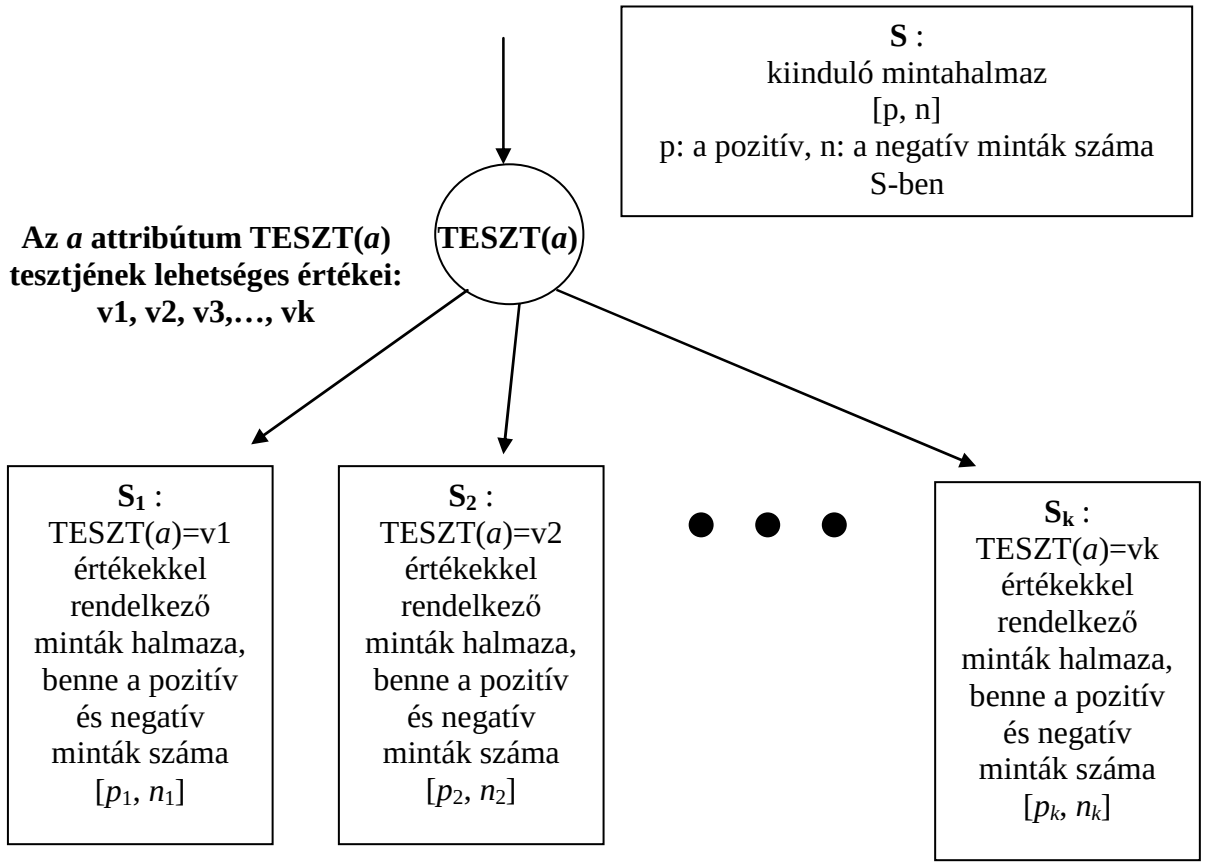
Gépi tanulás

Döntési fák szignifikancia alapú metszése

(Russel-Norvig: Mesterséges intelligencia, Modern megközelítésben, második kiadás, 764. oldal)

Ha egy irreleváns attribútumot tesztelünk, akkor azt várjuk, hogy a minták eloszlása a teszt után keletkező rész mintahalmazokban ugyanolyan lesz, mint a teszt előtti összesített mintahalmazban volt.
Ezt a hipotézisünket tudjuk az úgynevezett khi-négyzet teszttel vizsgálni. Vegyünk egy kétosztályos osztályozási példát. A minták vagy a C1 vagy a C2 osztályba tartoznak. Ezt gyakran úgy nevezzük el, hogy az egyikbe tartozó példákat pozitív példának nevezzük, a másik osztályba tartozókat pedig negatív példáknak. (Mondjuk legyenek a C1-be tartozók a pozitívak, de ez valójában mindegy.)

Tegyük fel, hogy a mintákat leíró attribútumok közül az „*a*" attribútumot teszteljük, és a tesztnek *k* különféle kimenetele (értéke) lehet. Az ábra mutatja a kialakuló döntési fa csomópontot, és a tesztet megelőző S mintahalmazt, illetve a teszt nyomán kialakuló *k* db diszjunkt minta részhalmazt.($S = \bigcup S_i$; $S_i \cap S_j = 0$ minden $i \neq j$).



Azzal az úgynevezett H_0 nullhipotézissel élünk, hogy ez az elvégzett teszt valójában irreleváns. Ennek eldöntésére azt vizsgáljuk, hogy a keletkező részhalmazokban a pozitív és negatív minták aránya eltér-e a kiinduló halmazbeli aránytól. Ha a kettő nem tér el jelentősen

egyik részhalmazban sem az eredetitől, akkor valószínűleg nem hozott érdemi újat ez az elvégzett teszt.
Ha az *i*-dik ágban keletkező részhalmaz mintái pontosan ugyanabban az arányban tartalmaznának pozitív és negatív mintákat, mint az eredeti halmazban tapasztalható arányok, akkor – mivel összesen p_i+n_i minta jutott erre az ágra – az itt található pozitív és negatív minták száma:

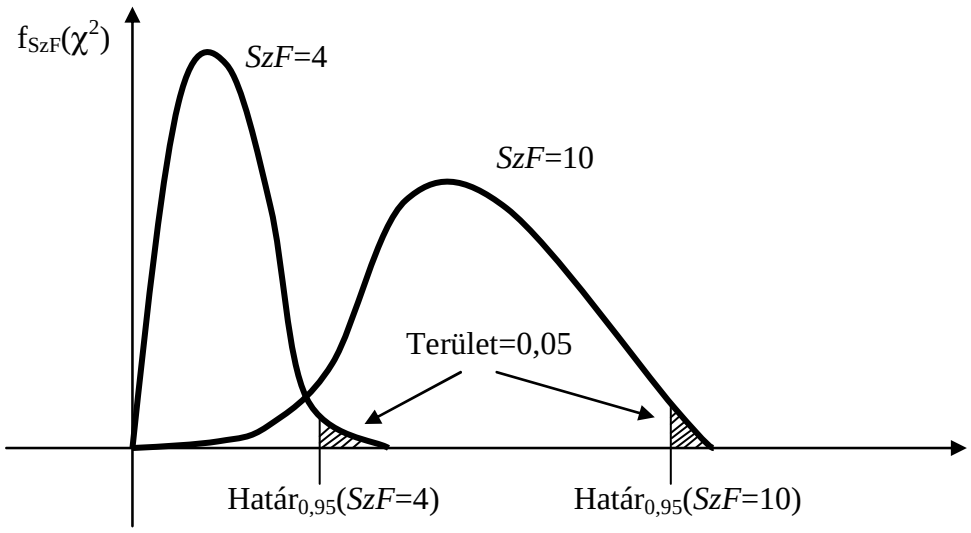
$$\hat{p}_i = \frac{p_i + n_i}{p + n} p \text{ , illetve } \hat{n}_i = \frac{p_i + n_i}{p + n} n \text{ lenne.}$$

Az összes ágban tapasztalható eltérés mértékének jellemzésére alkalmas szám:

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

Nyilván relatív eltérést kell figyelnünk, mert nem mindegy, hogy pl. 5 vagy 10000 mintaszám esetén tapasztalunk a várttól – mondjuk – 2 eltérést.
Természetesen, mivel véletlen mintáink vannak, ez sohasem lesz pontosan nulla, akkor sem, ha a teszt valóban semmi értelmes újat nem mond a problémáról. Ilyenkor *D* értéke nagy valószínűséggel a nulla környékén lesz (természetesen csak pozitív értékeket vehet fel), de kis eséllyel felvehet nagyobb értékeket is. Statisztikai vizsgálatok megmutatták, hogy a nullhipotézis feltételezésével *D* egy (*k*-1)-edfokú χ^2 (khí-négyzet) eloszlást követ.

Alábbiakban egy (kvalitatív – kézzel rajzolt) ábrán bemutatjuk a vizsgálat lényegét. Az ábrán két szabadságfok esetén – jellegre helyesen – a χ^2 (khí-négyzet) eloszlások sűrűségfüggvényét látható. A lenti ábrán pl. az SzF=4-hez tartozó sűrűségfüggvény azt mutatja, hogy egy 4 szabadságfokú χ^2 eloszlású valószínűségi változó milyen valószínűséggel vesz fel *x* és *x*+Δ*x* közti értékeket (a görbe alatti terület *x* és *x*+Δ*x* közt).
Így minél közelebb van nullához a mért értékünk, annál jobban hiszünk abban, hogy a nullhipotézisünk igaz (a TESZT nem releváns az adott osztályozási problémára). Hiszen e mellett a nullhipotézis mellett nagy valószínűséggel jönnek ki nulla környéki értékek. Viszont ha az elvileg várttól relatíve nagy eltéréseket tapasztalunk (*D* nagy), akkor kevéssé hiszünk abban, hogy a nullhipotézisünk helyes.



A vizsgált valószínűségeloszlás sűrűségfüggvényét mutatja az ábra két különböző szabadságfokra (*SzF*). A görbe alatti terület mindkét esetben 1, és a [0,*x*] feletti terület azt

fejezi ki,hogy az *x* értéket milyen valószínűséggel kaphatjuk az adott szabadsági foknál, ha a nullhipotézis teljesül. Tehát ha a felső végén 5%-nyi területet elkülönítünk, akkor a határ azt mutatja,hogy az esetek 95%-ában ezen határ alatti értékeket fogunk mérni, az 5%-ában ezen határ felettieket. Tehát, ha ezen határ feletti értéket kapunk, akkor 5%-nál kisebb az esélye,hogy a nullhipotézisünk teljesül. Ha egy 0,01 felső területet jelölnénk ki (nyilván a határ felfele tolódna), akkor az eseteknek csak 1%-ban találkozhatunk e feletti értékekkel, tehát a nullhipotézisünk 1%-nál kisebb valószínűséggel teljesül ezen határ fölötti érték mérése esetén.

χ^2 táblázat: az az érték (határ), amely fölé a megadott valószínűséggel esik a mért (számított) változó az adott szabadságfok mellett:

SzF	$\mathbf{P_{0,5}}$	$\mathbf{P_{0,2}}$	$\mathbf{P_{0,1}}$	$\mathbf{P_{0,05}}$	$\mathbf{P_{0,01}}$	$\mathbf{P_{0,005}}$	$\mathbf{P_{0,001}}$
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