UČENJE DREVES

Iščemo najboljši model v obliki drevesa. Prostor iskanja so načeloma vsa možna drevesa; velika kombinatorična zahtevnost. (U ... učna množica)

V praksi gradimo drevo T s požrešnim algoritmom:
- če vsi primeri iz U pripadajo istemu razredu R, naredi list in ga označi z R;
- sicer: izberi najbolj informativen atribut A in razdeli U na podmnožice glede na vrednosti atributa A. Rekurzivno izvedi algoritem na vsaki od pravkar generiranih podmnožic U.

Kako merimo informativnost atributa A?

Informacijski prispevek (*Information Gain*)

iz teorije informacije:
IG(Y|X) (v strojnem učenju: Y je razred, X je atribut)
Poslati moram sporočilo Y; koliko bitov v povprečju lahko prihranim, če obe strani poznata vrednost X?
IG(Y|X) = H(Y) - H(Y|X)

$$H(Y|X) = \sum_{v} p(X=v) * H(Y|X=v)$$

$$H(Y|X=v) = - \sum_{r} p(Y=r|X=v)*\log2(Y=r|X=v)$$

Oblike likov:
robota učimo koncepta "oblika". Robot ima senzorje za zaznavanje barve, pike in roba (atributi). Razred je oblika: trikotnik, kvadrat

		barva			
oblika		**rdeča**	**rumena**	**zelena**	
	trikotnik	2	0	4	**6**
	kvadrat	3	4	2	**9**
		5	**4**	**6**	**15**

H(oblika) = -6/15 log2(6/15) -9/15 log2(9/15) = 0.971

IG(barva) =H(oblika) –Ires(barva) = 0.971 – 0.690 = 0.281

Ires(barva) = 5/15 H(rdeča) + 4/15 H(rumena) + 6/15 H(zelena) =
= 1/3 * 0.971 + 4/15 *0 + 6/15 *0.918 = 0.690

H(**rdeča**) = -2/5 log2(2/5) -3/5 log2(3/5) = 0.971
H(**rumena**) = 0 log2(0) –4/4 log2(4/4) = 0
H(**zelena**) = -4/6 log2(4/6) -2/6 log2(2/6) = 0.918

H(barva) = -5/15 log2(5/15) -4/15 log2(4/15) -6/15 log2(6/15) = 1.565

RIG(barva) = IG(barva)/H(barva) = 0.281/1.565 = 0.179

		pika		
oblika		**da**	**ne**	
	trikotnik	3	3	**6**
	kvadrat	3	6	**9**
		6	**9**	**15**

H(oblika) = -6/15 log2(6/15) -9/15 log2(9/15) = 0.971

IG(pika) =H(oblika) –Ires(pika) = 0.971 – 0.951 = 0.02

Ires(pika) = 6/15 H(da) + 9/15 H(ne) =
= 2/5 * 1 + 3/5 *0.918 = 0.951

H(**da**) = -2* 1/2 log2(1/2) = 1
H(**ne**) = -1/3 log2(1/3) –2/3 log2(2/3) = 0.918

H(pika) = -6/15 log2(6/15) -9/15 log2(9/15) = 0.971

RIG(pika) = IG(pika)/H(pika) = 0.02/0.971 = 0.021

		rob		
oblika		**da**	**ne**	
	trikotnik	1	5	**6**
	kvadrat	6	3	**9**
		7	**8**	**15**

H(oblika) = -6/15 log2(6/15) -9/15 log2(9/15) = 0.971

IG(rob) =H(oblika) –Ires(rob) = 0.971 – 0.784 = 0.187

Ires(rob) = 7/15 H(da) + 8/15 H(ne) =
= 7/15 * 0.591 + 8/15 *0.954 = 0.784

H(**da**) = -1/7 log2(1/7) -6/7 log2(6/7) = 0.591
H(**ne**) = -5/8 log2(5/8) -3/8 log2(3/8) = 0.954

H(rob) = -7/15 log2(7/15) -8/15 log2(8/15) = 0.996

RIG(rob) = IG(rob)/H(rob) = 0.187/0.996 = 0.188