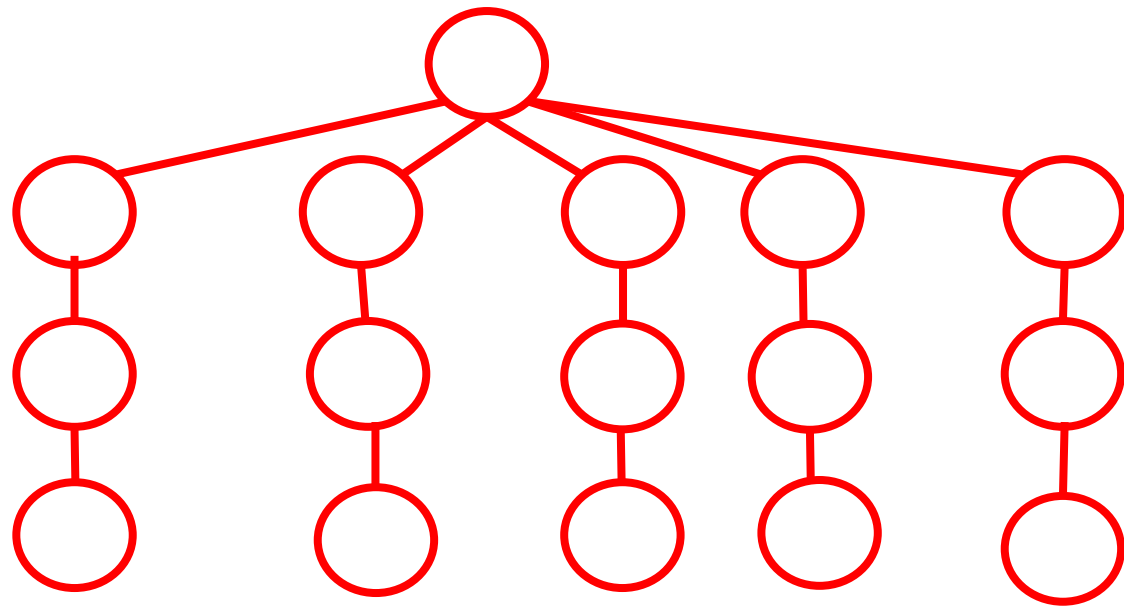
**példa besorolása:** a célváltozó értéke

**pozitív/negatív példa:** a célváltozó értéke igaz / hamis

**tanító halmaz:** a teljes példahalmaz

**triviális** fa, könnyen előállítható

- mindegyik példához egy önálló bejárési út tartozik
- a levél a példa besorolását adja
- a fa egyszerűen memorizálja a példát, nem alakít ki jellegzetes mintákat
- a példákból **nem általánosít** (nem tud ismeretlen példákra választ adni)

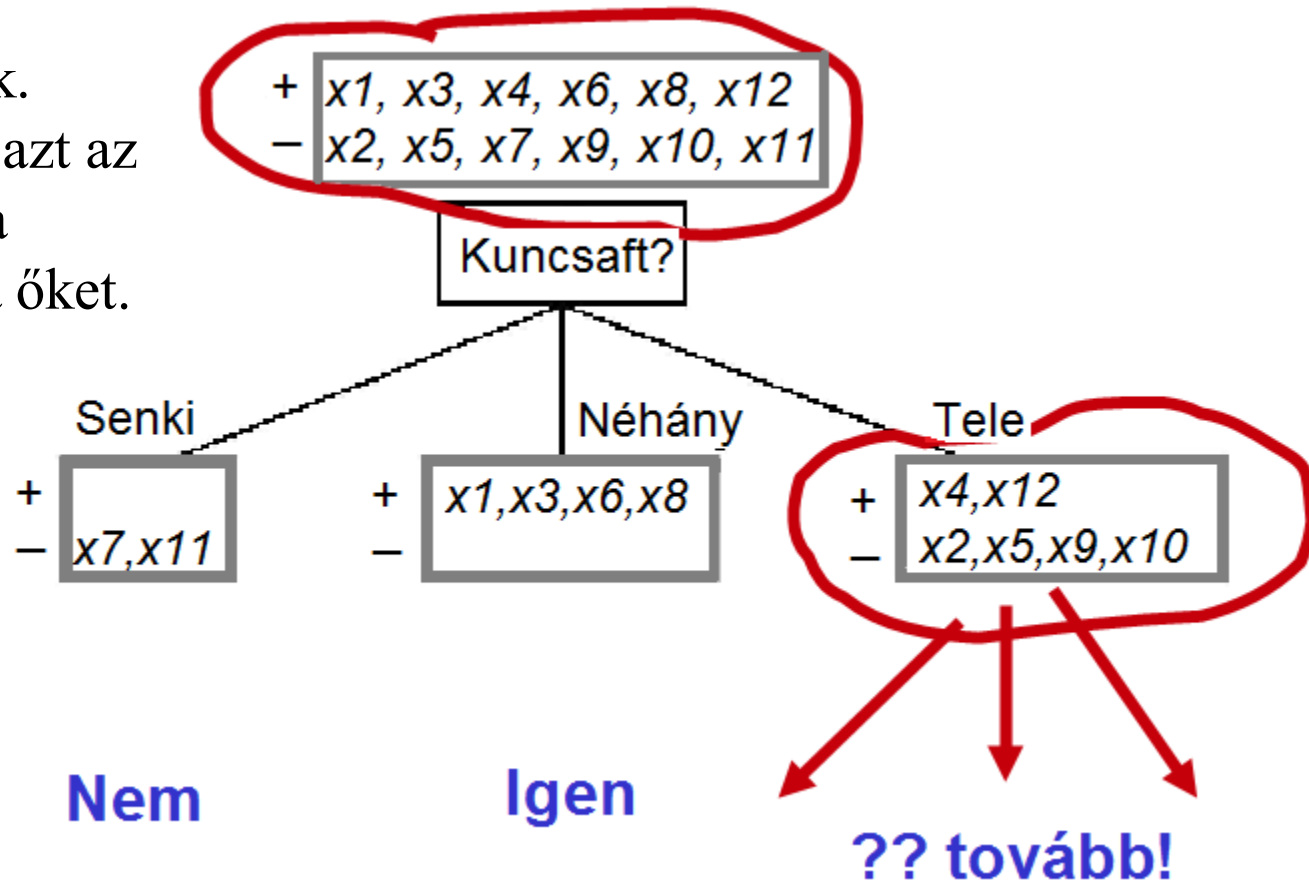


**„igazi”** fa - a jellegzetes minták kinyerése,  
a módszer képes nagyszámú esetet **tömör** formában leírni, és így  
az egyelőre ismeretlen példákra is **általánosítani**

# A döntési fa építése – általános jelenségek

Van néhány pozitív és néhány negatív példánk.

**Heurisztika:** válasszuk azt az attribútumot, amelyik a legjobban szétválasztja őket.



Ha az **összes** megmaradt eset **pozitív**, (vagy az **összes negatív**), akkor készen vagyunk azzal az ággal: a válasz **Igen** vagy **Nem**.

# Döntési fák kialakítása példák alapján

*A legkisebb döntési fa megtalálása - általánosságban nem megoldható!*

Heurisztika: **mohóság** - egy *meglehetősen egyszerű* (jó) fa is jó lesz, nem fogunk a tökéletes fa megtalálásával nagyon sok időt tölteni!  
(Sokszor kompromisszumot kell kössünk az eredmény jósága, és a jó megoldás megtalálásra fordított energia közt.)

Az alapötlet: először a „legfontosabb” attribútumot teszteljük.

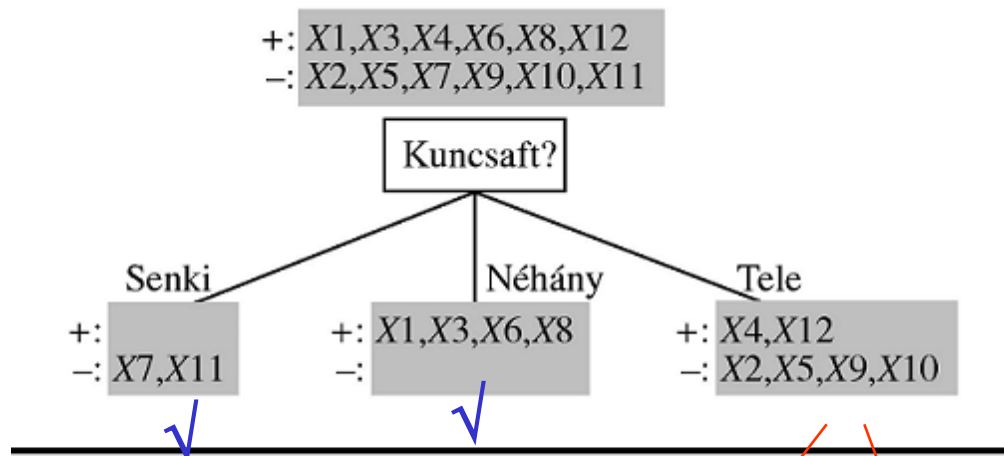
„**legfontosabb**” = a **legnagyobb eltérést** okozza példák besorolásában

Elvárás: **kisszámú teszt** alapján korrekt besoroláshoz jutunk:

a **bejárési utak rövidek** lesznek, és így az **egész fa kicsi** (tömör) lesz.

A *Kuncsaft* attribútum **nem** tökéletes, de elég jó.

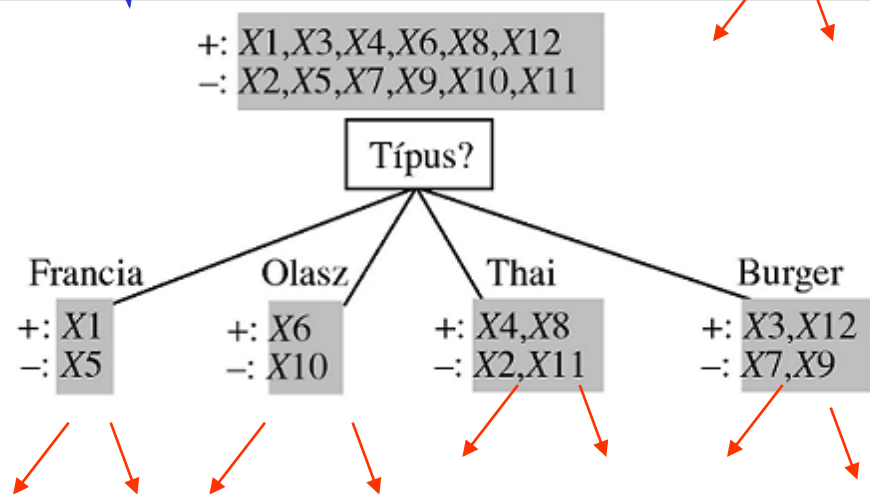
Ha a Kuncsaft tesztelése után döntünk, várhatóan legfeljebb 2-t hibázunk a 12 példából.



A *Típus* attribútum ehhez képest? (*vacak...*)

Ha a Típus tesztelése után döntünk, várhatóan 6-ot hibázunk a 12 példából.

Éppúgy, mint a teszt előtt!



Egy **nagymértékben haszontalan attribútum**, mint pl. a *Típus* olyan példahalmazokat hoz létre, amelyek nagyjából ugyanolyan arányban tartalmaznak pozitív és negatív példákat, mint az eredeti halmaz.

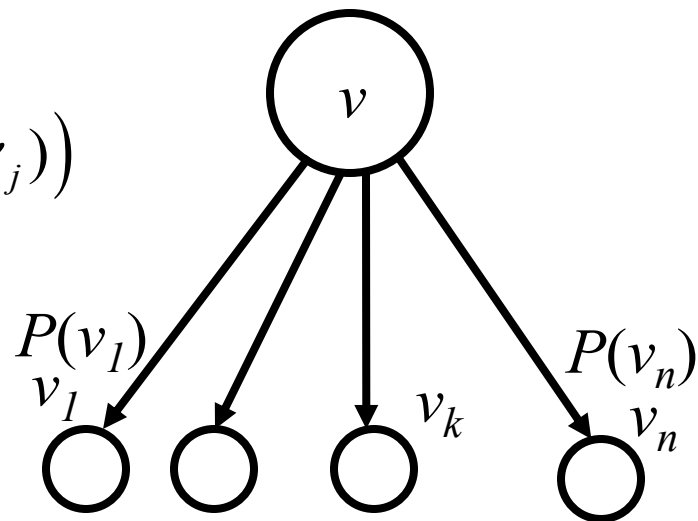
## „Elég jó?" „Nagymértékben haszontalan?"

A mérték legyen **maximális**: a **tökéletes** attribútumra,  
**minimális**: olyan attribútumra, aminek  
egyáltalán **nincs értéke** számunkra.

Egy megfelelő mérték: **az attribútum által adott információ  
várható értéke, információ átlagos tartalma, entrópia, ...**

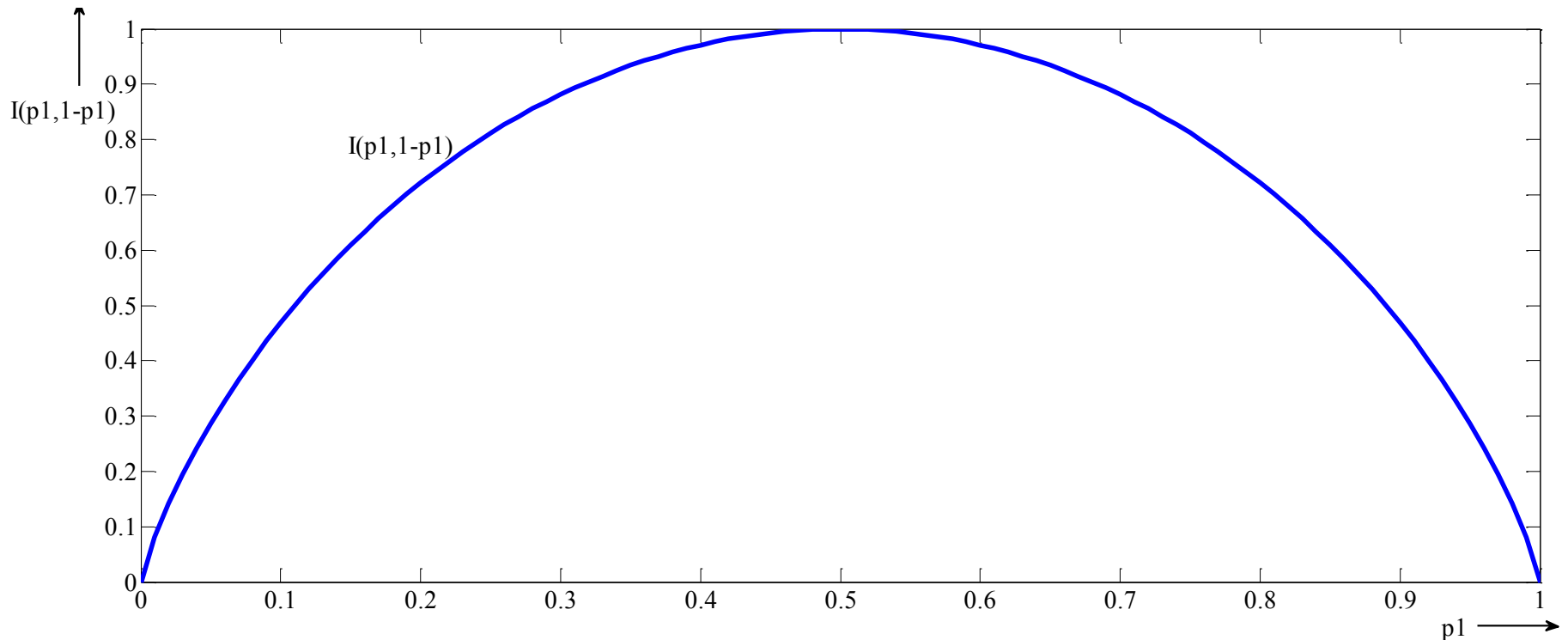
Ha a lehetséges  $v_k$  válaszok valószínűsége  $P(v_k)$ , akkor az adott  
kiinduló halmaz esetén a jó döntéshez szükséges  
információszükséglet:

$$I(P(v_1), P(v_2), \dots, P(v_k)) = \sum_{j=1}^k -P(v_j) \cdot \log_2(P(v_j))$$



pl. Fogadunk 1000 Ft-ba egy szabályos pénzérme dobása esetén, hogy „fej” jön ki. Mennyit ér nekünk az infó, ha valaki elárulja, hogy melyik fog kijönni? Mennyi infóra van szükségünk?

Mennyit ér az infó, ha tudjuk, hogy az érme aszimmetrikus, és 80%-ban „fej” jön ki? Mennyit ér nekünk az infó, ha az érme mindkét oldalán „fej” van, tehát 100%-ban ez jön ki?



## A döntési fa információtartalma - a tanítóhalmazban található pozitív és negatív példák aránya alapján

A tanítóhalmaz  $p$  pozitív és  $n$  negatív példát tartalmaz. A két válasz:  $v_1$  (pozitív),  $v_2$  (negatív), és a valószínűségük becslése:

$$P(v_1) \cong \frac{p}{p+n}, \quad P(v_2) \cong \frac{n}{p+n}$$

Ekkor a fa információtartalmának becslése:

$$I(P(v_1), P(v_2)) \cong I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \cdot \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \cdot \log_2\left(\frac{n}{p+n}\right)$$

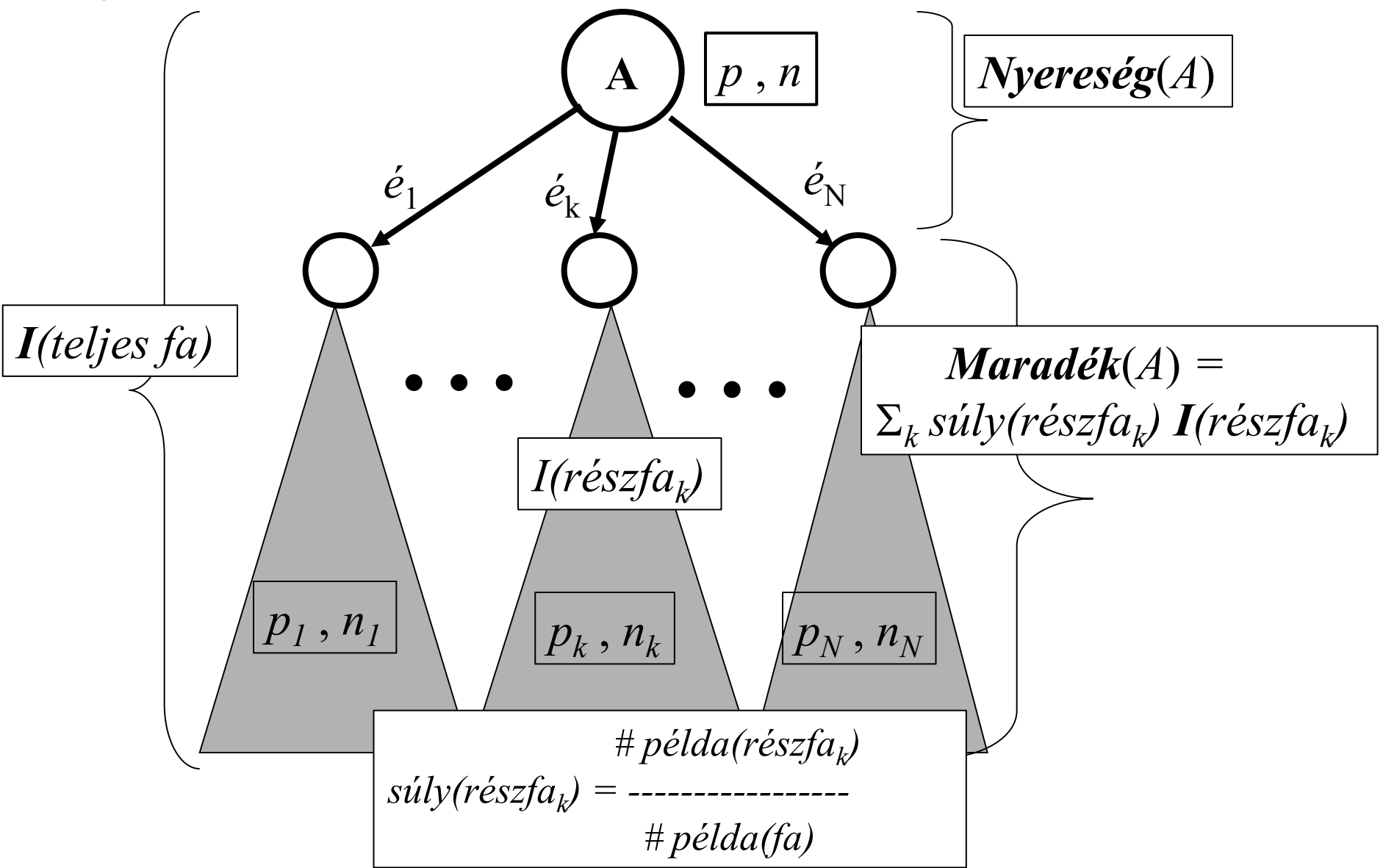
Hogyan válasszuk ki, hogy melyik attribútumot teszteljük?

Mennyi új információt ad nekünk egy attribútum tesztelése?

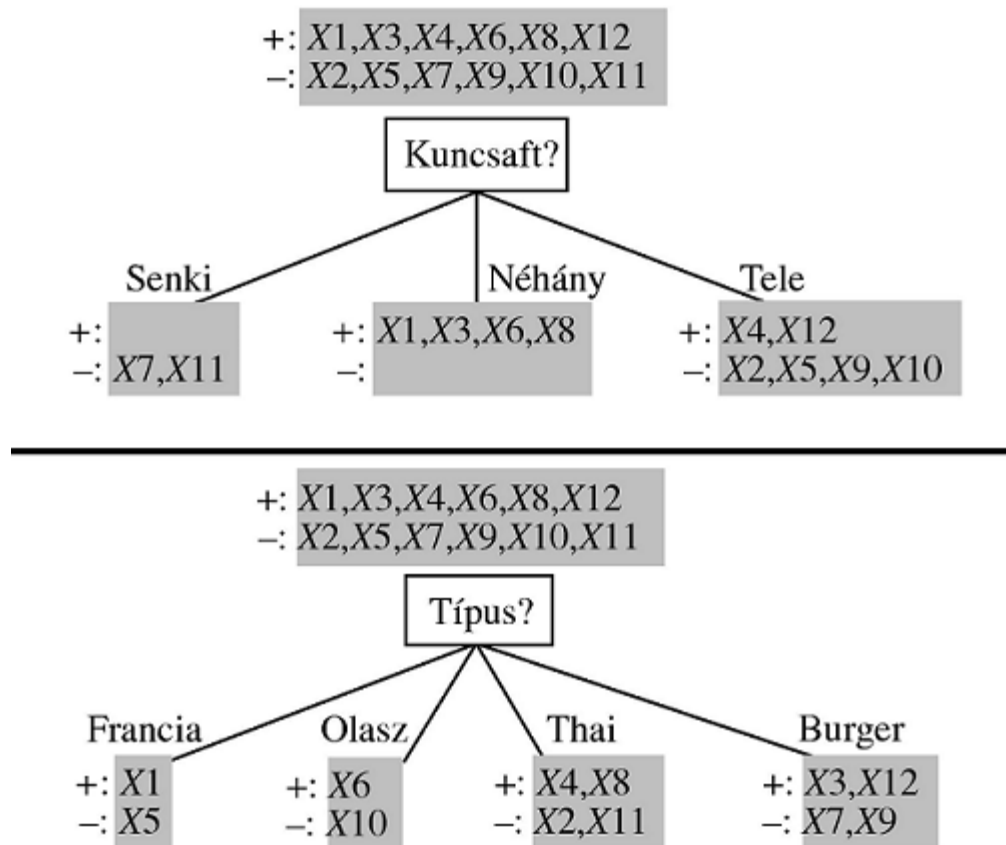
Mennyi információra volt szükségünk az attribútum tesztelése előtt, és mennyire van még szükségünk az attribútum tesztelése után? A kettő különbségét nyertük meg a teszttel!



Bármelyik  $A$  attribútum az  $E$  tanító halmazt  $E_1, E_2, \dots, E_N$  részhal-  
mazokra bontja az  $A$  tesztjére adott válaszoknak megfelelően, ha  $A$   
tesztje  $N$  különböző választ ad.



Nézzük meg a *Kuncsaft* és a *Típus* attribútumokat



$$\text{Nyeresség}(\text{Kuncsaft}) = 1 - \left[ \frac{2}{12} \cdot I(0,1) + \frac{4}{12} \cdot I(0,1) + \frac{6}{12} \cdot I\left(\frac{2}{4}, \frac{4}{4}\right) \right] \approx 0,54 \text{ bit}$$

$$\text{Nyeresség}(\text{Típus}) = 1 - \left[ \frac{2}{12} \cdot I\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{2}{12} \cdot I\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} \cdot I\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{4}{12} \cdot I\left(\frac{2}{4}, \frac{2}{4}\right) \right] = 0 \text{ bit}$$

Több attribútumra is kellene a számítás, de a Kuncsaft kimagaslóan jó.

# (Kuncsaft = Tele) ágon még fennmaradó példák

Példa                                      Attribútumok                                      Cél

*Altern   Bár   Pént   Éhes   Ár   Eső   Fogl   Típus   Becs   VárniFog*

|          |             |             |             |             |              |             |             |                |               |            |
|----------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|----------------|---------------|------------|
| $X_2$    | <i>Igen</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Igen</i> | <i>Olcsó</i> | <i>Nem</i>  | <i>Nem</i>  | <i>Thai</i>    | <i>30–60</i>  | <i>Nem</i> |
| $X_5$    | <i>Igen</i> | <i>Nem</i>  | <i>Igen</i> | <i>Nem</i>  | <i>Drága</i> | <i>Nem</i>  | <i>Igen</i> | <i>Francia</i> | <i>&gt;60</i> | <i>Nem</i> |
| $X_9$    | <i>Nem</i>  | <i>Igen</i> | <i>Igen</i> | <i>Nem</i>  | <i>Olcsó</i> | <i>Igen</i> | <i>Nem</i>  | <i>Burger</i>  | <i>&gt;60</i> | <i>Nem</i> |
| $X_{10}$ | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Drága</i> | <i>Nem</i>  | <i>Igen</i> | <i>Olasz</i>   | <i>10–30</i>  | <i>Nem</i> |

|          |             |             |             |             |              |            |            |               |              |             |
|----------|-------------|-------------|-------------|-------------|--------------|------------|------------|---------------|--------------|-------------|
| $X_4$    | <i>Igen</i> | <i>Nem</i>  | <i>Igen</i> | <i>Igen</i> | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i> | <i>Thai</i>   | <i>10–30</i> | <i>Igen</i> |
| $X_{12}$ | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i> | <i>Burger</i> | <i>30–60</i> | <i>Igen</i> |

Részfában:  $p_1 = 2$ ,  $n_1 = 4$ ,  $p_1/(p_1+n_1) = 1/3$ ,  $n_1/(p_1+n_1) = 2/3$

# (Kuncsaft = Tele) ág

$$I = I(\text{részfa}) = I(2/6, 4/6) = 0,9183 \quad (2 \text{ pozitív és } 4 \text{ negatív példa})$$

Ha pl. Alternatíva-t teszteljük, akkor a nyereség:

$$Ny(\text{Altern}) = I - \underbrace{[5/6 I(2/5, 3/5)]}_{\text{Altern = Igen}} + \underbrace{[1/6 I(0, 1)]}_{\text{Altern = Nem}}$$

Altern = Igen

$$\begin{aligned} p2 &= 2, \\ n2 &= 3 \end{aligned}$$

Altern = Nem

$$\begin{aligned} p3 &= 0 \\ n3 &= 1 \end{aligned}$$

} 6 példa

$$Ny(\text{Altern}) = I - [5/6 I(2/5, 3/5) + 1/6 \cancel{I(0, 1)}] \approx 0,92 - 0,81 = 0,11$$

# (Kuncsaft = Tele) ág

$$Ny(\text{Altern}) = I - [5/6 I(2/5, 3/5) + 1/6 I(0,1)] \approx 0,92 - 0,81 = 0,11$$

$$Ny(\text{Bár}) = I - [1/2 I(1/3, 2/3) + 1/2 I(1/3, 2/3)] = 0,92 - 0,92 = 0$$

$$Ny(\text{Péntek}) = I - [5/6 I(2/5, 3/5) + 1/6 I(0,1)] \approx 0,92 - 0,81 = 0,11$$

$$Ny(\text{Éhes}) = I - [4/6 I(1/2, 1/2) + 2/6 I(0,1)] \approx 0,92 - 0,66 = 0,25 \text{ Válasszuk pl. ezt!}$$

$$Ny(\text{Ár}) = I - [4/6 I(1/2, 1/2) + 2/6 I(0,1)] \approx 0,92 - 0,66 = 0,25$$

$$Ny(\text{Eső}) = I - [5/6 I(2/5, 3/5) + 1/6 I(0,1)] \approx 0,92 - 0,81 = 0,11$$

$$Ny(\text{Foglalt}) = I - [2/6 I(0,1) + 4/6 I(1/2, 1/2)] \approx 0,92 - 0,66 = 0,25$$

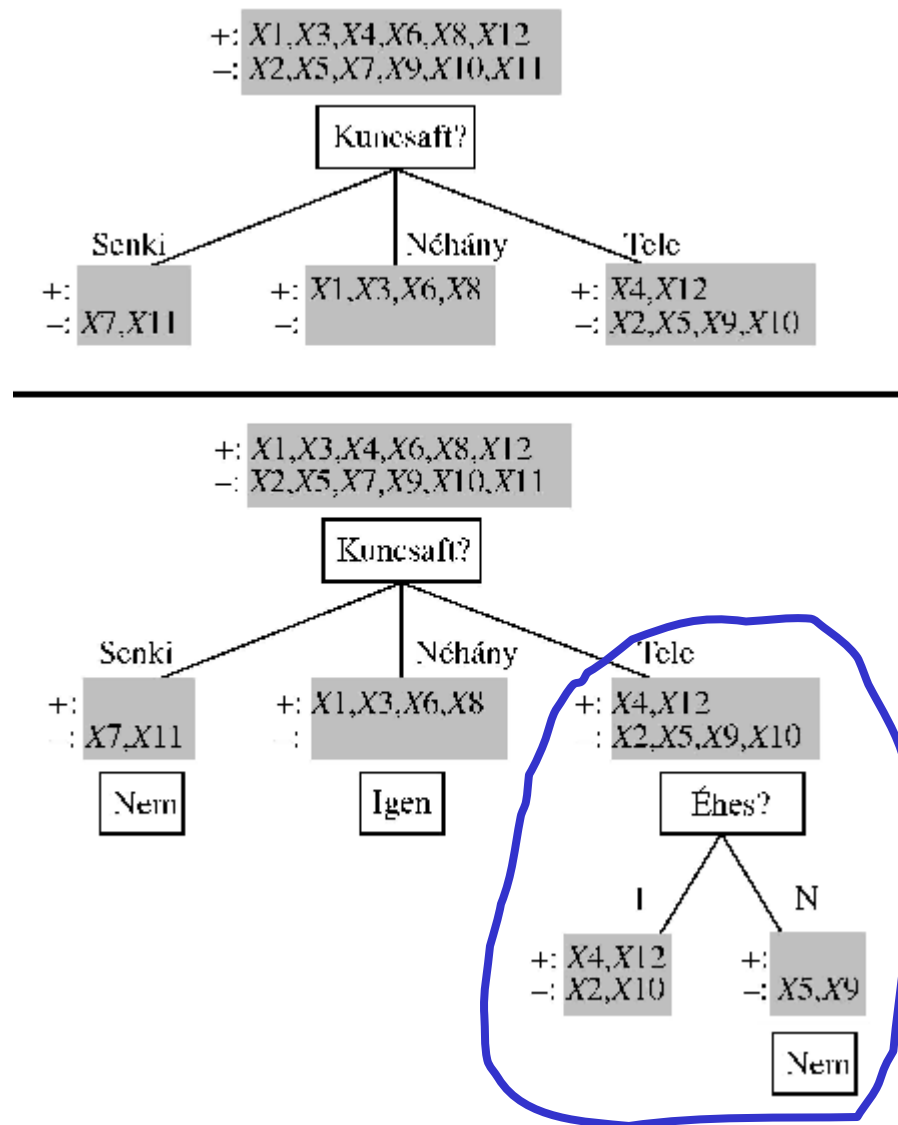
$$Ny(\text{Típus}) = I -$$

$$[2/6 I(1/2, 1/2) + 1/6 I(0,1) + 2/6 I(1/2, 1/2) + 1/6 I(0,1)] \approx 0,92 - 0,66 = 0,25$$

$$Ny(\text{Becs}) = I -$$

$$[2/6 I(1/2, 1/2) + 2/6 I(0,1) + 2/6 I(1/2, 1/2)] \approx 0,92 - 0,66 = 0,25$$

# (Kuncsaft = Tele) ÉS Teszt(Éhes?)



(Kuncsaft = Tele) ÉS (Éhes?=Igen) ágon még fennmaradó példák

| Példa    | Attribútumok |             |             |              |            |             |              |              | Cél             |
|----------|--------------|-------------|-------------|--------------|------------|-------------|--------------|--------------|-----------------|
|          | <i>Alt</i>   | <i>Bár</i>  | <i>Pént</i> | <i>Ár</i>    | <i>Eső</i> | <i>Fogl</i> | <i>Típus</i> | <i>Becs</i>  | <i>VárniFog</i> |
| $X_2$    | <i>Igen</i>  | <i>Nem</i>  | <i>Nem</i>  | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i>  | <i>Thai</i>  | <i>30–60</i> | <i>Nem</i>      |
| $X_{10}$ | <i>Igen</i>  | <i>Igen</i> | <i>Igen</i> | <i>Drága</i> | <i>Nem</i> | <i>Igen</i> | <i>Olasz</i> | <i>10–30</i> | <i>Nem</i>      |

|          |             |             |             |              |            |            |               |              |             |
|----------|-------------|-------------|-------------|--------------|------------|------------|---------------|--------------|-------------|
| $X_4$    | <i>Igen</i> | <i>Nem</i>  | <i>Igen</i> | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i> | <i>Thai</i>   | <i>10–30</i> | <i>Igen</i> |
| $X_{12}$ | <i>Igen</i> | <i>Igen</i> | <i>Igen</i> | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i> | <i>Burger</i> | <i>30–60</i> | <i>Igen</i> |

Részfában:  $p_1 = 2, n_1 = 2, \quad p_1/(p_1+n_1) = 1/2, \quad n_1/(p_1+n_1) = 1/2$   
Információsükséglet a következő teszt előtt ezen az ágon: 1 bit!

Különböző attribútumok nyeresége a **(Kunsaft = Tele)** ÉS **(Éhes?=Igen)**

$$Ny(Alt) = I - [1/1 I(1/2, 1/2) + 0] = 1 - 1 = 0$$

$$Ny(Bár) = I - [1/2 I(1/2, 1/2) + 1/2 I(1/2, 1/2)] = 1 - 1 = 0$$

$$Ny(Pént/Szo) = I - [1/4 I(0,1) + 3/4 I(1/3,2/3)] \approx 1 - 0,92 = 0,08$$

$$Ny(Ár) = I - [1/4 I(0,1) + 3/4 I(1/3,2/3)] \approx 1 - 0,92 = 0,08$$

$$Ny(Eső) = I - [1/1 I(1/2, 1/2) + 0] = 1 - 1 = 0$$

$$Ny(Foglalt) = I - [1/4 I(0,1) + 3/4 I(1/3,2/3)] \approx 1 - 0,92 = 0,08$$

$$Ny(Típus) = I - [1/4 I(0,1) + 1/4 I(0,1) + 1/2 I(1/2, 1/2)] = 0,5 \quad \underline{\text{Válasszuk pl. ezt!}}$$

$$Ny(Becs) = I - [1/2 I(1/2, 1/2) + 1/2 I(1/2, 1/2)] = 1 - 1 = 0$$



# (Kuncsaft = Tele) ÉS (Éhes?=Igen) ÉS (Típus = Thai) ágon még fennmaradó példák

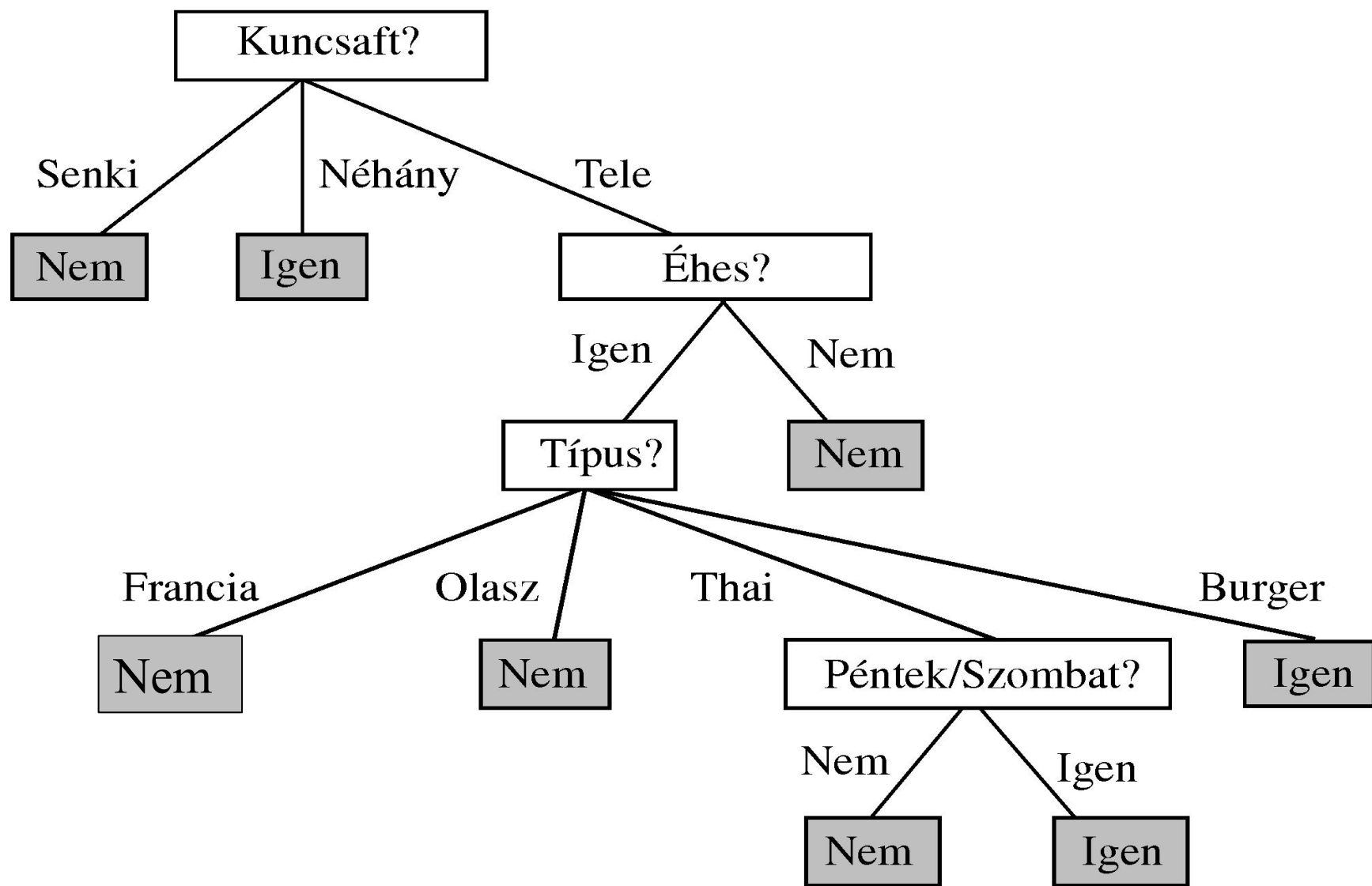
| Példa | Attribútumok |            |                 |              |            |             | Cél          |                 |
|-------|--------------|------------|-----------------|--------------|------------|-------------|--------------|-----------------|
|       | <i>Alt</i>   | <i>Bár</i> | <i>Pént/Szo</i> | <i>Ár</i>    | <i>Eső</i> | <i>Fogl</i> | <i>Becs</i>  | <i>VárniFog</i> |
| $X_2$ | <i>Igen</i>  | <i>Nem</i> | <i>Nem</i>      | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i>  | <i>30–60</i> | <i>Nem</i>      |
| $X_4$ | <i>Igen</i>  | <i>Nem</i> | <i>Igen</i>     | <i>Olcsó</i> | <i>Nem</i> | <i>Nem</i>  | <i>10–30</i> | <i>Igen</i>     |

Részfában:  $p_1 = 1$ ,  $n_1 = 1$ ,  $p_1/(p_1+n_1) = 1/2$ ,  $n_1/(p_1+n_1) = 1/2$

A Pént/Szomb és a Becs jó, a többi nem jó jellemző.

Hiszen esetükben a teszt után ugyanúgy  $\frac{1}{2}$  -  $\frac{1}{2}$  lesz a pozitív és negatív példák aránya. (Pl. a *Bár* = *Nem* vagy *Eső* = *Nem* vagy *Ár* = *Olcsó* stb. mentén (azaz ezekbe az irányokba nem érdemes építeni a fát).

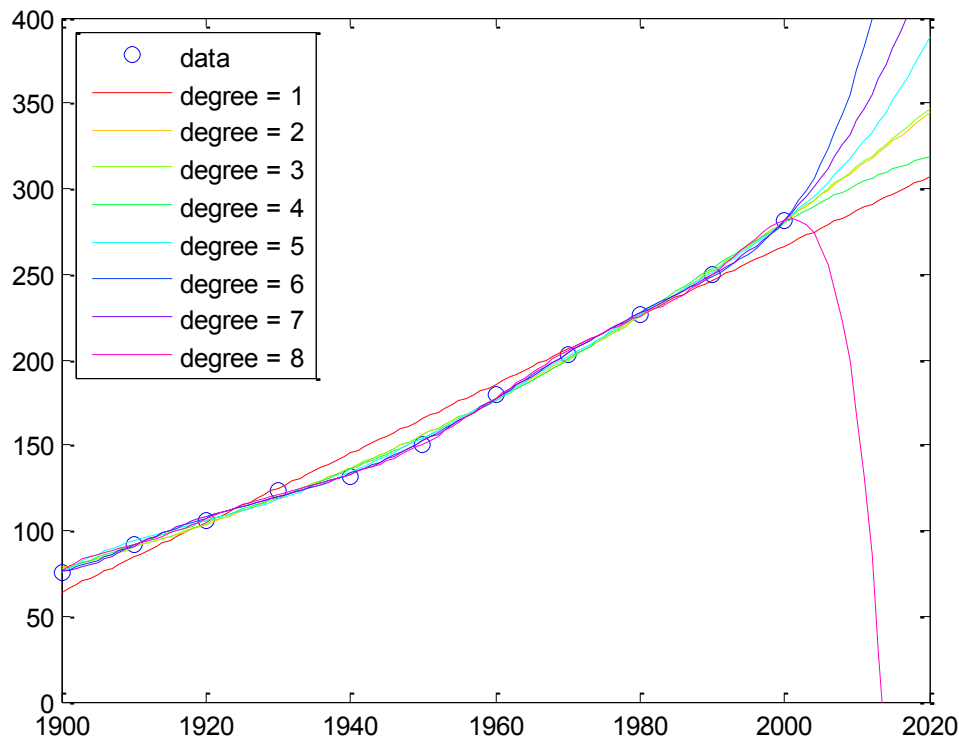
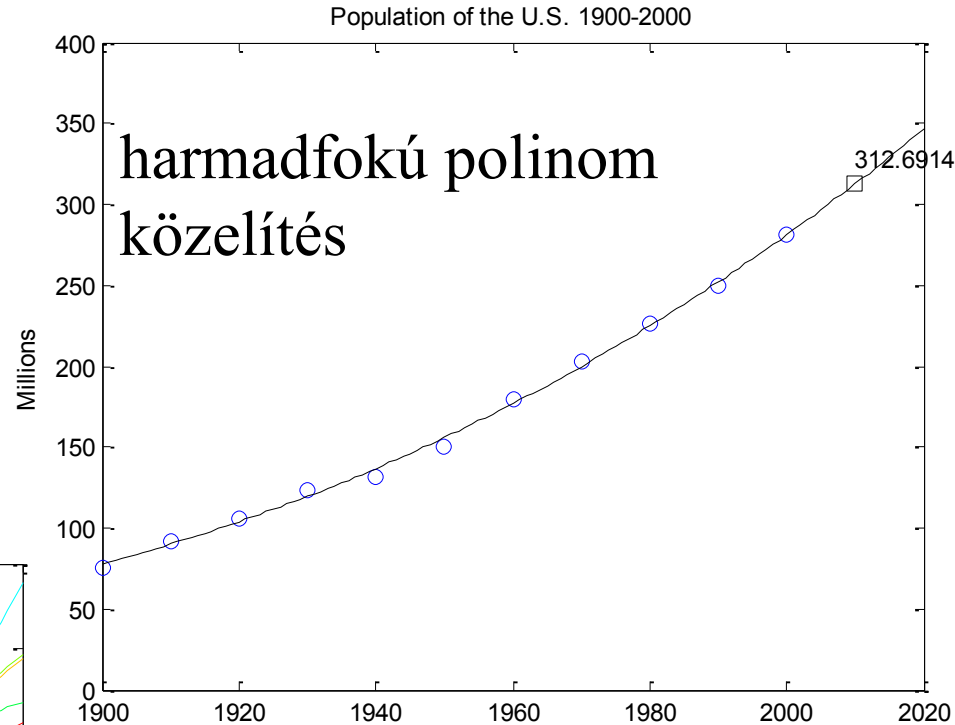
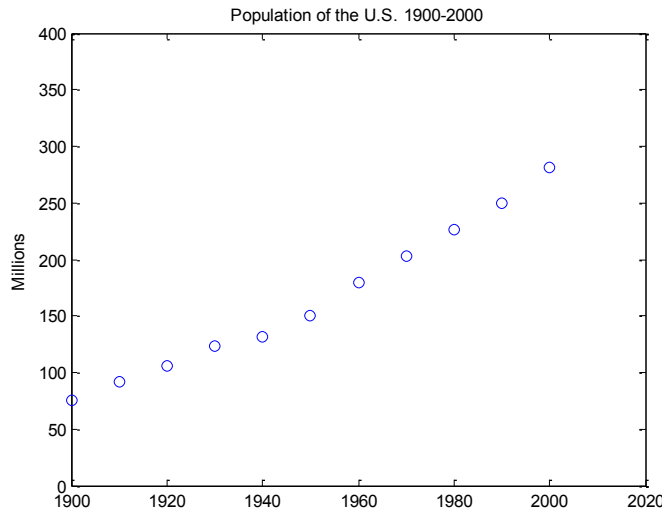
# A kialakított döntési fa



Ha pl. a Pént/Szomb helyett a Becs attribútumot választjuk – ugyanúgy konzisztens a kialakuló fa mindegyik példánkkal, DE máshogy általánosít! (Az ismeretlen új mintákra esetleg más választ ad.)

# A mintákat torzító zaj hatása – **túltanulást** okozhat!

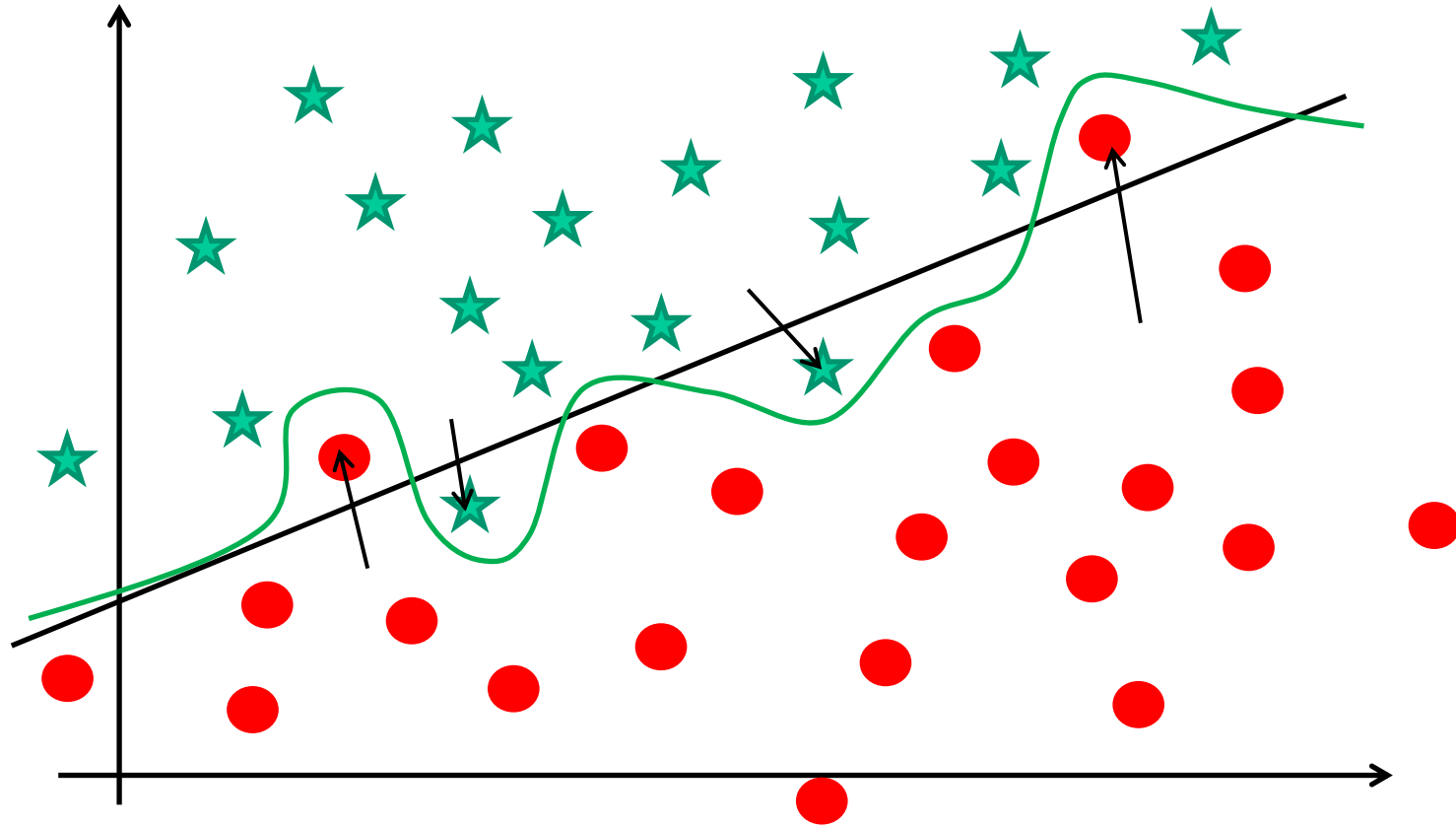
(a faék dicsérete! ☺)



Matlab demó: az 1900-2000 közötti népszámlálási adatokból becsüljük 2010-es USA népességet. (Regressziós feladat, de a jelenség a döntéseknél hasonló hatást okoz.)

A nagyon „okos” = túl komplex eszköz a zajt is megtanulja, rosszabbul becsül, mint az egyszerűbb!

A döntési határ a zaj miatt bonyolultabbá válhat

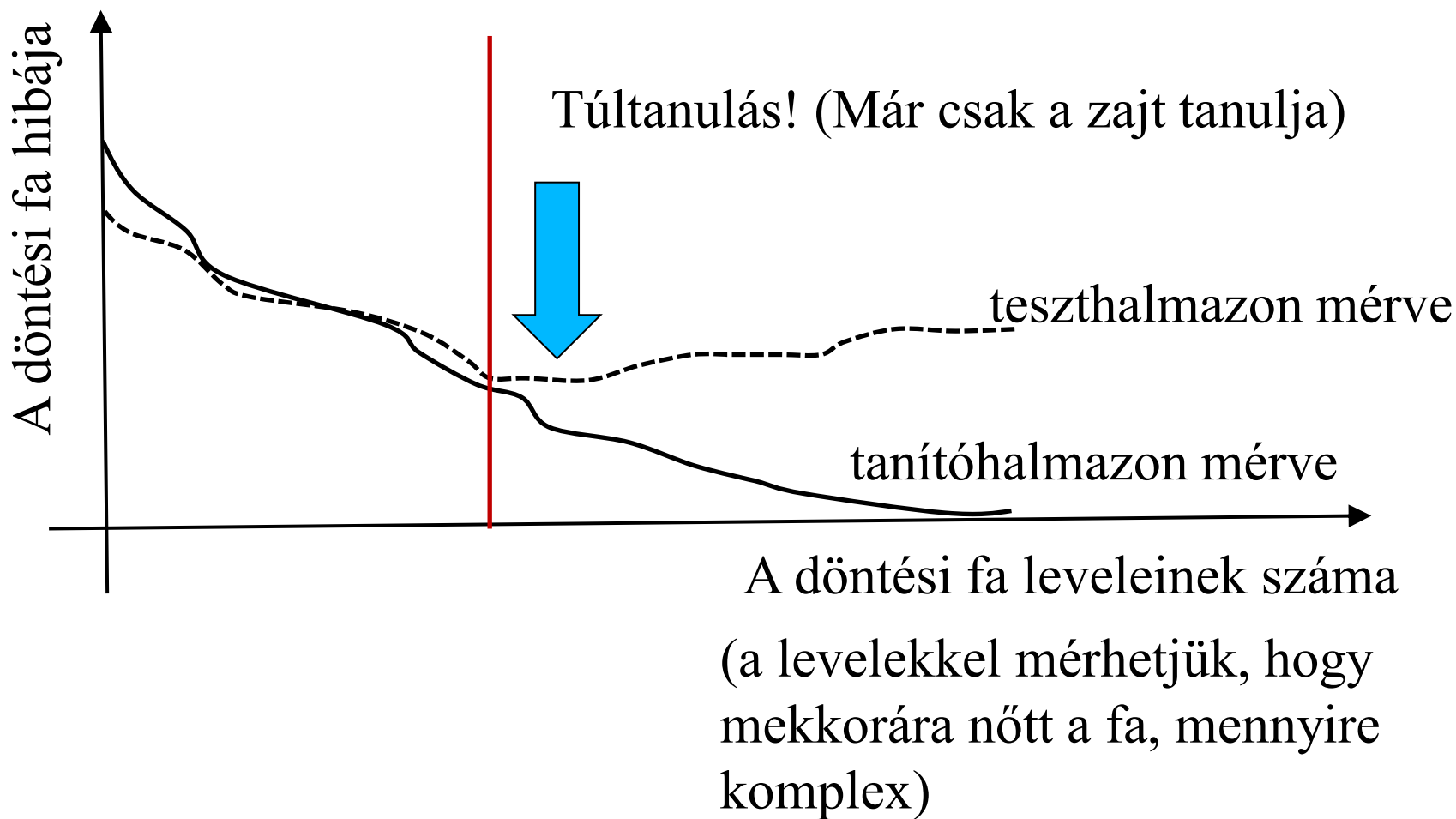


Ezt csak egy komplex (összetett) eszköz képes megtanulni.  
DE ha ez csak a zaj miatt ilyen bonyolult, akkor NEM JÓ, ha megtanulja az eszköz. A lényeges infót (egyenes vagy más sima elválasztó határ) csak torzítja.

# Mikor álljunk le a döntési fa növesztésével?

Alapvetően 2 stratégia:

## 1. Korai leállás



# Döntési fa nyesése

## 2. Engedjük túlnőni, majd visszavágjuk, visszanyessük

Ismerjük fel, ha nem releváns attribútumot használtunk, és az adott ágat ne fejtsük ki tovább, sőt messük le a kialakított részfat.

Irreleváns attribútum: példahalmaz kettévágásánál a kapott részhalmazok kb. azonos arányban tartalmazznak pozitív és negatív példát, mint az eredeti halmaz.

Az információ nyereség ilyenkor közel 0.

Az információ nyereség hiánya az irrelevancia jó jelzése.

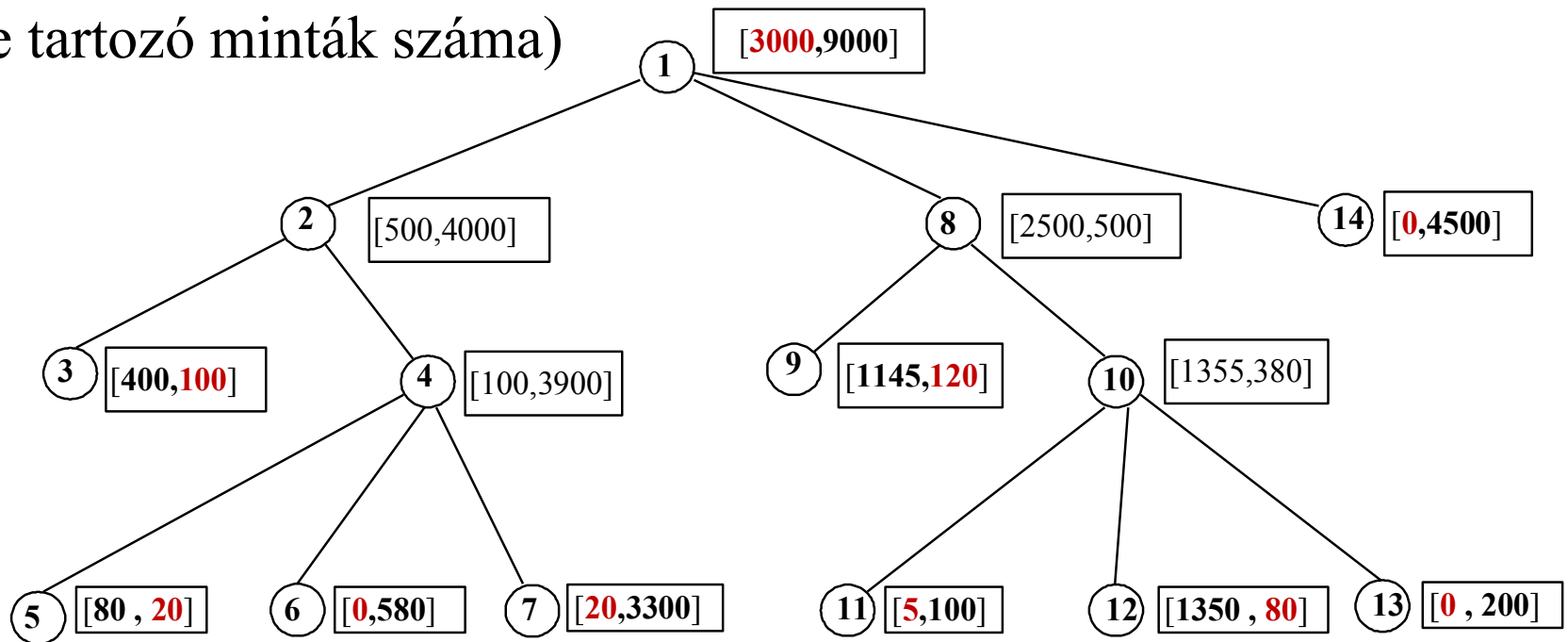
Milyen nagy legyen az információ nyereség, hogy egy attribútumot mégis felhasználjunk a példahalmaz megosztására?

## Döntési fa nyesése:

## Hibaarány – komplexitás kompromisszumon alapuló metszés

Egy kétosztályos (bináris) osztályozási példa kapcsán

(Baloldali szám a C1 osztályba tartozó minták száma, jobboldali a C2-be tartozó minták száma)



**Komplexitás** – legyen a levelek száma (lehetne más is), a példánkban:

a gyökérnél 1, a kifejtett fánál 9

**Hibaarány** (használhatnánk az információsükségletet is):

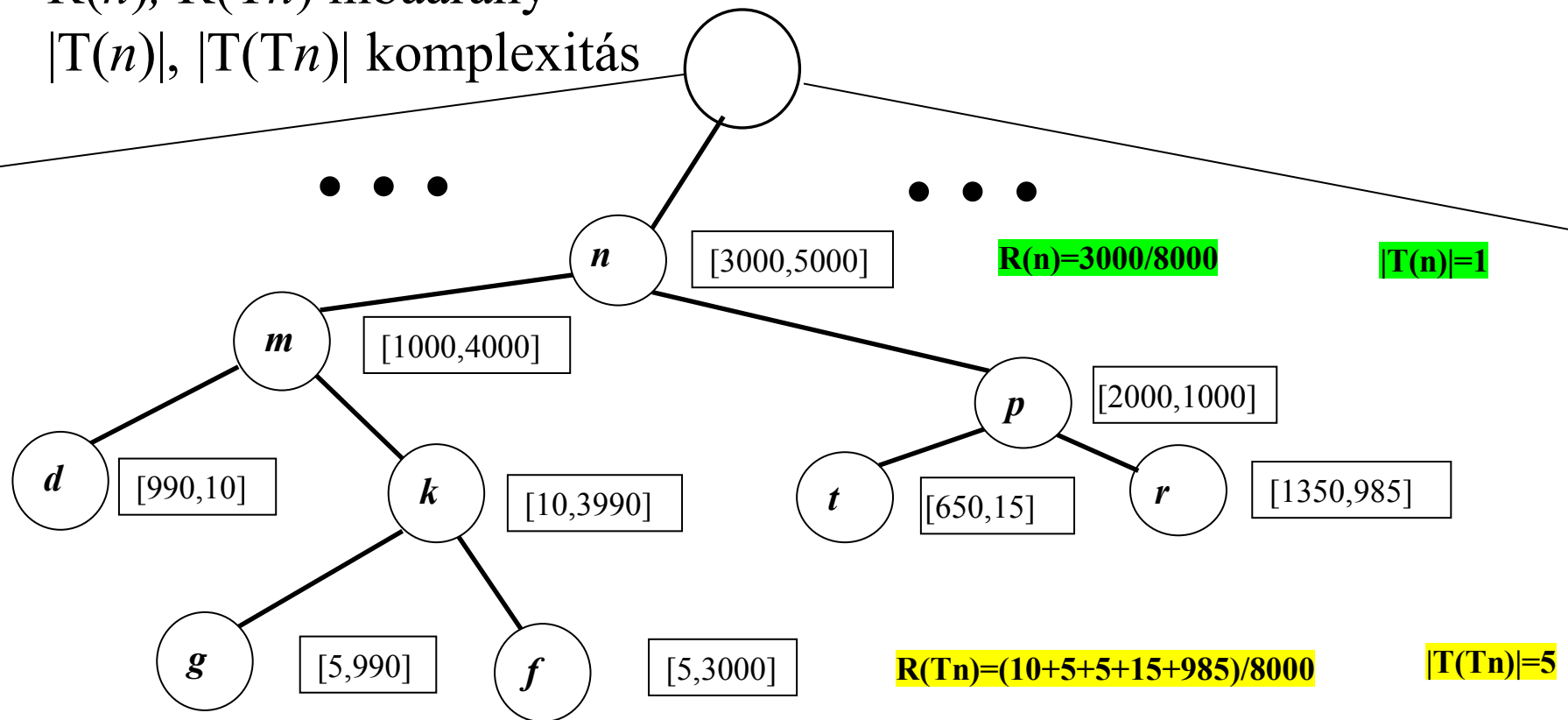
a gyökérnél  $3000/12000=0,25$  (hiszen a jobb válaszunk C2 lenne),

a kifejtett fánál:  $(100+20+0+20+120+5+80+0+0)/12000=0,029$

# Hibaarány – komplexitás kompromisszumon alapuló metszés

Számozzuk a csomópontokat, jelölés:

- az  $n$ -dik csomópontot önmagában (mintha levél lenne) jelölje  $n$
- az  $n$ -edik csomópontból kiinduló részfa  $T_n$
- $R(n)$ ,  $R(T_n)$  hibaarány
- $|T(n)|$ ,  $|T(T_n)|$  komplexitás



*Hogyan hasonlítsuk össze a komplexitást a hibával??*



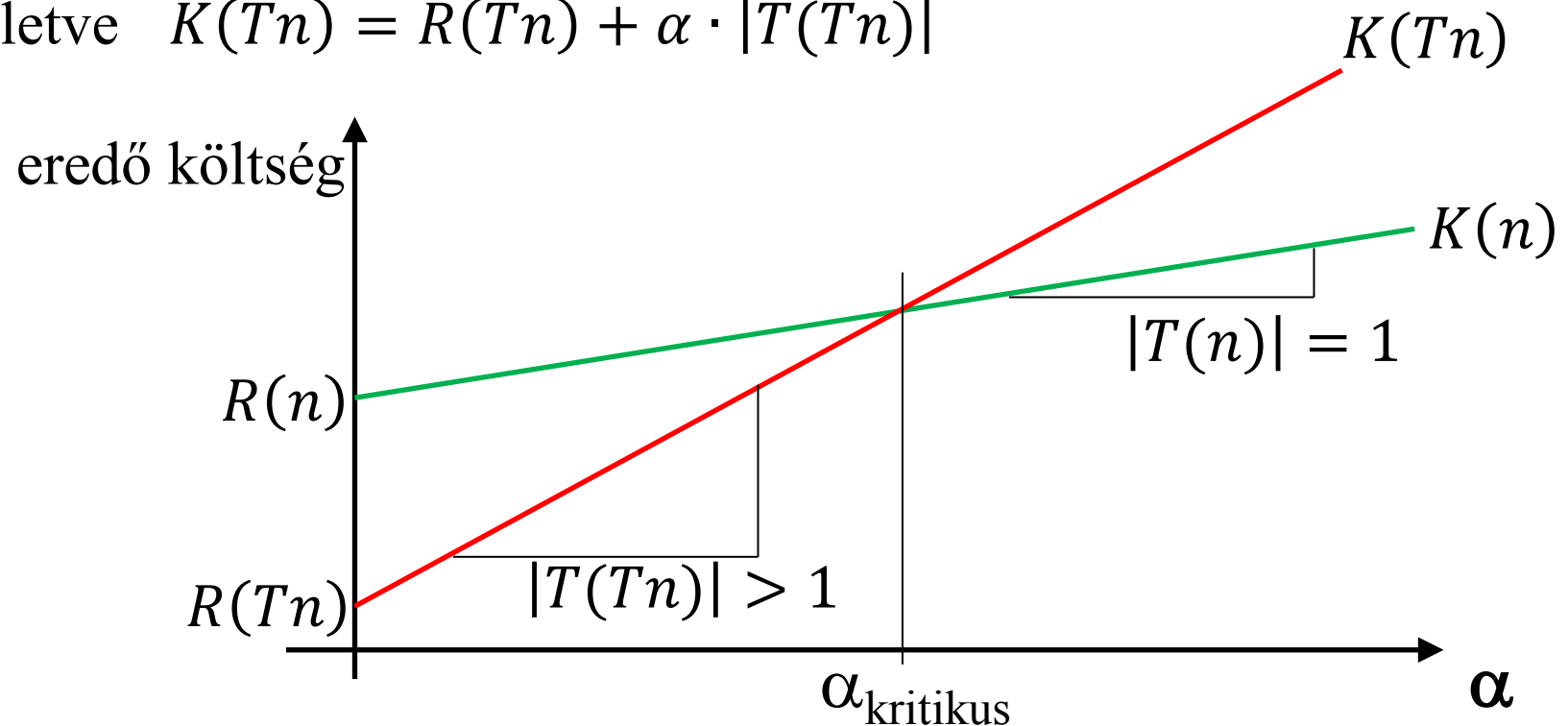
Legyen az összetett, eredő költség (pl. az  $n$ -dik csomópontra):

$$K^*(n) = \text{EgységnyiHibaKöltsége} \cdot R(n) + \text{EgységnyiKomplexitásKöltsége} \cdot |T(n)|$$

Legyen a két költség aránya:  $\alpha = \frac{\text{EgységnyiKomplexitásKöltsége}}{\text{EgységnyiHibaKöltsége}}$

$$\text{Így} \quad K(n) = \frac{K^*(n)}{\text{EgységnyiHibaKöltsége}} = R(n) + \alpha \cdot |T(n)|$$

illetve  $K(Tn) = R(Tn) + \alpha \cdot |T(Tn)|$



$\alpha_{\text{kritikus}}$  amikor mindegy, hogy levágjuk-e a részfát, az összetett költség azonos az  $n$ -edik csomópontra, mint levélre, és az  $n$ -dik csomópontból kiinduló  $T_n$  részfára

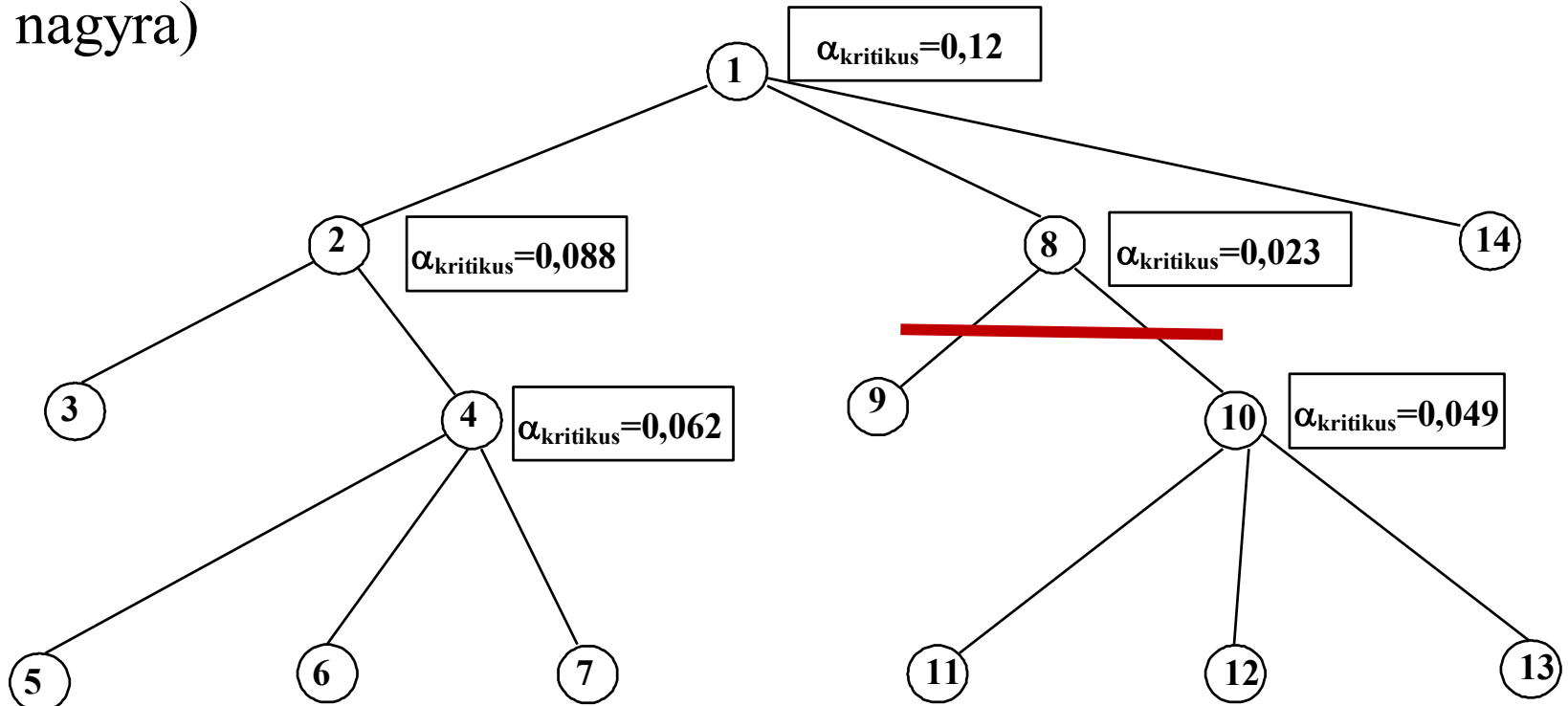
$$R(n) + \alpha_{\text{kritikus}} \cdot |T(n)| = R(Tn) + \alpha_{\text{kritikus}} \cdot |T(Tn)|$$

$$\alpha_{\text{kritikus}} = \frac{R(n) - R(Tn)}{|T(Tn)| - |T(n)|} = \frac{R(n) - R(Tn)}{|T(Tn)| - 1}$$

1. Ha  $\alpha=0$ , akkor a komplexitásnak nincs ára, soha nem érdemes visszametszeni
2.  $0 < \alpha < \alpha_{\text{kritikus}}$ , akkor még mindig jobb nem visszametszeni
3. Ha  $\alpha = \alpha_{\text{kritikus}}$ , akkor mindegy, hogy visszavágjuk-e
4. Ha  $\alpha_{\text{kritikus}} < \alpha$ , akkor érdemes visszavágni, olcsóbb abbahagyni az  $n$ -dik levélnél

Gondoljuk végig, hogy mi történik, ha  $\alpha$ -t folyamatosan növeljük 0-tól indulva!

**➡** amelyik csomópont  $\alpha_{\text{kritikus}}$  értékét először érjük el, ott érdemes *leginkább* metszeni a fát! (Persze ha a hiba nem nő túl nagyra)



A számok légből kapottak (csak a példa kedvéért)