Klasifikavimas: sprendimų medžiai

Kontroliuojamas mokymas

Kontroliuojamo mokymo uždavinio sprendimas gali būti pavaizduotas tokia schema

$$\text{Atributai } (X_1, \ldots, X_k) \Longrightarrow \boxed{\text{Sukonstruotos taisyklės}} \Longrightarrow \text{Kintamojo } Y \text{ reikšmė}$$

Priklausomai nuo kintamojo Y tipo, skirsime klasifikavimo ir skaitinės prognozės uždavinius

Kai priklausomas kintamasis Y yra kategorinis, tada pagal jo reikšmę imties įrašas (x_i, y_i) priskiriamas klasei, kuriai $Y = y_i$. Todėl Y vadinamas klasės kintamuoju, o klasių skaičių apsprendžia jo galimų reikšmių aibės dydis.

Stuburinių klasifikacijos pavyzdys

Pavadinimas	Kraujo tipas (X_1)	Odos danga (X_2)	Gyva-vedis (X_3)	Gyvena vandenyje (X_4)	Skraido (X_5)	Turi kojas (X_6)	Klasė (Y)
žmogus	šiltas	plaukai	taip	ne	ne	taip	žinduolis
pitonas	šaltas	žvynai	ne	ne	ne	ne	roplys
lašiša	šaltas	žvynai	ne	taip	ne	ne	žuvis
banginis	šiltas	plaukai	taip	taip	ne	ne	žinduolis
varlė	šaltas	nėra	ne	kartais	ne	taip	varliagyvis
komodo varanas	šaltas	žvynai	ne	ne	ne	taip	roplys
šikšnosparnis	šiltas	plaukai	taip	ne	taip	taip	žinduolis
balandis	šiltas	plunksnos	ne	ne	taip	taip	paukštis
katė	šiltas	kailis	taip	ne	ne	taip	žinduolis
tigrinis ryklys	šaltas	žvynai	taip	taip	ne	ne	žuvis
vėžlys	šaltas	žvynai	ne	kartais	ne	taip	roplys
pingvinas	šiltas	plunksnos	ne	kartais	ne	taip	paukštis
dygliuotis	šiltas	dygliai	taip	ne	ne	taip	žinduolis
ungurys	šaltas	žvynai	ne	taip	ne	ne	žuvis
salamandra	šaltas	nėra	ne	kartais	ne	taip	varliagyvis

Klasifikavimo modelis

Apibrėžimas. *Tegul A_X yra nepriklausomų kintamųjų $X = (X_1, X_2, \ldots, X_k)$ galimų reikšmių aibė, o baigtinė aibė A_Y sudaryta iš visų galimų klasės kintamojo Y reikšmių. Klasifikavimo uždavinys yra pagal imties duomenis sukonstruoti funkciją*

$$f : \ A_X \mapsto A_Y \ .$$

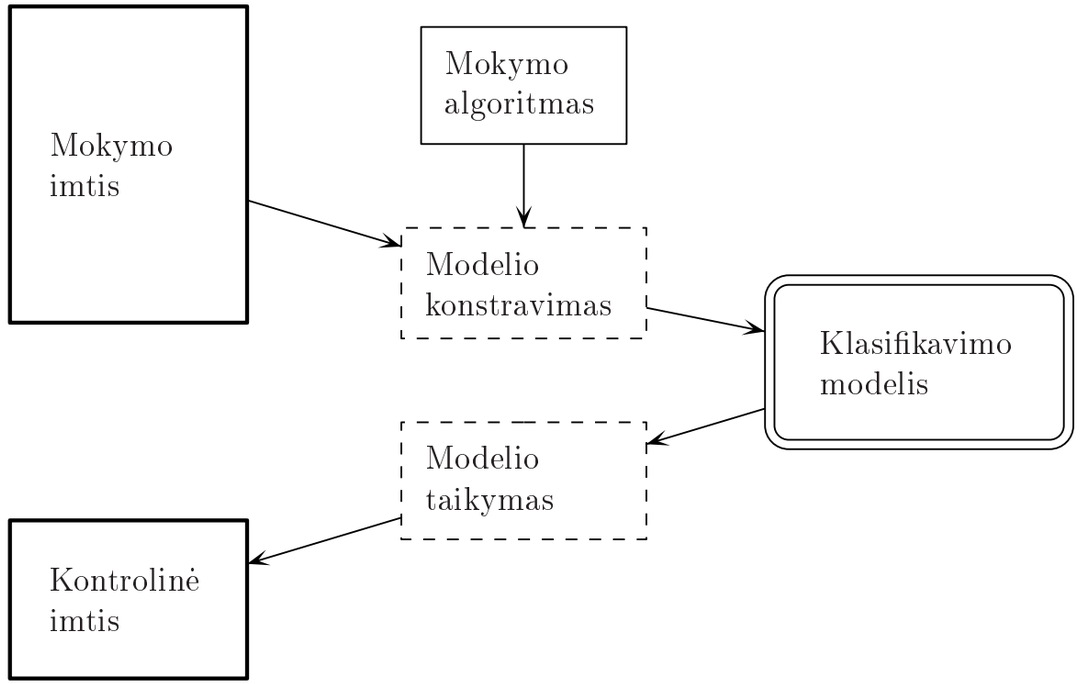
Ši funkcija vadinama klasifikavimo modeliu.

Klasifikavimo modelio pagalba sprendžiami uždaviniai

1. **Turimų duomenų klasifikavimas.** Klasifikavimo modelis naudojamas kaip priemonė, leidžianti nustatyti kriterijus, kuriais remiantis, objektas priskiriamas vienai ar kitai klasei.
2. **Nežinomų duomenų prgnozė.** Turėdami naujo objekto (imties įrašo) atributų reikšmes, klasifikavimo modelio pagalba priskiriame jį vienai iš galimų klasių.

Pavadinimas	Kraujo tipas (X_1)	Odos danga (X_2)	Gyva-vedis (X_3)	Gyvena vandenyje (X_4)	Skraido (X_5)	Turi kojas (X_6)	Klasė (Y)
šiurpusis nuodadantis	šaltas	žvynai	ne	ne	ne	taip	?

Klasifikavimo modelio konstravimo schema



Nesutapimų matrica

Kontrolinės imties binarinio klasifikavimo nesutapimų matrica

		Prognozuojama klasė	
		0	**1**
Tikroji klasė	**0**	n_{00}	n_{01}
	1	n_{10}	n_{11}

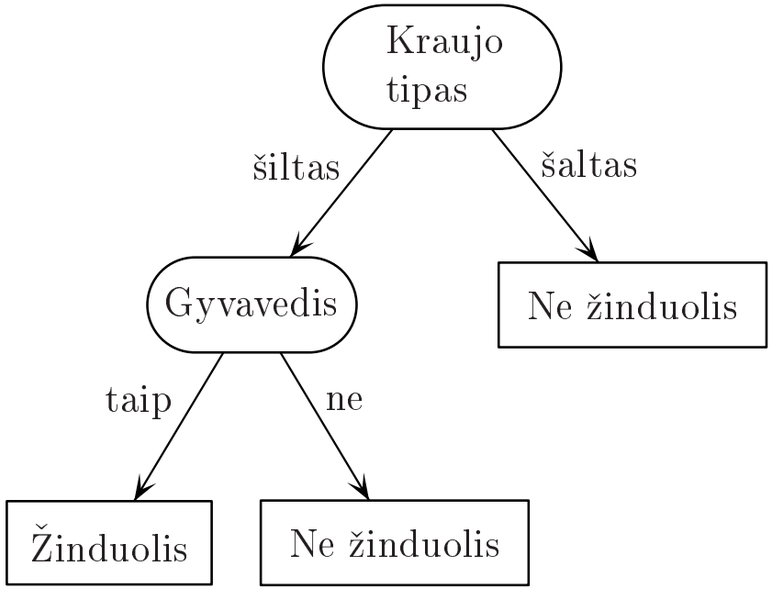
Neteisingai klasifikuotų kontrolinės imties įrašų dalis

$$e = \frac{n_{01} + n_{10}}{n_{00} + n_{01} + n_{10} + n_{11}}$$

vadinama modelio klaidos koeficientu.

Sprendimų medis

Sprendimų medis, klasifikuojantis stuburinius gyvūnus



Kaip suformuluoti gerus klausimus ?

Kaip įvertinti klasifikatoriaus patikimumą?

Hunt’o algoritmas (1966)

Tarkime D_T yra su medžio viršūne T susijusi mokymo imties įrašų aibė, o A_Y - visų klasių aibė. Kitaip sakant, A_Y yra sudaryta iš visų galimų priklausomo (klasės) kintamojo Y reikšmių. Algoritmą sudaro du žingsniai.

H1. Jei visi aibėje D_T esantys įrašai priklauso vienai klasei $y_T \in A_Y$, tai viršūnė T yra lapas, žymintis klasę y_T.

H2. Jei aibėje D_T yra įrašų, priklausančių skirtingoms klasėms, tai T tampa vidine medžio viršūne, kurios vaikams priskiriami aibės D_T poaibiai. Poaibių skaičius ir sudėtis priklauso nuo pasirenkamos atributų reikšmių tikrinimo sąlygos, t.y. nuo suformuluoto klausimo. Toliau algoritmas kartojamas kiekvienam viršūnės T vaikui.

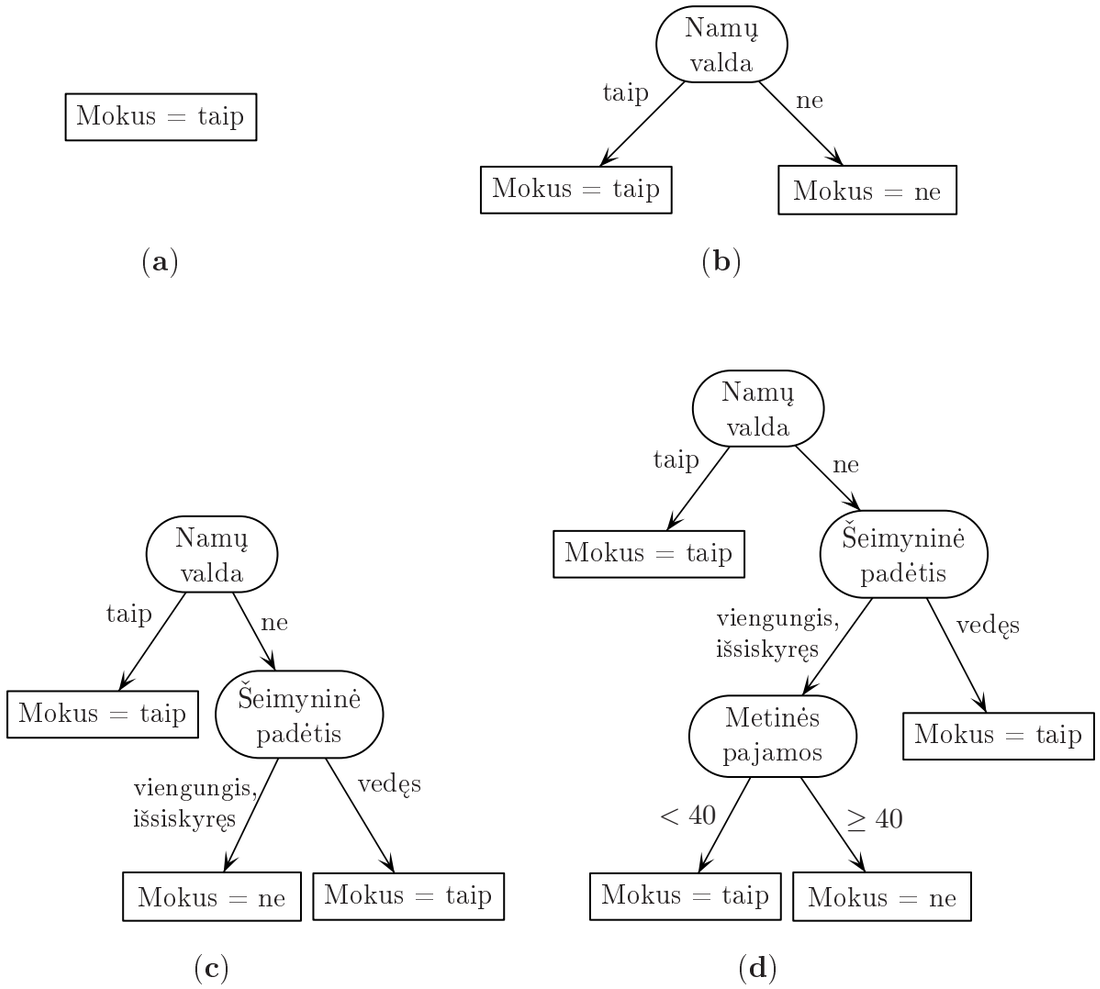
Antrąjame žingsnyje atributų tikrinimo sąlygų parinkimas nedetalizuojamas ir priklauso nuo konkrečios algoritmo modifikacijos.

Banko klientų klasifikacija

	Namų valda (X_1)	Šeimyninė padėtis(X_2)	Metinės pajamos (tūkst.Lt)(X_3)	Mokus klientas(Y)
1	taip	vedęs	65	taip
2	ne	vedęs	50	taip
3	taip	viengungis	35	taip
4	taip	vedęs	60	taip
5	ne	išsiskyręs	47	ne
6	ne	viengungis	30	taip
7	taip	išsiskyręs	110	taip
8	ne	viengungis	42	ne
9	taip	vedęs	37	taip
10	ne	viengungis	45	ne

Banko klientų klasifikacija

Hunt'o algoritmas sprendimų medžio konstrukcijai



Pastabos apie Hunt’o algoritmą

1. Kuriai nors viršūnei T, nėra ją atitinkančių įrašų, t.y. $D_T = \emptyset$. Tokiu atveju T paskelbiama lapu, atitinkančiu klasę, dažniausiai sutinkamą aibėje $D_{T'}$, čia T' žymi viršūnės T tėvą.
2. Visi aibės D_T įrašai, turi vienodas nepriklausomų kintamųjų reikšmes, bet priklauso ne vienai klasei. Tai reiškia, kad H2 žingsnyje situacija pradėtų kartotis. Tokiu atveju T paskelbiama lapu, atitinkančiu klasę, dažniausiai sutinkamą aibėje D_T.
3. Kaip skaidyti mokymo imtį? Turi būti parinktas objektyvus ir pagrįstas skaidinio "gerumo matas", kiekviename žingsnyje leidžiantis pasirinkti geriausią skaidinį.
4. Kada sustoti? Pagal Hunt’o algoritmą, viršūnė nebeskaidoma tik kai visi jos įrašai priklauso vienai klasei arba turi vienodas atributų reikšmes. Tačiau, labai "šakotas" medis ne visada yra geras. Todėl reikalingi kriterijai, leidžiantys anksčiau sustabdyti medžio auginimo procesą.

Skaidymo būdai

Pagrindinis sprendimų medžio konstravimo algoritmo elementas yra mokymo imties įrašų skaidymas į dalis, priklausomai nuo pasirinktų atributų galimų reikšmių. Todėl tiek pčios sąlygos formulavimas, tiek galimi atsakymai (o tuo pačiu ir skaidomos viršūnės vaikų skaičius) priklauso nuo atributų tipo.

Binariniai kintamieji. Tai paprasčiausias kintamasis, kartais dar vadinamas "taip-ne" atributu. Medžio viršūnę skaidant tokio kintamojo atžvilgiu, visada gaunami du vaikai, vaizduojantys du galimus atsakymus.

Skaidymas pagal vardinius kintamuosius

Jei vardinio kintamojo galimų reikšmių skaičius yra k, tai jo atžvilgiu turimą įrašų aibę galima suskaidyti $B_k - 1$ būdų. Čia B_k yra kombinatorikoje žinomi Belo skaičiai, nesunkiai randami iš rekurentinio sąryšio

$$B_{k+1} = \sum_{m=0}^{k} \binom{k}{m} B_m \,, \quad B_0 = 1 \,.$$

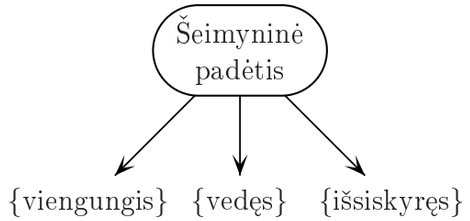
Iš viso turėsime $2^{k-1} - 1$ binarinių skaidinių, o likusieji bus daugianariai.

Pastaba. Kai $k = 1, 2, ..., 10,$ Belo skaičiai B_k yra

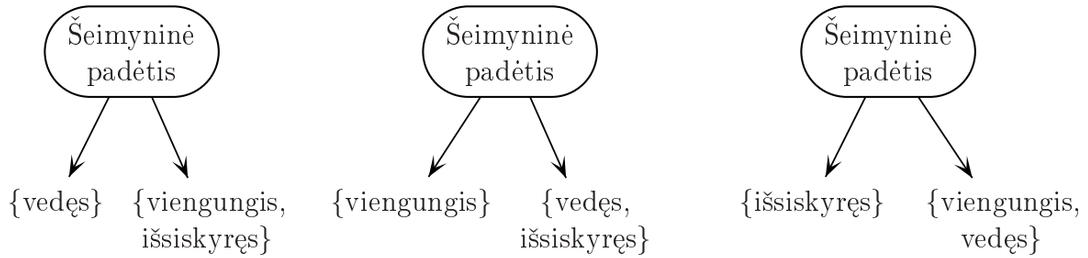
$$1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975.$$

Skaidymas pagal vardinius kintamuosius

Vardinis kintamasis X_2 (`Šeimyninė padėtis`) įgyja $k = 3$ skirtingas reikšmes. Todėl X_2 atžvilgiu atitinkamą medžio viršūnę galima išskaidyti 4 būdais: vienas skaidinys yra daugianaris, o kiti 3 - binariniai.



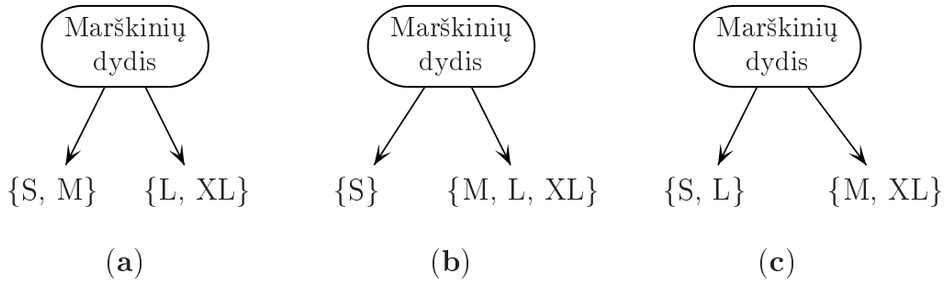
(**a**) Daugianaris skaidinys



(**b**) Binariniai skaidiniai

Skaidymas pagal ranginius kintamuosius

Panašiai kaip ir vardiniai kintamieji, ranginiai atributai gali generuoti tiek binarinius tiek daugianarius skaidinius. Tik šiuo atveju, grupuojant reikšmes, dažniausiai atsižvelgiama į jų natūralią tvarką.



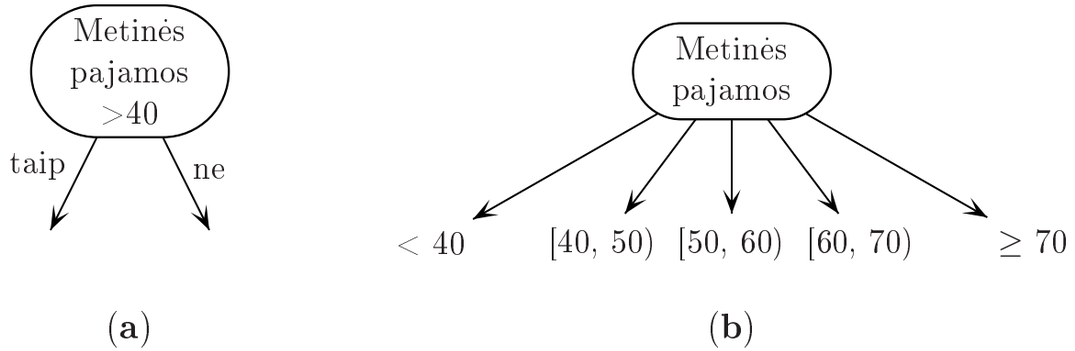
Paveiksle pavaizduoti trys galimi imties įrašų grupavimo būdai ranginio kintamojo `Marškinių dydis` atžvilgiu. Natūrali tokio atributo reikšmių tvarka yra `S, M, L, XL`. Kaip matome, tik grupavimas (c) šią tvarką pažeidžia.

Skaidymas pagal tolydžiuosius kintamuosius

Mokymo imties įrašai grupuojami tolydžiojo kintamojo X atžvilgiu, prieš tai jį diskretizavus.

Jei reikalingas binarinis skaidinys, galima paprasčiausiai tikrinti sąlygas $X < x$ arba $X \geq x$, tinkamai parinkus x.

Paveiksle pavaizduoti du skaidiniai kintamojo X_3 (`Metinės pajamos`) atžvilgiu



(a) (b)

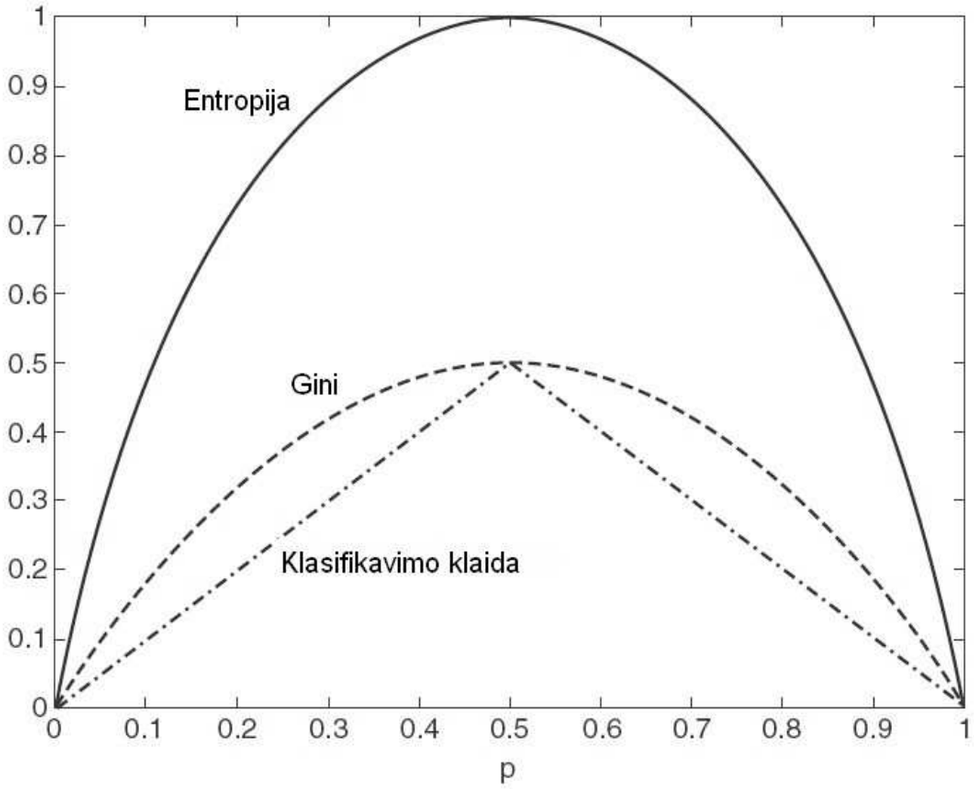
Skaidinių įverčiai

Lyginami neapibrėžtumo pokyčiai iki ir po padalinimo. Geriausiu pripažįstamas tas skaidinys, kuris labiausiai sumažina neapibrėžtumą.
Tegul T yra sprendimų medžio viršūnė, o galimų klasių aibė $A_Y = \{y_0, y_1, \ldots, y_{c-1}\}$. Žymėsime $P_T(i) = P_T(Y = y_i)$ - klasės y_i santykinį dažnį viršūnę T atitinkančioje įrašų aibėje D_T. Neapibrėžtumą viršūnėje T apibūdina dydžiai: entropija $H_T(Y)$, Gini indeksas $G_T(Y)$ ir klasifikavimo klaida $E_T(Y)$:

$$
\begin{aligned}
H_T(Y) &= -\sum_{i=0}^{c-1} P_T(i) \log_2 P_T(i)\,, \\
G_T(Y) &= 1 - \sum_{i=0}^{c-1} P_T^2(i)\,, \\
E_T(Y) &= 1 - \max_{0 \le i \le c-1} P_T(i)\,.
\end{aligned}
$$

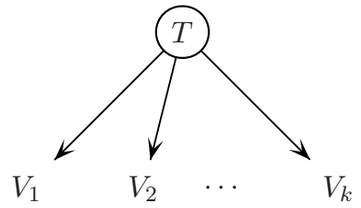
Binarinių skaidinių įverčiai

Sprendžiant binarinio klasifikavimo uždavinį, entropija,Gini indeksas ir klasifikavimo klaida yra ekvivalentūs neapibrėžtumo matai. Šiuo atveju tai yra kintamojo $p = P_T(0) = 1 - P_T(1)$ funkcijos.



Neapibrėžtumo pokytis

Tarkime, kad viršūnė T skaidoma kurio nors kintamojo X atžvilgiu ir įgyja k vaikų $V_1, V_2, \ldots, V_k$.



Viršūnę T atitinkančių įrašų skaičių žymėsime $N(T)$. Aišku, kad $N(T) = N(V_1) + N(V_2) + \cdots + N(V_k)$. Pažymėkime $F_V(Y)$ pasirinktąjį neapibrėžtumo matą viršūnėje V. Tada vidutinis neapibrėžtumo pokytis bus matuojamas dydžiu

$$I_T(Y, X) = F_T(Y) - \sum_{j=1}^{k} \frac{N(V_j)}{N(T)} F_{V_j}(Y) .$$

Kai $F_T(Y) = H_T(Y)$, gauname tarpusavio informaciją.

Pvz.:Banko klientai

Kl.kodas	X_1	X_2	X_3	Y
1	1	2	65	1
2	0	2	50	1
3	1	1	35	1
4	1	2	60	1
5	0	3	47	0
6	0	1	30	1
7	1	3	110	1
8	0	1	42	0
9	1	2	37	1
10	0	1	45	0

Konstruosime binarinį sprendimų medį pagal Gini indeksą. Kadangi iš 10 įrašų tik 3 priklauso nulinei klasei ($Y = 0$), tai šaknies T Gini indeksas

$$G_T(Y) = 1 - 0,3^2 - 0,7^2 = 0,42\,.$$

Pvz.:Banko klientai

Binarinius skaidinius pagal kintamuosius X_1 ir X_2 atitinkantys skaičiavimai

	X_1	
	0	1
$Y=0$	3	0
$Y=1$	2	5
$G_V(Y)$	0,48	0
$G_T(Y\|X_1)$	0,24	
$I_T(Y,X_1)$	0,18	

	X_2	
	{1, 2}	{3}
$Y=0$	2	1
$Y=1$	6	1
$G_V(Y)$	0,375	0,5
$G_T(Y\|X_2)$	0,4	
$I_T(Y,X_2)$	0,02	

	X_2	
	{1, 3}	{2}
$Y=0$	3	0
$Y=1$	3	4
$G_V(Y)$	0,5	0
$G_T(Y\|X_2)$	0,3	
$I_T(Y,X_2)$	0,12	

	X_2	
	{2, 3}	{1}
$Y=0$	1	2
$Y=1$	5	2
$G_V(Y)$	0,27778	0,5
$G_T(Y\|X_2)$	0,366666667	
$I_T(Y,X_2)$	0,053333333	

Pvz.:Banko klientai

Geriausią binarinį skaidinį tolydžiojo kintamojo X_3 atžvilgiu rasime tikrindami sąlygą $X_3 < x$. Dalinimo tašką x rasime imdami tarpinius taškus tarp dviejų gretimų X_3 reikšmių.

Y	1	1	1	0	0	0	1	1	1	1
X_3	30	35	37	42	45	47	50	60	65	110

x	25		32		36		40		43		46		48		55		62		90		120	
	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$	<	$\geq$
Y = 0	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
Y = 1	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
$G_V(Y)$	0	0,42	0	0,44	0	0,47	0	0,49	0,38	0,44	0,48	0,32	0,5	0	0,49	0	0,47	0	0,44	0	0,42	0
$G_T(Y\|X_3)$	0,42		0,4		0,375		0,34286		0,41667		0,4		**0,3**		0,34286		0,375		0,4		0,42	
$I_T(Y,X_3)$	0		0,02		0,045		0,07714		0,00333		0,02		**0,12**		0,07714		0,045		0,02		0	

Taigi, binarinį sprendimų medį reikia pradėti konstruoti nuo skaidinio pagal X_1, nes

$$I_T(Y, X_1) > \max I_T(Y, X_2) = \max I_T(Y, X_3)\,.$$

Santykinė tarpusavio informacija

Kai kurie algoritmai, pavyzdžiui C4.5, neapibrėžtumo pokyčiui matuoti vietoje tarpusavio informacijos naudoja santykinę tarpusavio informaciją

$$\tilde{I}_T(Y,X) = \frac{I_T(Y,X)}{H(V)}\,.$$

Čia $H(V)$ žymi viršūnės T skaidinio entropiją

$$H(V) = -\sum_{j=1}^{k} \frac{N(V_j)}{N(T)} \log_2 \frac{N(V_j)}{N(T)}\,.$$

Tokia modifikacija leidžia neutralizuoti per daug susmulkinto skaidinio įtaką.

Sprendimų medį konstruojančio algoritmo schema

Algoritmas rekursyviai konstruoja medį, pagal mokymo imties įrašus E ir jų atributų aibę F .

```
AuginkMedi(E, F)
 1.  if stopSąlyga(E, F) = True then
 2.    lapas = naujaViršūnė()
 3.    lapas.vardas = klasifikavimas()
 4.    return lapas
 5.  else
 6.    viršūnėT = naujaViršūnė()
 7.    viršūnėT.sąlyga= geriausiasSkaidinys(E, F)
 8.    V = {v | v yra galimas viršūnėT.salyga tikrinimo rezultatas}
 9.    for v ∈ V do
10.      E_v = {e | viršūnėT.salyga(e)= v ∧ e ∈ E}
11.      vaikas= AuginkMedi(E_v, F)
12.      medis papildomas viršūne vaikas, kurios tėvas viršūnėT,
         jas jungianti briauna (vaikas→viršūnėT) žymima v
13.    end for
14.  end if
15.  return viršūnėT
```

Sprendimų medį konstruojančio algoritmo procedūros

1. Funkcija `naujaViršūnė()` prijungia prie jau turimo medžio naują viršūnę, tarkime, *nviršūnė*. Jei tai yra lapas, jis žymi klasę *nviršūnė.vardas*. Priešingu atveju *nviršūnė.sąlyga* reiškia skaidinio, kurį vaizduoja *nviršūnė*, sąlygą.

Sprendimų medį konstruojančio algoritmo procedūros

1. Funkcija `naujaViršūnė()` prijungia prie jau turimo medžio naują viršūnę, tarkime, *nviršūnė*. Jei tai yra lapas, jis žymi klasę *nviršūnė.vardas*. Priešingu atveju *nviršūnė.sąlyga* reiškia skaidinio, kurį vaizduoja *nviršūnė*, sąlygą.
2. Funkcija `geriausiasSkaidinys()` randa sąlygą, pagal kurią turi būti skaidomi mokymo imties įrašai. Tai priklauso nuo pasirinkto neapibrėžtumo mato. Čia dažnai naudojami entropija ir Gini indeksas.

Sprendimų medį konstruojančio algoritmo procedūros

1. Funkcija `naujaViršūnė()` prijungia prie jau turimo medžio naują viršūnę, tarkime, *nviršūnė*. Jei tai yra lapas, jis žymi klasę *nviršūnė.vardas*. Priešingu atveju *nviršūnė.sąlyga* reiškia skaidinio, kurį vaizduoja *nviršūnė*, sąlygą.
2. Funkcija `geriausiasSkaidinys()` randa sąlygą, pagal kurią turi būti skaidomi mokymo imties įrašai. Tai priklauso nuo pasirinkto neapibrėžtumo mato. Čia dažnai naudojami entropija ir Gini indeksas.
3. Funkcija `klasifikavimas()` randa lapą žyminčią klasę. Dažniausiai tai yra klasė, kuriai priklauso dauguma lapą *lapas* atitinkančių mokymo imties įrašų. Kartais dažniai $P_{lapas}(i)$ dar naudojami įvertinti tikimybėms, kad viršūnei *lapas* priskirtas įrašas yra klasėje y_i.

Sprendimų medį konstruojančio algoritmo procedūros

1. Funkcija `naujaViršūnė()` prijungia prie jau turimo medžio naują viršūnę, tarkime, *nviršūnė*. Jei tai yra lapas, jis žymi klasę *nviršūnė.vardas*. Priešingu atveju *nviršūnė.sąlyga* reiškia skaidinio, kurį vaizduoja *nviršūnė*, sąlygą.
2. Funkcija `geriausiasSkaidinys()` randa sąlygą, pagal kurią turi būti skaidomi mokymo imties įrašai. Tai priklauso nuo pasirinkto neapibrėžtumo mato. Čia dažnai naudojami entropija ir Gini indeksas.
3. Funkcija `klasifikavimas()` randa lapą žyminčią klasę. Dažniausiai tai yra klasė, kuriai priklauso dauguma lapą *lapas* atitinkančių mokymo imties įrašų. Kartais dažniai $P_{lapas}(i)$ dar naudojami įvertinti tikimybėms, kad viršūnei *lapas* priskirtas įrašas yra klasėje y_i.
4. Funkcija `stopSąlyga()` naudojama medžio auginimo procesui sustabdyti. Tai reikėtų daryti, jei visi likę įrašai priklausytų vienai klasei arba turėtų vienodas atributų reikšmes. Kartais procesas stabdomas ir anksčiau. Pavyzdžiui, kai likusių įrašų skaičius pasidaro mažesnis už tam tikrą, iš anksto nustatytą ribą.