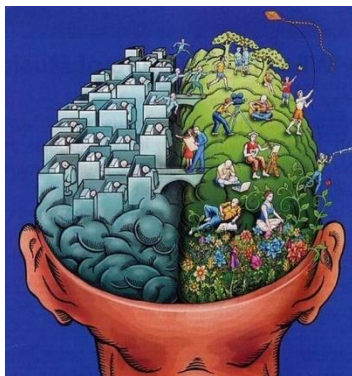




Budapesti Műszaki és Gazdaságtudományi Egyetem

Méréstechnika és Információs rendszerek Tanszék



Felügyelt tanulás, döntési fák

Előadó: Hullám Gábor

Előadás anyaga: Dobrowiecki Tadeusz
Pataki Béla

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

Nagyon gyakran a számítógép tanulásához rendelkezésünkre áll egy csomó mintapélda, 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 – az, hogy hogyan használjuk fel a mintákat – bizonyos mértékben problémafüggetlen.

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

Egy algoritmust, pl. egy döntést (ágensfüggvényt) alapvetően két módon lehet megvalósítani?

(1) Analitikus tervezés:

Begyűjteni az analitikus (fizikai, kémiai stb. összefüggésekből felépített) modelleket az adott problémára.

Megtervezni analitikusan a konkrét mechanizmust, és azt algoritmusként implementálni.

(2) Tanulás:

Megtervezni a tanulás mechanizmusát, és azt algoritmusként implementálni.

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

A következőkben a (2)-vel foglalkozunk

Tanulás alapvető fajtái:

felügyelt tanulás egy esetről mind a bemenetet,
mind a kimenetet észlelni tudjuk (bemeneti minta + kívánt válasz)

megerősítéses tanulás az ágens az általa végrehajtott
tevékenység csak bizonyos értékelését kapja meg, esetleg nem is
minden lépésben (jutalom, büntetés, **megerősítés**)

felügyelet nélküli tanulás semmilyen információ sem áll
rendelkezésünkre a helyes kimenetről (az észlelések közötti
összefüggések tanulása)

féligellenőrzött tanulás a tanításra használt esetek egy részénél mind
a bemenetet, mind a kimenetet észlelni tudjuk (bemeneti minta +
kívánt válasz), a másik – tipikusan nagyobb – részénél csak a bemeneti
leírás ismert

Felügyelt tanulás (pl. induktív következtetés):

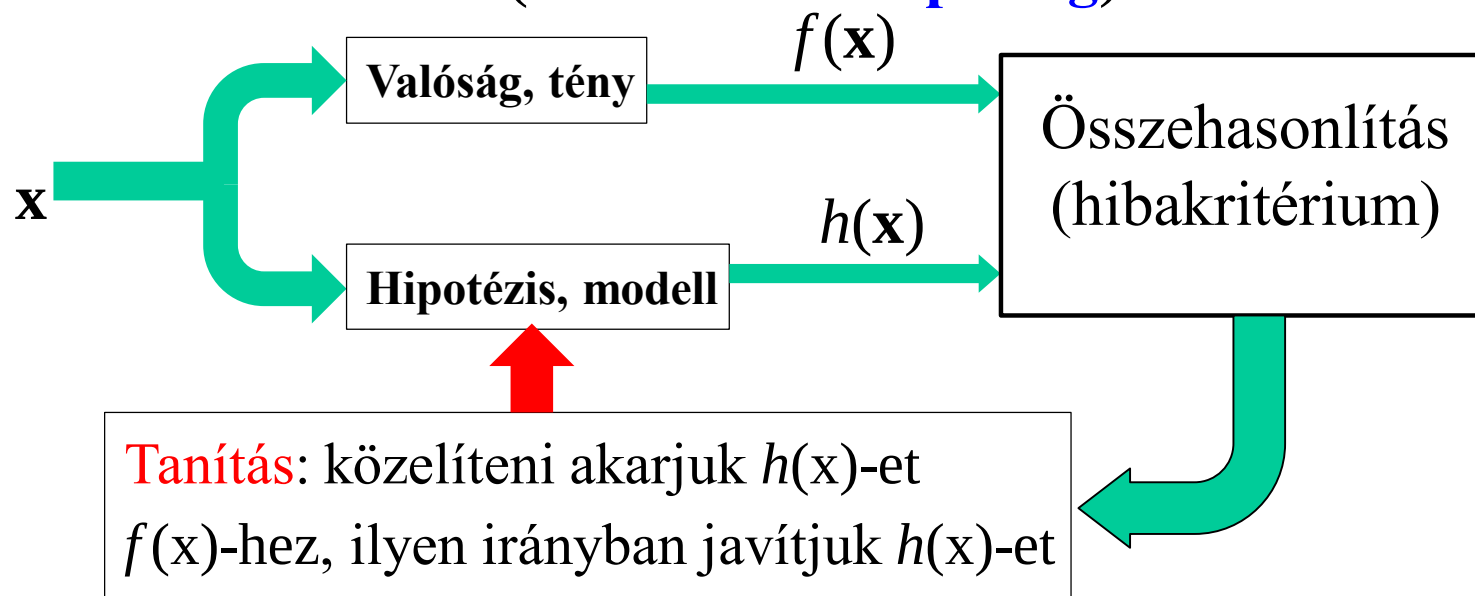
tanulási példa: $(\mathbf{x}, f(\mathbf{x}))$ adatkár, ahol $f(\mathbf{x})$ ismeretlen

tanulás célja: $f(\mathbf{x})$ értelmes közelítése egy $h(\mathbf{x})$ hipotézissel

$h(\mathbf{x}) = f(\mathbf{x})$, $\mathbf{x}$ – ismert példákban gyakran teljes pontosság

$h(\mathbf{x}') \cong f(\mathbf{x}')$, $\mathbf{x}'$ – a tanulás közben még nem látott

esetek (**általánosító képesség**)



Feladat: az ismeretlen f -re vonatkozó példák egy halmaza alapján (*tanítóhalmaz*), adjon meg egy olyan h függvényt (*hipotézist*), amely tulajdonságaiban jól közelíti az f -et (amit *teszthalmazon* verifikálunk).

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: döntés, hogy egy adott étteremben várjunk-e asztalra.

Cél: előállítani a döntés logikai függvényét

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 bementett várakozási idő
(0-10 perc, 10-30, 30-60, >60 perc)

A **döntés eredményét** a döntési ágensfüggvény adja meg:

VárniFog = $f(\textit{Alternativa}, \textit{Bár}, \textit{Pén/Szom}, \textit{Éhes}, \textit{Kuncsaft}, \textit{Ár}, \textit{Eső}, \textit{Foglalás}, \textit{Típus}, \textit{BecsültVár}, \dots)$

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	Igen	Nem	Nem	Igen	Néhány	Drága	Nem	Igen	Francia	0–10	Igen
X_2	Igen	Nem	Nem	Igen	Tele	Olcsó	Nem	Nem	Thai	30–60	Nem
X_3	Nem	Igen	Nem	Nem	Néhány	Olcsó	Nem	Nem	Burger	0–10	Igen
X_4	Igen	Nem	Igen	Igen	Tele	Olcsó	Nem	Nem	Thai	10–30	Igen
X_5	Igen	Nem	Igen	Nem	Tele	Drága	Nem	Igen	Francia	>60	Nem
X_6	Nem	Igen	Nem	Igen	Néhány	Közep	Igen	Igen	Olasz	0–10	Igen
X_7	Nem	Igen	Nem	Nem	Senki	Olcsó	Igen	Nem	Burger	0–10	Nem
X_8	Nem	Nem	Nem	Igen	Néhány	Közep	Igen	Igen	Thai	0–10	Igen
X_9	Nem	Igen	Igen	Nem	Tele	Olcsó	Igen	Nem	Burger	>60	Nem
X_{10}	Igen	Igen	Igen	Igen	Tele	Drága	Nem	Igen	Olasz	10–30	Nem
X_{11}	Nem	Nem	Nem	Nem	Senki	Olcsó	Nem	Nem	Thai	0–10	Nem
X_{12}	Igen	Igen	Igen	Igen	Tele	Olcsó	Nem	Nem	Burger	30–60	Igen

Mi benne a viselkedésünk 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

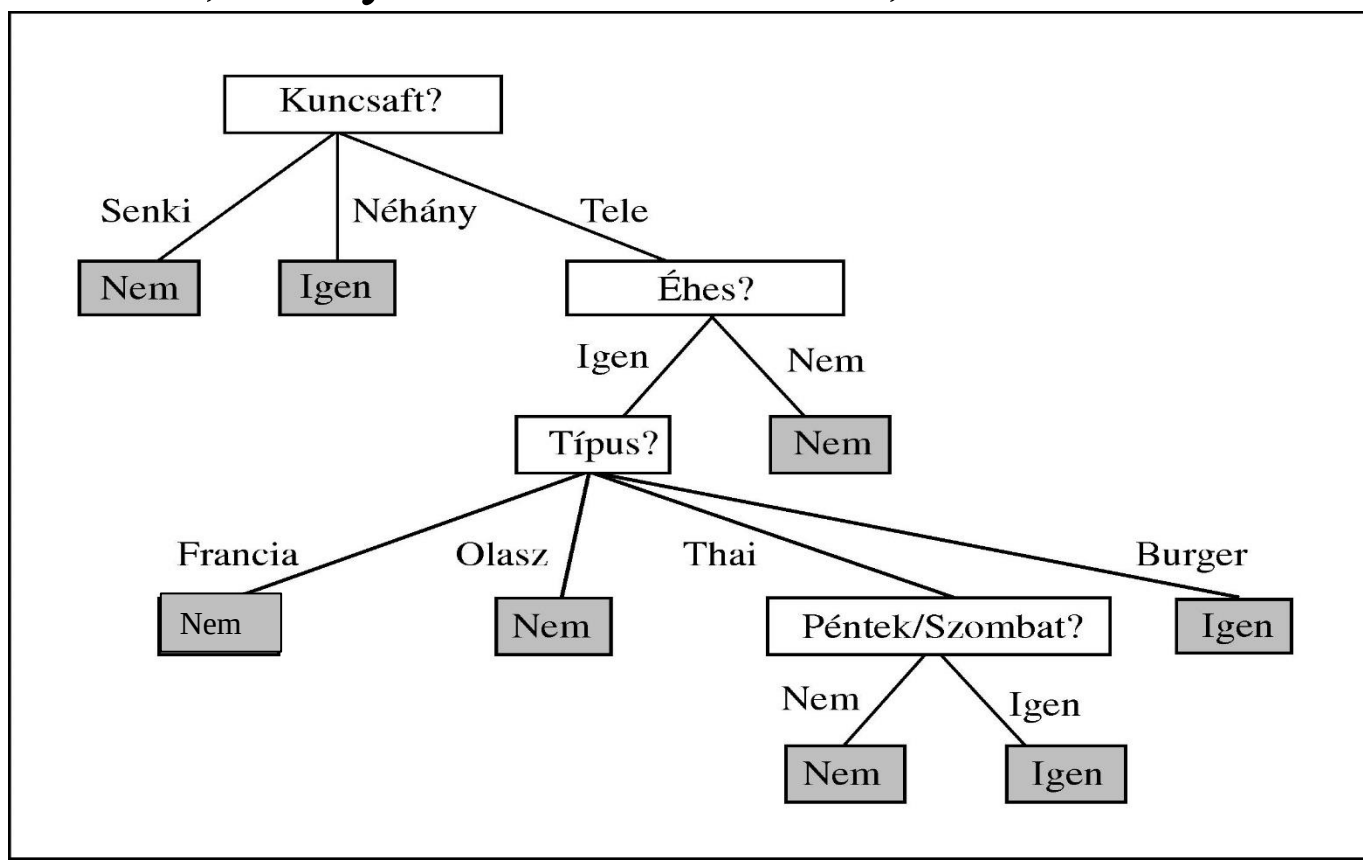
bemenete: egy tulajdonsághalmazzal leírt objektum vagy szituáció

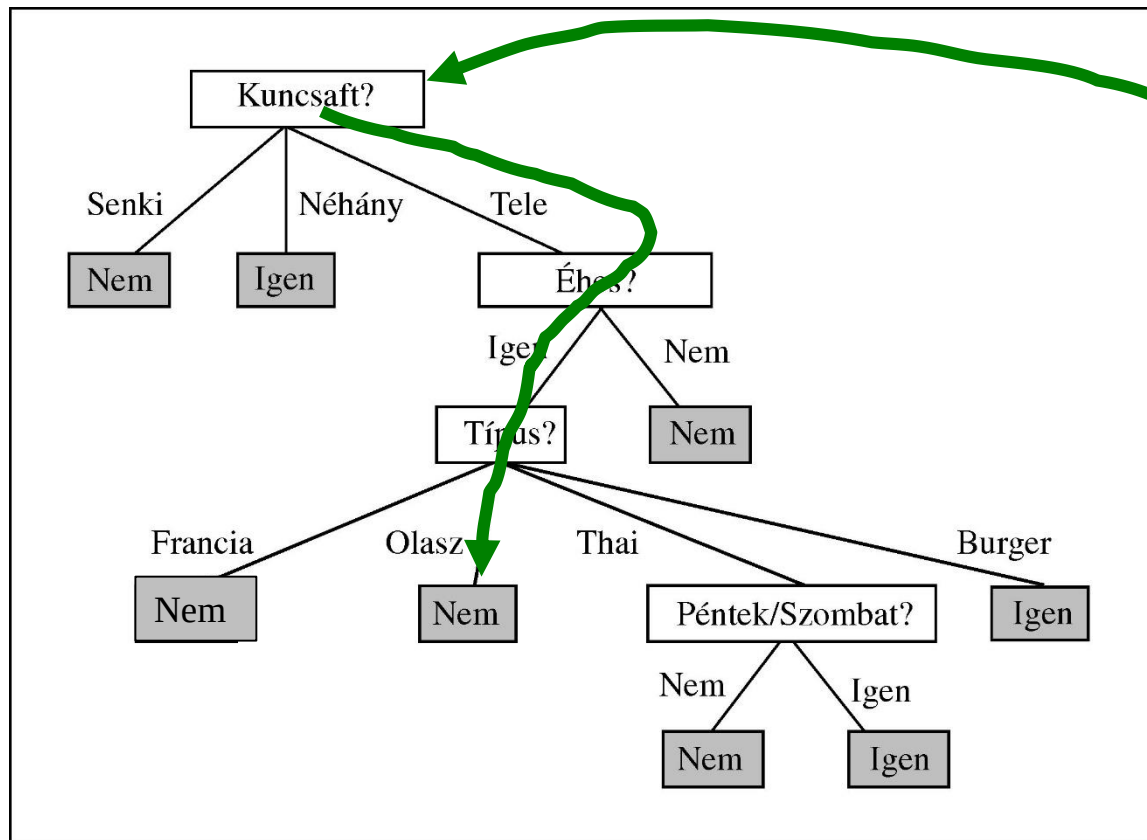
kimenete: egy igen/nem „döntés/cselekvés”

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
Típus = Olasz

VárniFog = (Kuncsaft = Néhány) V

((Kuncsaft = Tele) ∧ Éhes ∧ (Típus = Thai) ∧ Pént/Szomb) V

((Kuncsaft = Tele) ∧ Éhes ∧ (Típus = Burger))

Döntési fák kifejezőképessége

teljes - minden (ítélet)logikai függvény felírható döntési faként.

(az igazságtábla minden sora = a fa egy bejárása,
de így az igazságtábla mérete = a fa mérete = exponenciális!)

Döntési fák kialakítása példák alapján

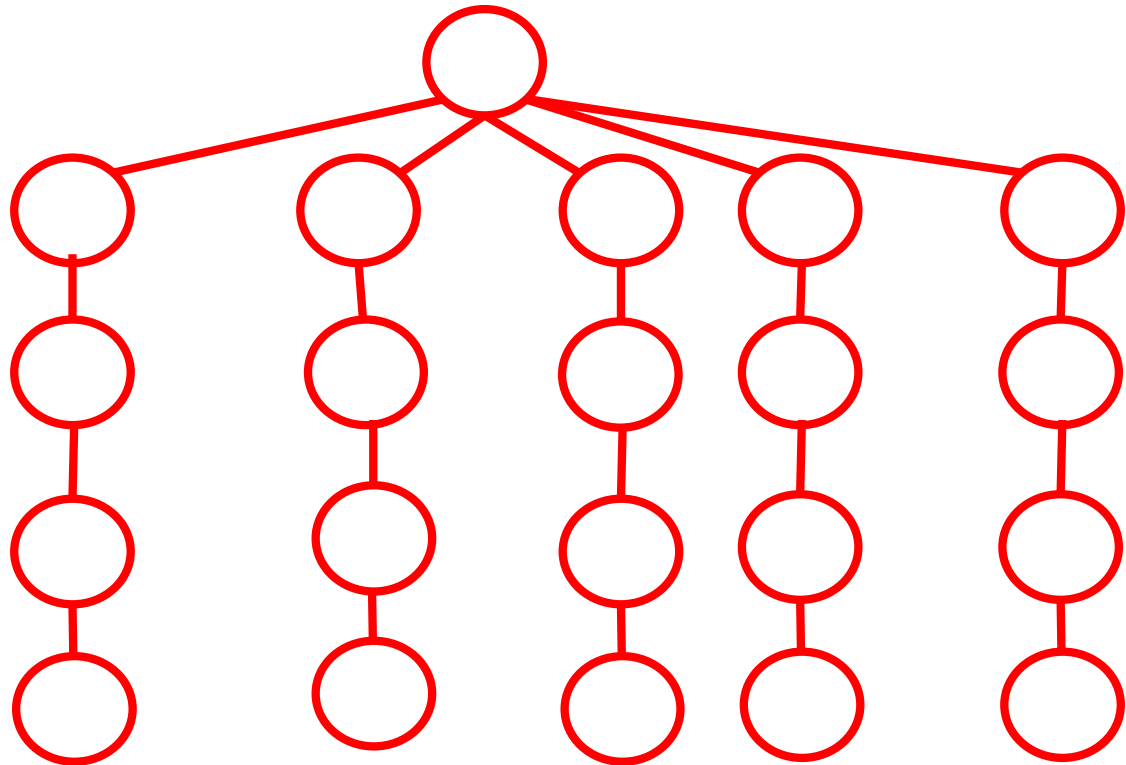
példa: (attribútumok értékei, célpredikátum értéke)

példa besorolása: a célpredikátum értéke

pozitív/negatív példa: a célpredikátum értéke igaz / hamis

tanító halmaz: a teljes példahalmaz

triviális fa, könnyű - mindegyik példához egy önálló bejárési út
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
extrapolálni)

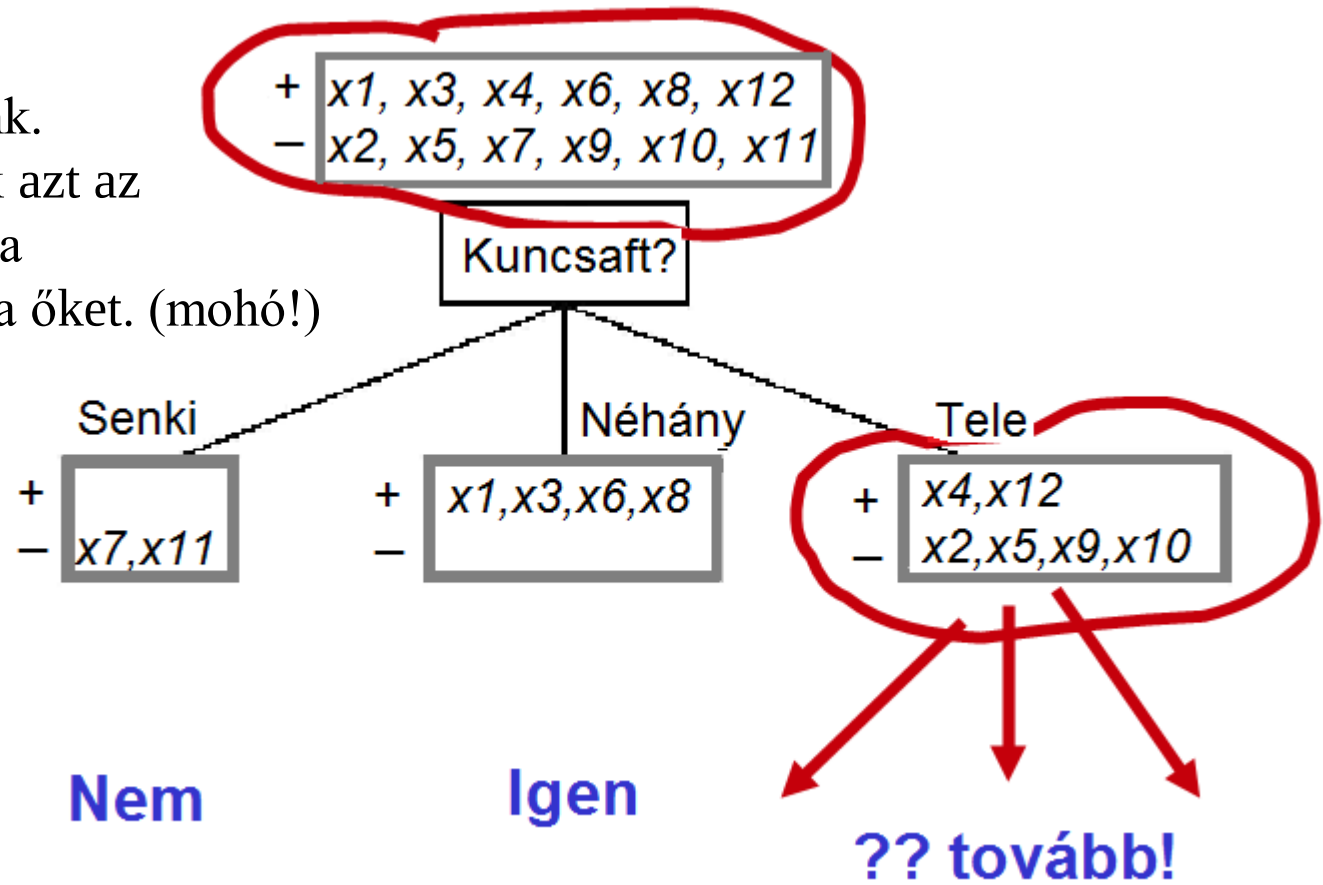


„igazi” fa - a jellegzetes minták kinyerése,
a módszer képes nagyszámú esetet **tömör** formában leírni, és így
az egyelőre ismeretlen példákra is **általánosítani**

A döntési fa építése – általános jelenségek

Van néhány pozitív és néhány negatív példánk.

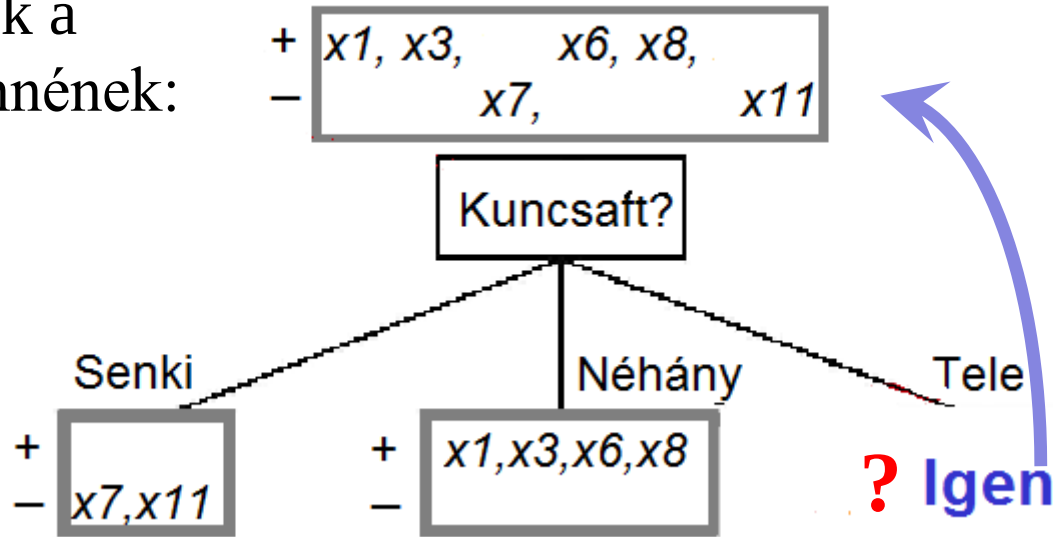
Heurisztika: válasszuk azt az attribútumot, amelyik a legjobban szétválasztja őket. (mohó!)



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

A döntési fa építése – általános jelenségek

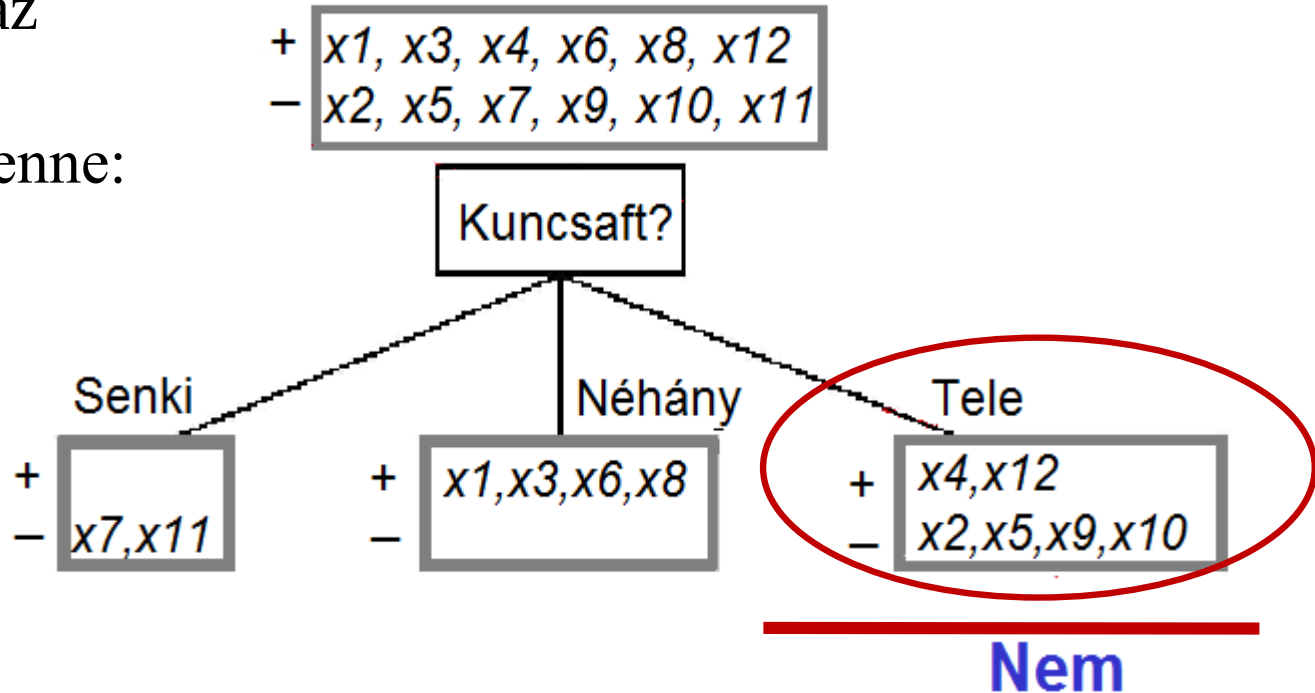
Ha csak ezek a
példáink lennének:



Ha **nem maradt egyetlen példa sem**, ez azt jelenti, hogy ilyen példát nem figyeltünk meg eddig...., de a jövőben mégis jelentkezhet. Ilyenkor valamilyen alapértéket adunk vissza, amelyet rendszerint a csomópont szülőjének többségi válaszából származtatunk.

A döntési fa építése – általános jelenségek

Ha csak ez az egyetlen attribútum lenne:



Nem maradt már teszteletlen attribútum, de maradtak még pozitív és negatív példák. Baj van.

Ezeknek a példáknak pontosan megegyezik a leírása, de különböző a besorolásuk.

Néhány adat tehát nem korrekt: a **zaj** torzítja az adatokat – vagy hiányzik még attribútum (infó).

Megoldás? Pl. a **többségi szavazás** használata.

Döntési fák kialakítása példák alapján

példa: (attribútumok értékei, célpredikátum értéke)

példa besorolása: a célpredikátum értéke

pozitív/negatív példa: a célpredikátum értéke igaz / hamis

tanító halmaz: a teljes példahalmaz

A legkisebb döntési fa megtalálása - általánosságban nem megoldható

Heurisztika: mohóság - egy meglehetősen egyszerű (jó) fa is jó lesz!

Az alapötlet: először a „legfontosabb” attribútumot teszteljük.

„legfontosabb” = a legnagyobb eltérést okozza példák besorolásában

Elvárás: **kisszámú példa** alapján korrekt besoroláshoz jussunk:

a bejárési utak rövidek legyenek, és így az egész fa kicsi (tömör) legyen.

Döntési fák kialakítása példák alapján

A **tökéletes attribútum** a példákat két csoportra bontja,
az egyikbe **csak pozitív**,
a másikba **csak negatív** példák tartoznak.

Ezzel be is lehetne fejezni a fa tanítását – de tökéletes attribútum általában nincs!

Olyan attribútumot válasszunk, amely a lehető legmesszebbre jut a pontos besorolás biztosításában.

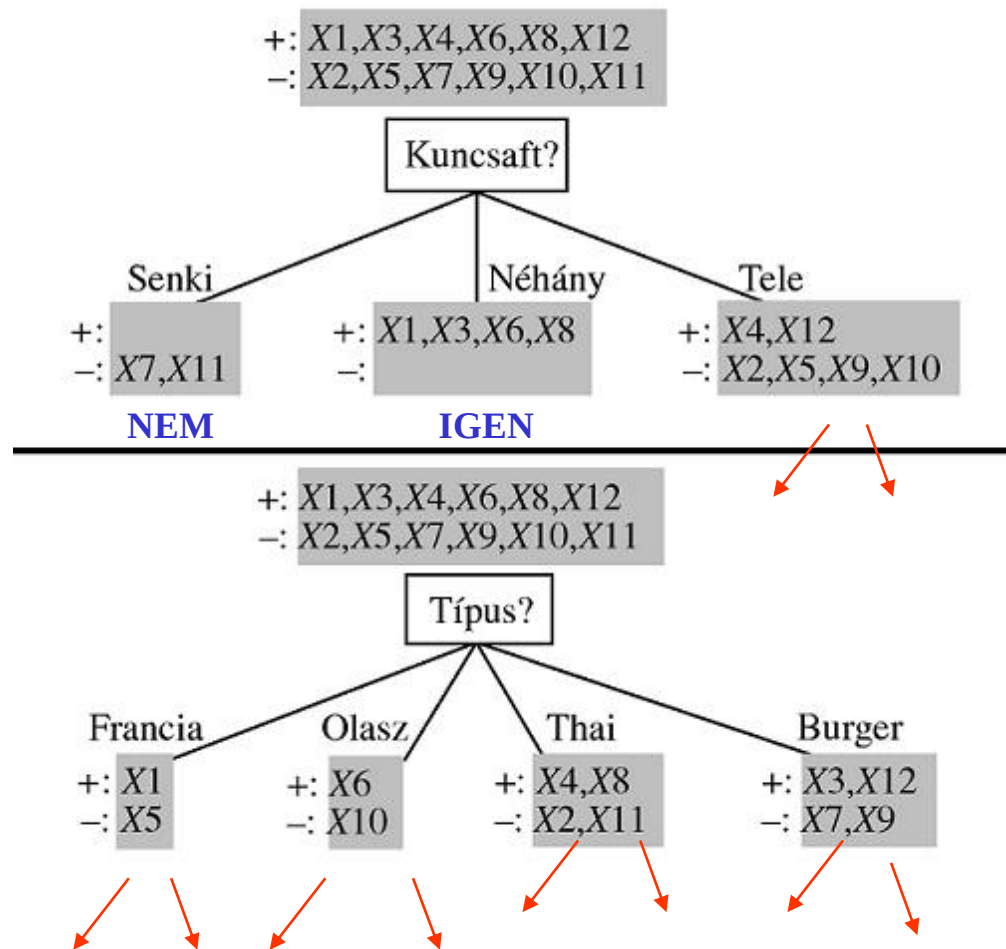
A *Kuncsaft* attribútum **nem** tökéletes, de elég jó.

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.

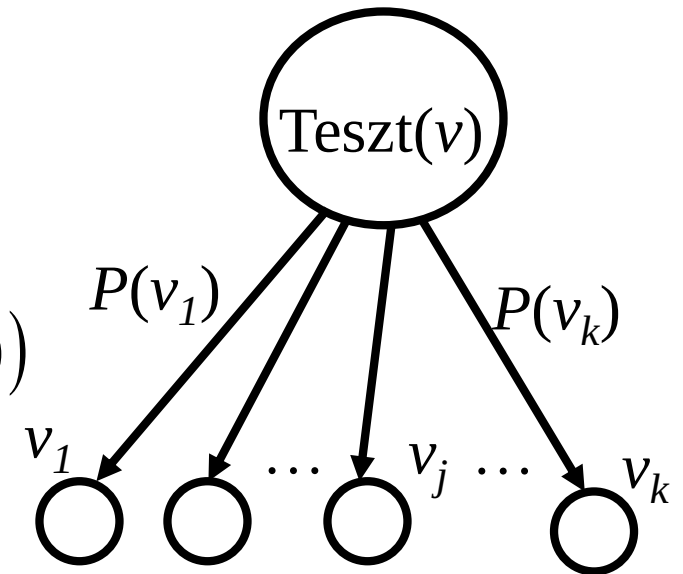
„Elég jó?" „Nagymértékben haszontalan?„

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

Egy megfelelő mérték: egy helyzetben az információsükségletünk a helyes megoldáshoz, az attribútum által adott információ várható értéke, információ átlagos tartalma, entrópia, ...

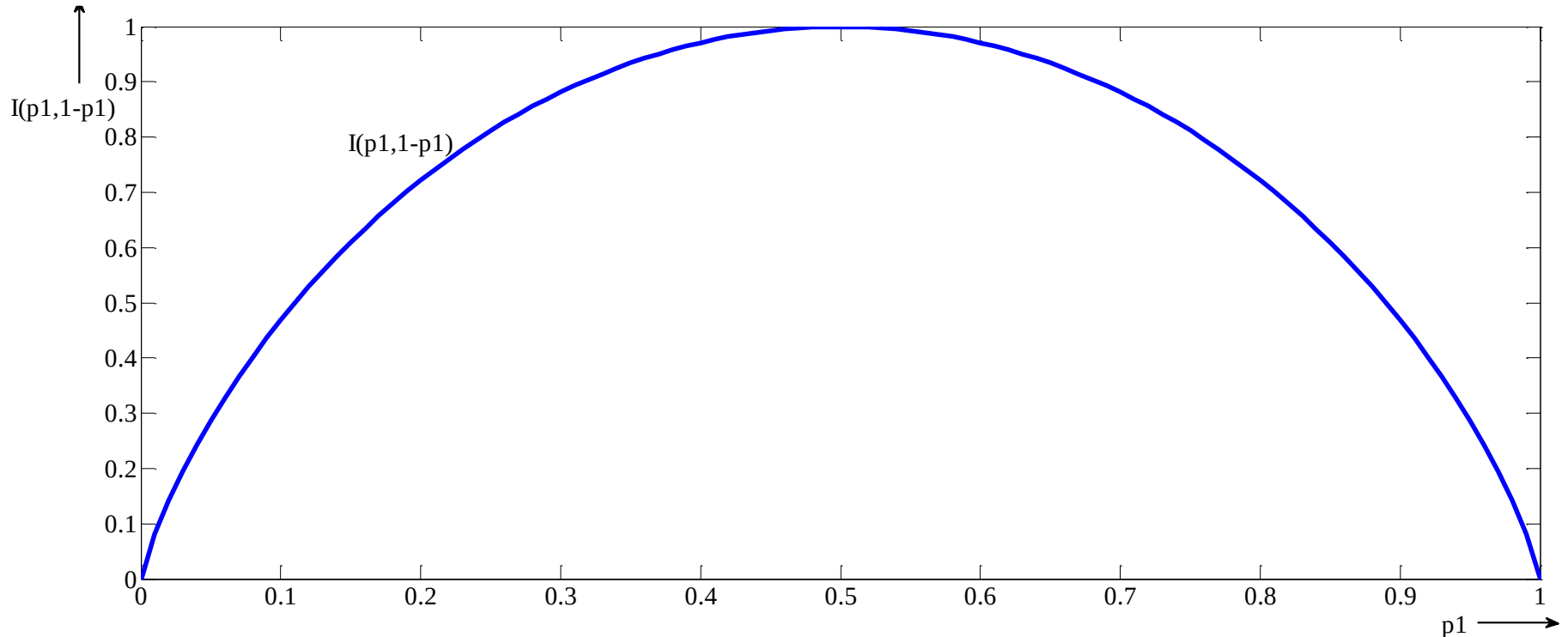
Ha a lehetséges v_j válaszok valószínűsége $P(v_j)$,
akkor az adott kiinduló halmaz esetén a jó
döntéshez szükséges információsükséglet:

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



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:

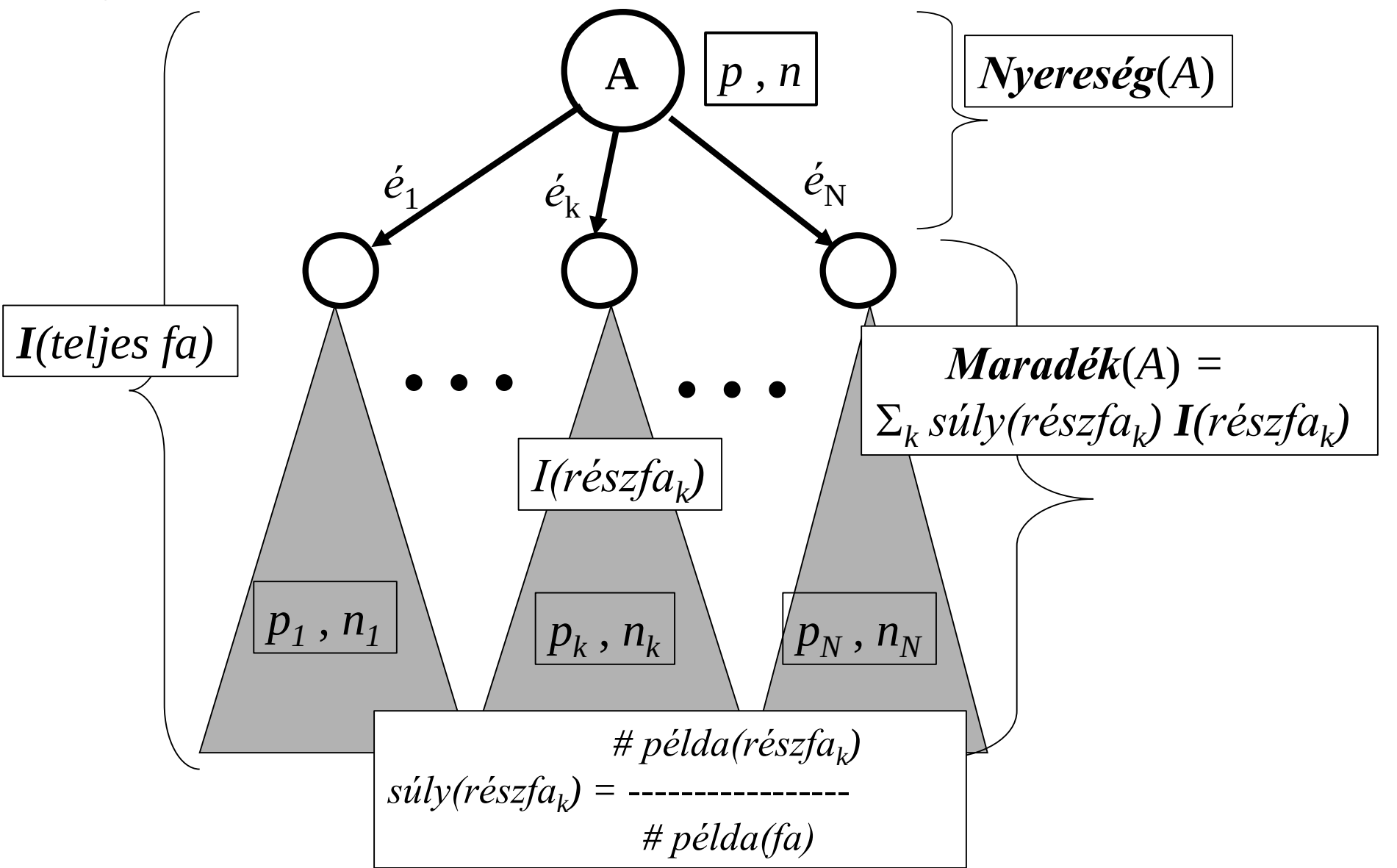
$$I(P(v_1), P(v_2)) \cong I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \cdot \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \cdot \log_2\left(\frac{n}{p+n}\right)$$

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.



Hogyan válasszunk attribútumot?

$$Maradék(A) = \sum_{k=1}^N \frac{p_k + n_k}{p + n} \cdot I \left(\frac{p_k}{p_k + n_k}, \frac{n_k}{p_k + n_k} \right)$$

Az attribútum tesztjének **információnyeresége** az **eredeti** információigény és a **teszt utáni** új információigény **különbsége**:

$$Nyereség(A) = I \left(\frac{p}{p + n}, \frac{n}{p + n} \right) - Maradék(A)$$

Válasszuk a pillanatnyilag legnagyobb nyereséget adó attribútumot!
(Mohó eljárás)

Közbevetés: mi az információszükséglet, ha 100%-os az egyik (és 0%-os a másik) eset?

$$I(0,1) = ?$$

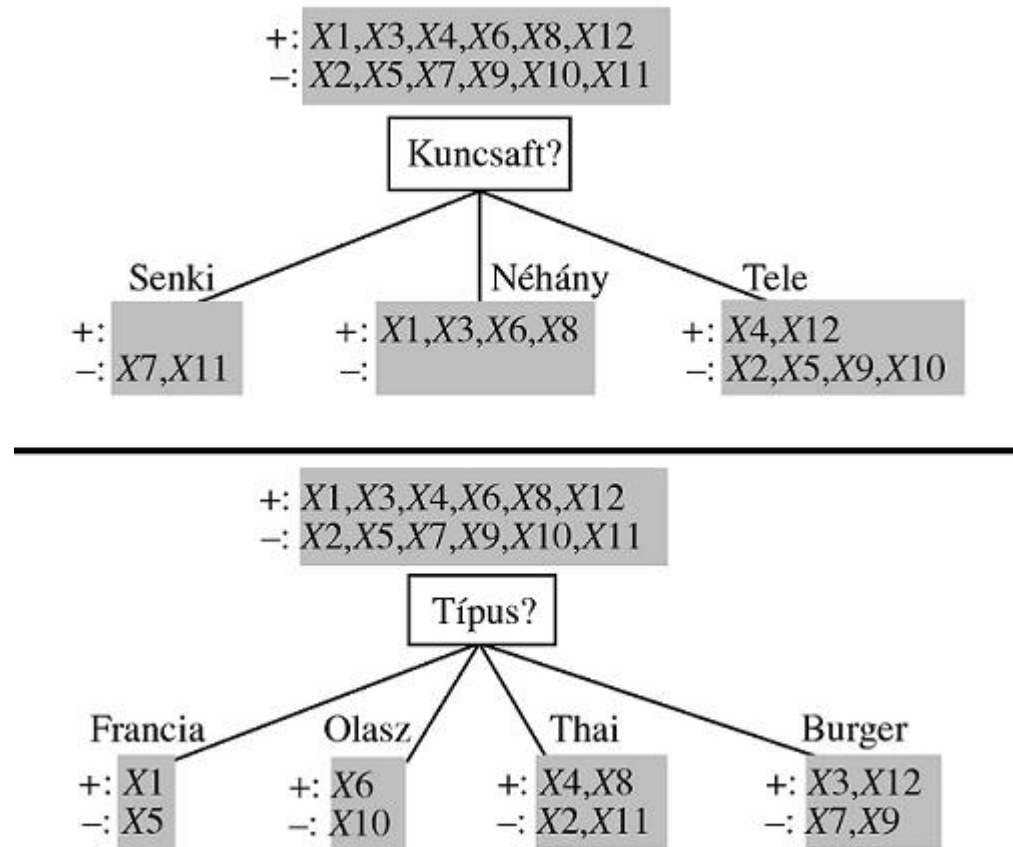
$\log_2()$ helyett használhatunk $\ln()$ -t, hiszen csak egy konstanssal való szorzásban különböznek: $\log_2(x) = \log_2(e) \cdot \ln(x)$

$$0 \cdot \ln(0) = \lim_{x \rightarrow 0} \frac{\ln(x)}{1/x}$$

helyette a deriváltak hányadosa vizsgálható (l'Hopital szabály)

$$0 \cdot \ln(0) = \lim_{x \rightarrow 0} \frac{\ln(x)}{1/x} = \lim_{x \rightarrow 0} \frac{1/x}{-1/x^2} = \lim_{x \rightarrow 0} (-x) = 0$$

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



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

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

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

(Kuncsaft = Tele) ágon még fennmaradó példák

Példák	Attribútumok									Cél
	<i>Altern</i>	<i>Bár</i>	<i>Pént</i>	<i>Éhes</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>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>>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>>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

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

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

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



Altern = Igen

$$p2 = 2, p3 = 0$$

$$n2 = 3$$



Altern = Nem

$$n3 = 1$$

6 példa

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

$$= .92 - .81 \approx .11$$

(Kuncsaft = Tele) ág

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

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

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

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

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

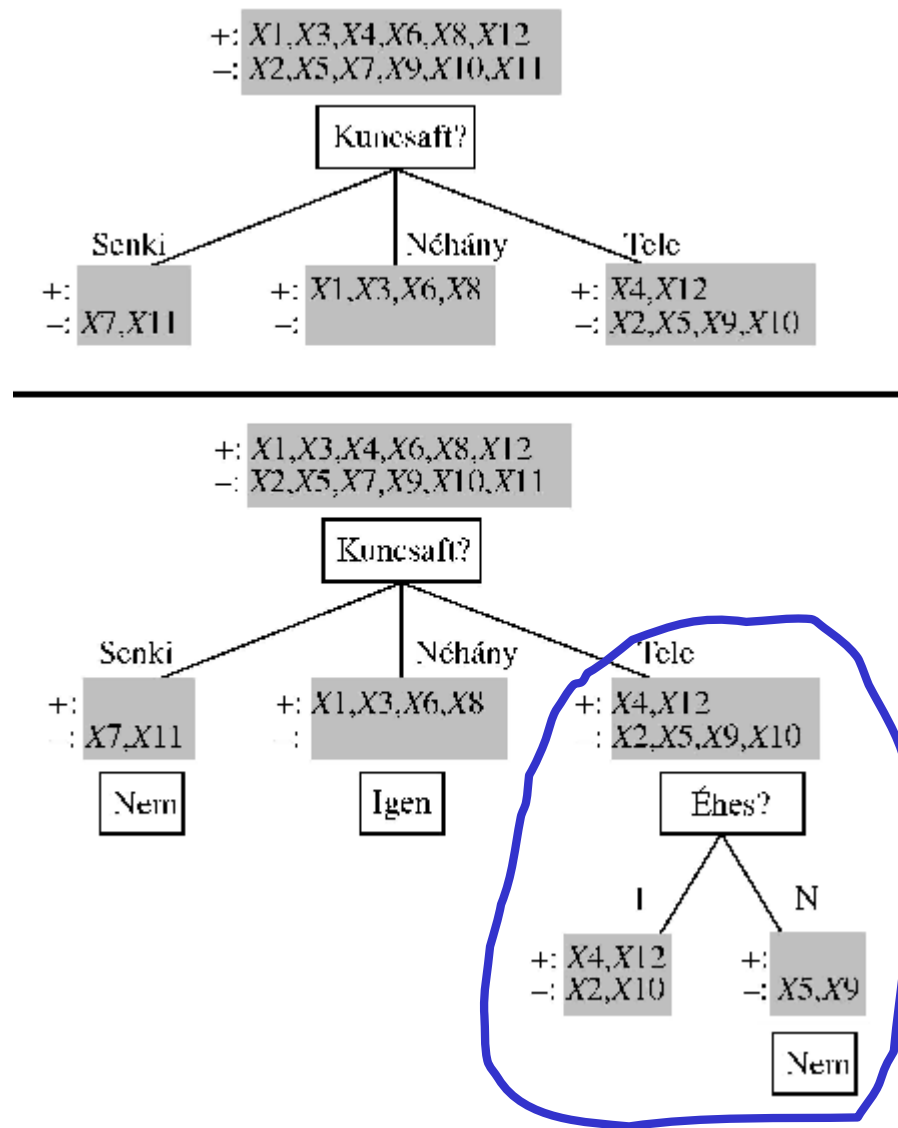
$$\text{Ny(Eső)} = I - [5/6 I(2/5, 3/5) + 1/6 I(0,1)] = .92 - .81 \approx .11$$

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

$$\begin{aligned} \text{Ny(Típus)} &= I - \\ &[2/6 I(1/2, 1/2) + 1/6 I(0,1) + 2/6 I(1/2, 1/2) + 1/6 I(0,1)] = .92 - .66 \approx .25 \end{aligned}$$

$$\begin{aligned} \text{Ny(Becs)} &= I - \\ &[2/6 I(1/2, 1/2) + 2/6 I(0,1) + 2/6 I(1/2, 1/2)] = .92 - .66 \approx .25 \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

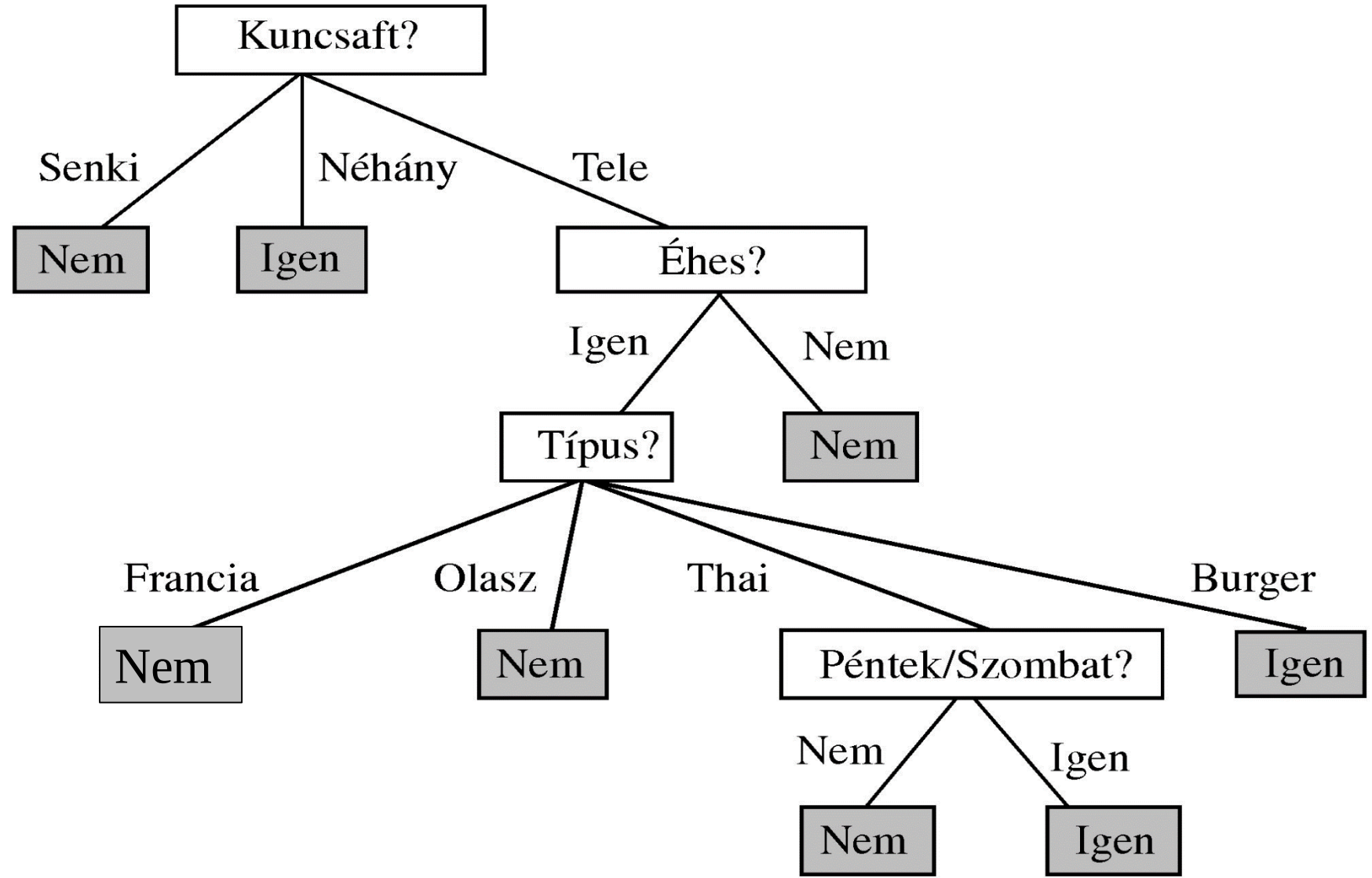
Példa	Attribútumok						Cél	
	<i>Alt</i>	<i>Bár</i>	<i>Pént/Szo</i>	<i>Ár</i>	<i>Eső</i>	<i>Fogl</i>	<i>Becs</i>	<i>VárniFog</i>
X_2	<i>Igen</i>	<i>Nem</i>	<i>Nem</i>	<i>Olcsó</i>	<i>Nem</i>	<i>Nem</i>	<i>30–60</i>	<i>Nem</i>
X_4	<i>Igen</i>	<i>Nem</i>	<i>Igen</i>	<i>Olcsó</i>	<i>Nem</i>	<i>Nem</i>	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

A többi nem jó jellemző

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

És ez ugyanúgy marad pl. a *Bár* = *Nem*, vagy *Eső* = *Nem*,
vagy *Ár* = *Olcsó*, stb. mentén (azaz ezekbe az irányokba nem
érdemes építeni a fát).

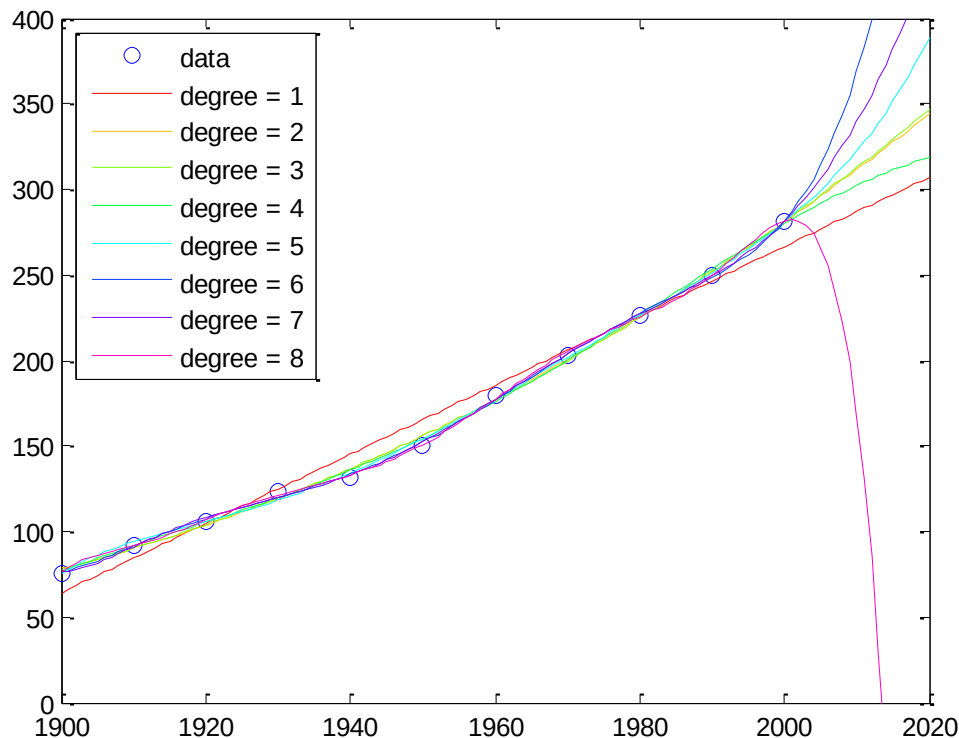
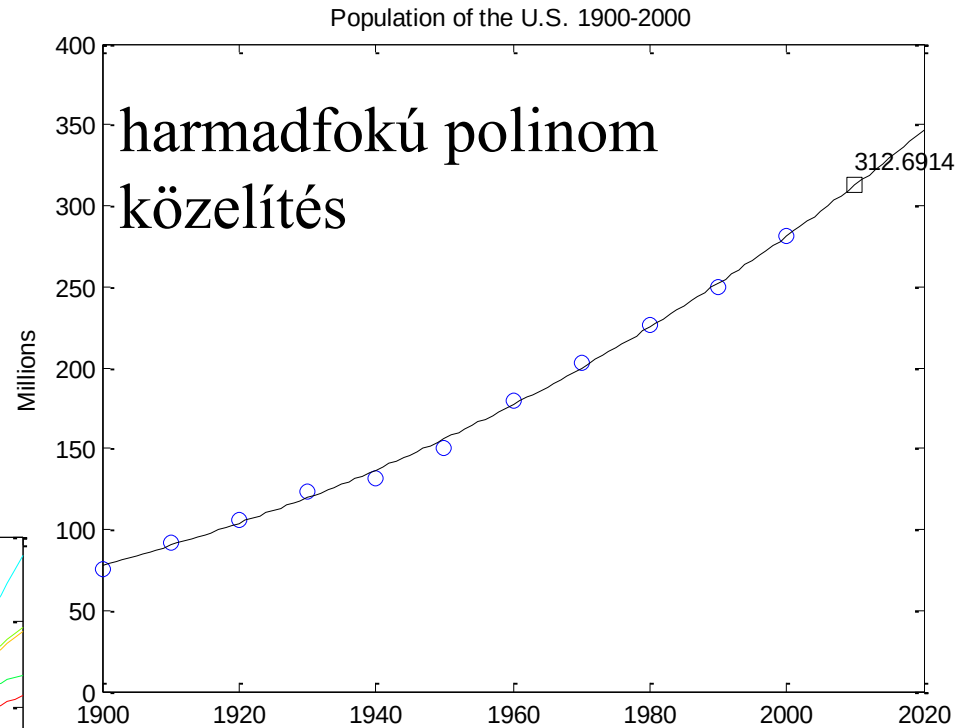
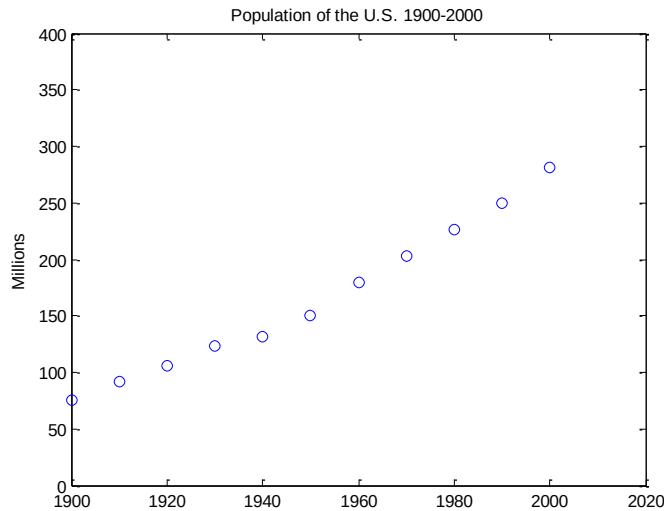
A kialakított döntési fa



Ha pl. a Pént/Szomb helyett a Becs attribútumot választjuk – ugyanúgy **konzisztens** a kialakuló fa mindegyik példánkkal, **DE máshogy** általánosít!

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

(a faék egyszerűségű gondolkodás dicsérete! ☺)



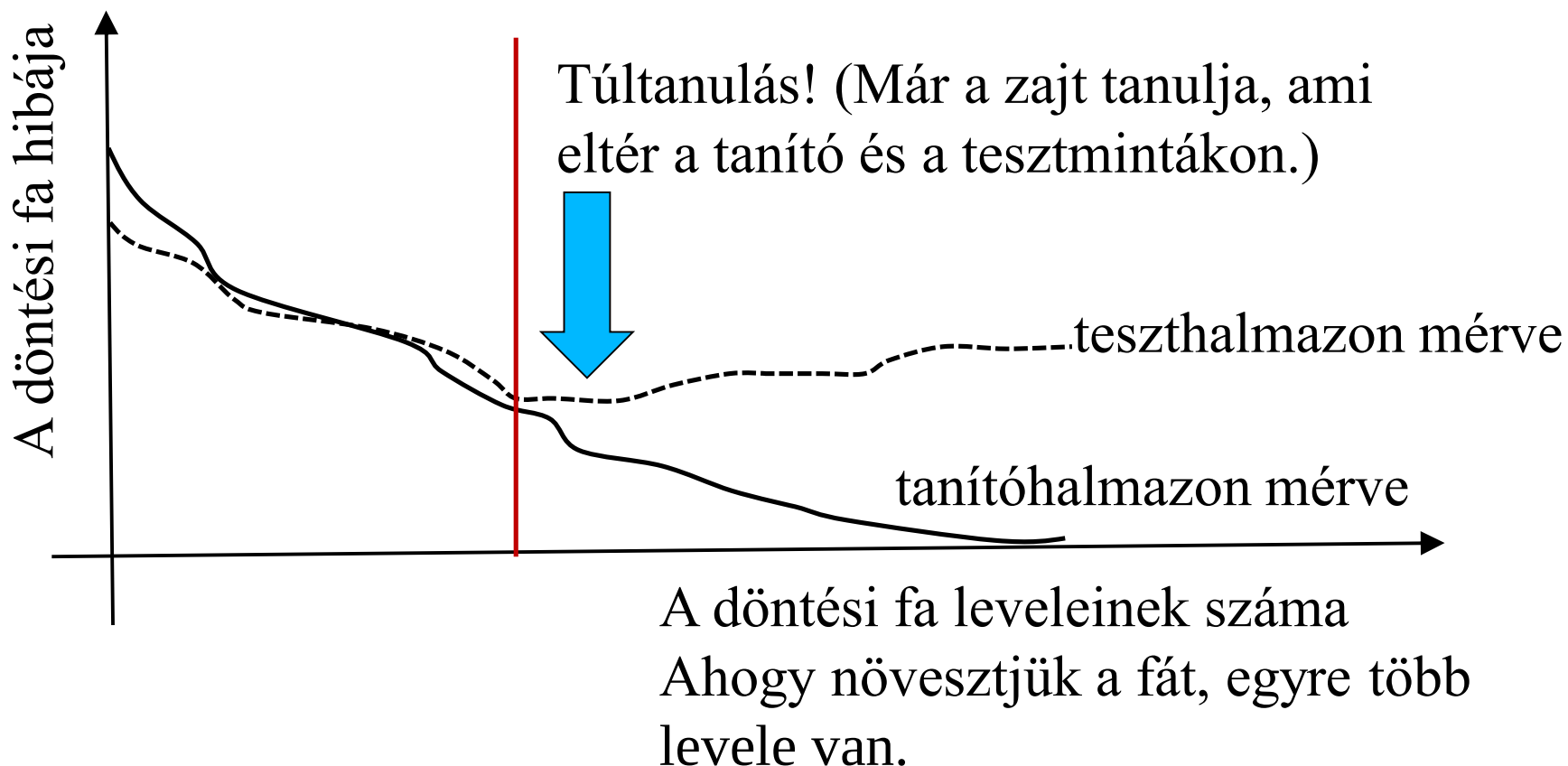
Matlab demó: az 1900-2000 közti népszámlálási adatokból becsüljük 2010-es USA népességet. (Regressziós feladat, de a jelenség a döntéseknél hasonló hatást okoz.)

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

Mikor álljunk le a döntési fa növesztésével?

Alapvetően 2 stratégia:

1. Korai leállás



(A levelekkel mérhetjük, hogy mekkorára nőtt a fa, mennyire komplex.)

Döntési fa nyesése (*pruning*):

- (1) Beállított max. mélység (mint a mélységkorlátozott keresésnél), ha elérjük, akkor ezen a szinten mindenképpen leállunk, nem fejtjük ki tovább az adott ágot.
- (2) Vizsgáljuk meg az elvégzett teszteket, ismerjük fel a nem releváns (semmitmondó) felhasznált attribútumot, és az adott ágot ott fejezzük be (metsszük vissza a fát).

Döntési fa nyesése

2. Engedjük túlnőni, majd visszavágjuk, visszanyessük

Első lehetőség

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 tartalmaznak pozitív és negatív példát, mint az eredeti halmaz.

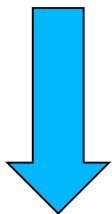
Az információ nyereség ilyenkor közel 0.

Az információ nyereség hiánya az irrelevancia jó jelzése.

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

Statisztikai tesztet végzünk: feltesszük, hogy nincs jellegzetes minta = ún. **nulla hipotézis**.

Statisztikai teszt – fordítva gondolkodunk, feltesszük, hogy nincs a halmazban jellegzetes minta = ún. **nulla hipotézis**. Utána ezt megpróbáljuk cáfolni –vagy legalább belátni, hogy nagyon valószínűtlen!



Ha a nullhipotézis igaz, akkor elvileg 0 lenne az információnyereség, az ettől való eltérés csak a véletlen műve.

Ha az eltérés **nagy** (jóval nagyobb, mint 0) $\Rightarrow$ **nem valószínű**, hogy az észlelt eltérés véletlen ingadozásból következik $\Rightarrow$ a nulla hipotézis nem valószínű: **vélhetőleg lényeges minta** van az adatokban. (Valamilyen valószínűséggel cáfolni tudjuk a nulla hipotézist.)

S :
kiinduló mintahalmaz
[p, n]
p: a pozitív, n: a negatív minták száma
S-ben

Az *a* attribútum TESZT(A)
tesztjének lehetséges értékei:
 $v_1, v_2, v_3, \dots, v_k$



S₁ :
TESZT(A)= v_1
értékekkel
rendelkező
minták halmaza,
benne a pozitív
és negatív
minták száma
 $[p_1, n_1]$

S₂ :
TESZT(A)= v_2
értékekkel
rendelkező
minták halmaza,
benne a pozitív
és negatív
minták száma
 $[p_2, n_2]$



S_k :
TESZT(A)= v_k
értékekkel
rendelkező
minták halmaza,
benne a pozitív
és negatív
minták száma
 $[p_k, n_k]$

$$\hat{p}_i = \frac{p_i + n_i}{p + n} p$$

$$\hat{n}_i = \frac{p_i + n_i}{p + n} n$$

$$D = \sum_{i=1}^k \left(\frac{(p_i - \hat{p}_i)^2}{\hat{p}_i} + \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right)$$

A D változó χ^2 eloszlást követ. Adott k (lehetséges teszteredmények száma) mellett a táblázat azt mutatja, hogy mi annak a valószínűsége, hogy a kiszámított relatív eltérés (D) eléri a megadottat, bár valójában irreleváns dolgot teszteltünk, ez csupán a véletlen műve:

SzF = $k-1$	$P_{0,5}$	$P_{0,2}$	$P_{0,1}$	$P_{0,05}$	$P_{0,01}$	$P_{0,005}$	$P_{0,001}$
1	0,46	1,64	2,71	3,84	6,64	7,88	10,83
2	1,39	3,22	4,61	5,99	9,21	10,60	13,82
3	2,37	4,64	6,25	7,82	11,35	12,84	16,27
4	3,36	5,99	7,78	9,49	13,28	14,86	18,47
5	4,35	7,29	9,24	11,07	15,09	16,75	20,52
6	5,35	8,56	10,65	12,59	16,81	18,55	22,46
7	6,35	9,80	12,02	14,07	18,48	20,28	24,32
8	7,34	11,03	13,36	15,51	20,09	21,96	26,12
9	8,34	12,24	14,68	16,92	21,67	23,59	27,88
10	9,34	13,44	15,99	18,31	23,21	25,19	29,59

Például: 400-at dobtunk pénzérmékkel és ebből 192 fej lett, 208 írás.

Kiderült, hogy

- ebből 200-at egy adott százforintossal dobtunk, ebből 88 a fej, 112 az írás,
- utána adott húszforintossal dobtunk 200-at, ebből 104 a fej, 96 az írás.

Állíthatjuk-e 99% biztonsággal, hogy az érmék különbözőek (a fej/írás kimenetel szempontjából), számít, hogy melyikkel dobunk?

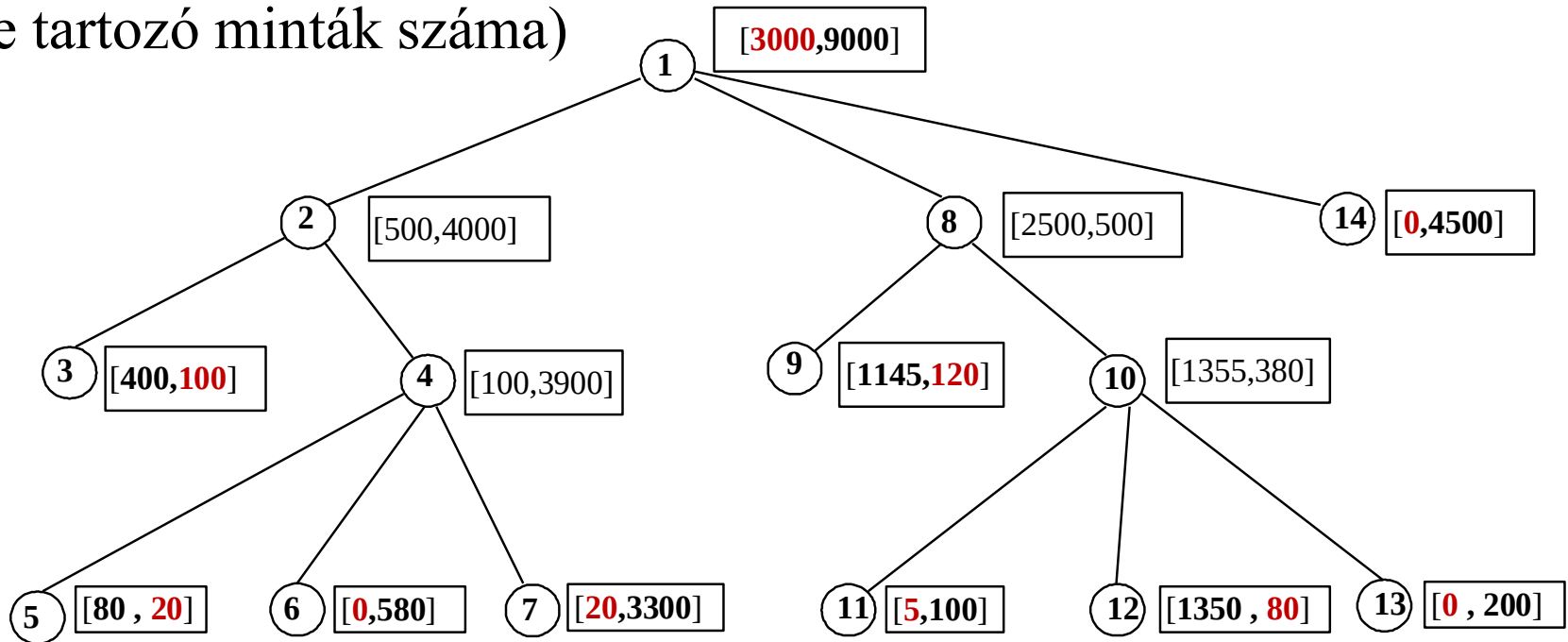
$$\Rightarrow D=2,56$$

Döntési fa nyesése: *második lehetőség*

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 gyökéknél 1, a kifejtett fánál 9

Hibaarány:

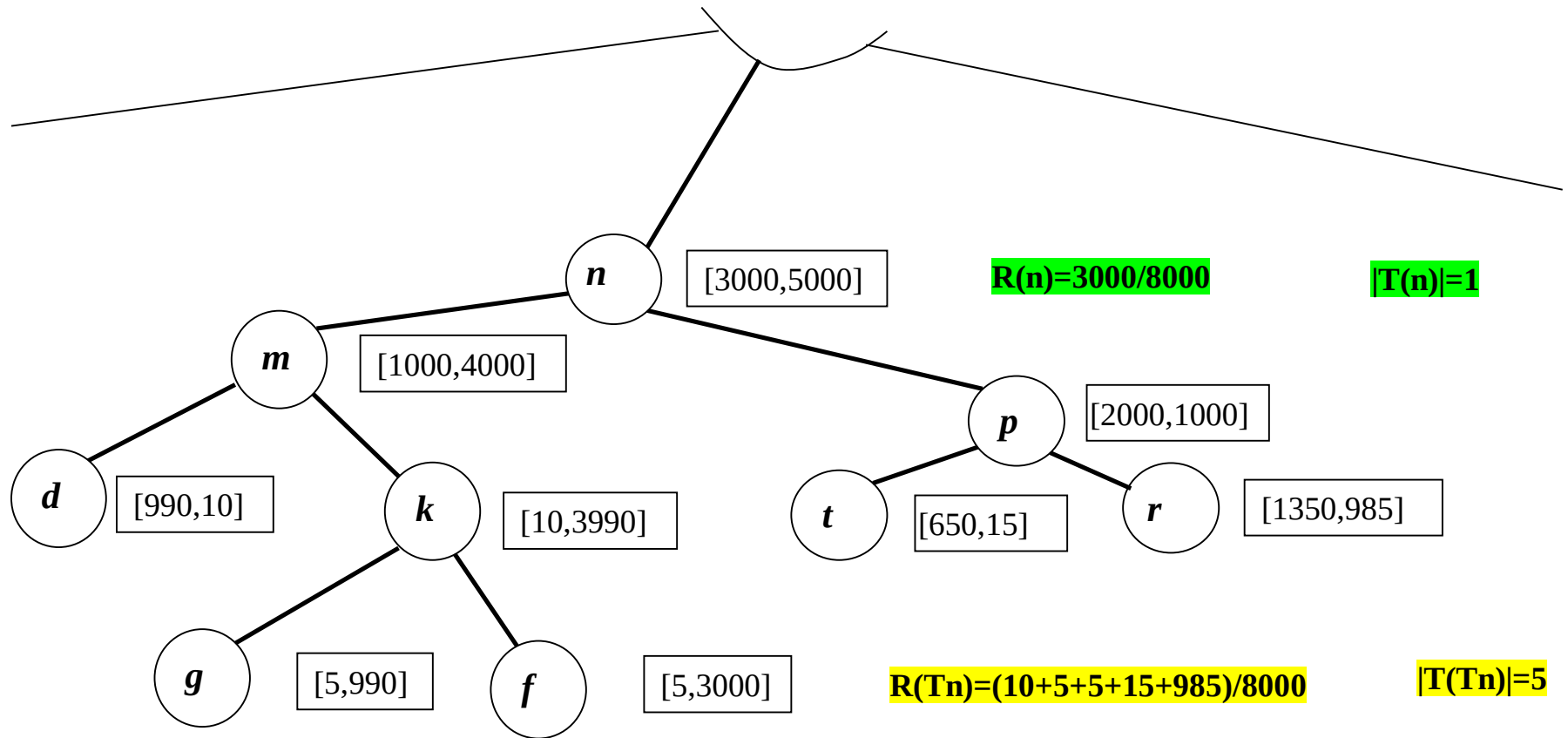
a gyökéknél 3000/12000 (hiszen a jobb válaszunk C2 lenne),

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

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



Hogyan hasonlítsuk össze a komplexitást a hibával??

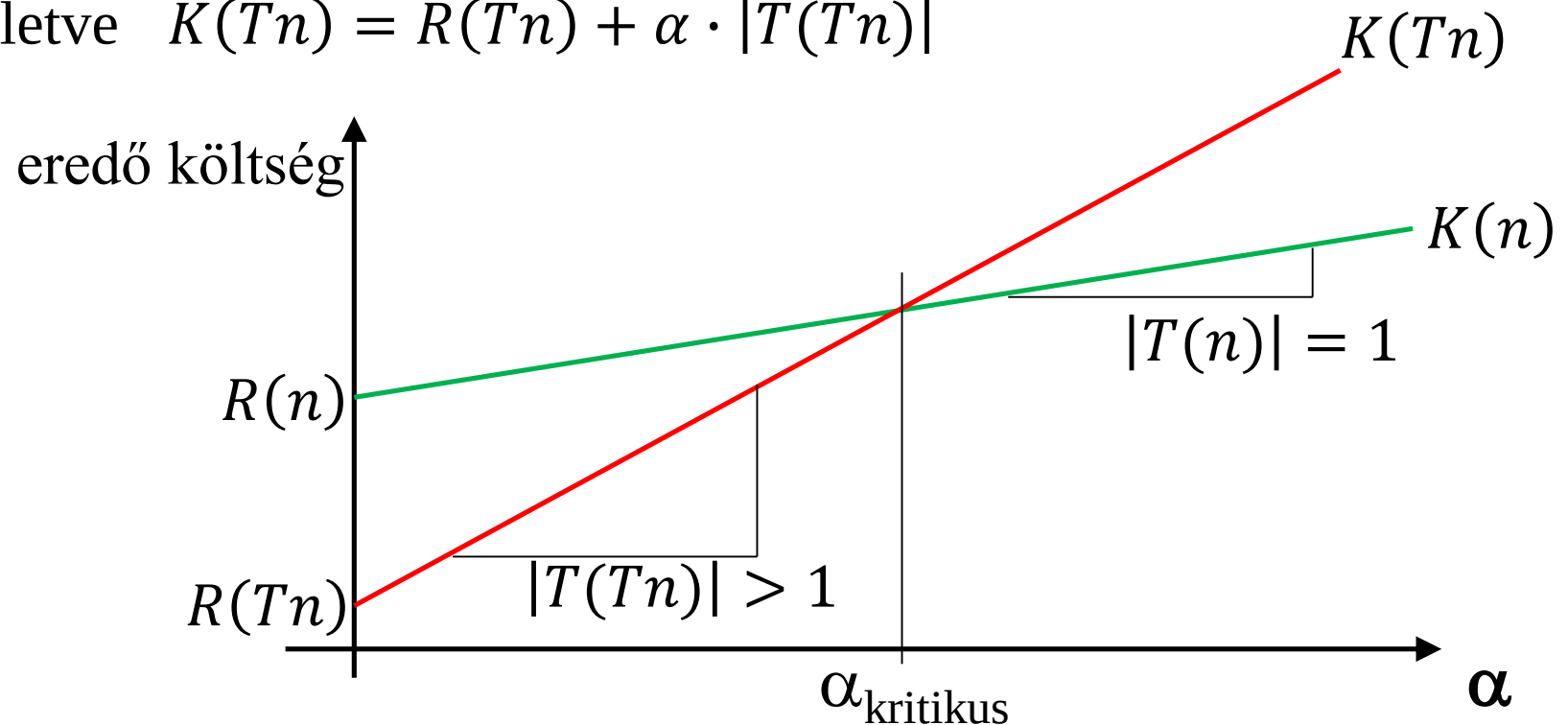
Legyen az összetett, eredő költség (pl. az n -dik csomópontra):

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

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

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

$$\text{illetve} \quad K(Tn) = R(Tn) + \alpha \cdot |T(Tn)|$$



α_{kritikus} amikor mindegy, hogy levágjuk-e a részfát, az összetett költség azonos az n -edik csomópontra, mint levélre, és az n -dik csomópontból kiinduló Tn részfára

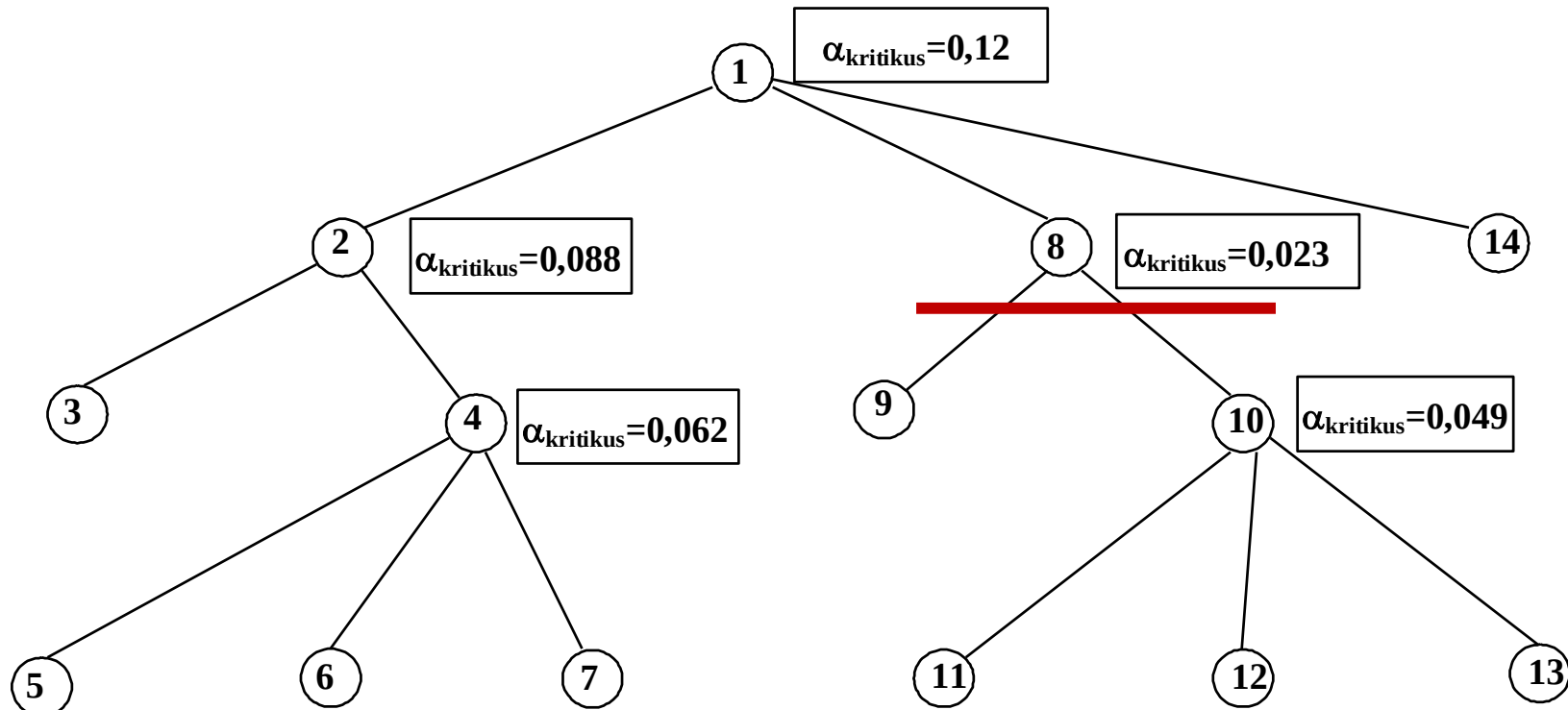
$$R(n) + \alpha_{\text{kritikus}} \cdot |T(n)| = R(Tn) + \alpha_{\text{kritikus}} \cdot |T(Tn)|$$

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

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, nem kifejezni az innen induló részfát

Gondoljuk végig, hogy mi történik, ha α -t folyamatosan növeljük 0-tól indulva!

➡ amelyik csomópont α_{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)