K-vidurkių algoritmas

1 Kas yra *K*-vidurkių algoritmas ir kaip jis veikia

Iki šiol nagrinėti algoritmai buvo tokie, kuriuose įtraukiami duomenys žinant jų įvestis ir išvestis, t. y. egzistuoja duomenys, kuriuos naudojant mašininio mokymo algoritmas gali mokytis, kaip nusakyti, koks tai objektas, kokia tai klasė, kokia reikšmė, pan. Tačiau egzistuoja uždavinių, kai mes turime įvesties duomenis apie objektus, tačiau nežinome, kokie tai objektai (1 pav.). Bet turime aiškų uždavinį – duomenis reikia suklasifikuoti, tarkime, į *n* skirtingų klasterių. Būtent todėl, kad nėra galimybės mokytis iš pavyzdinių duomenų, tokių uždavinių sprendimo algoritmai yra neprižiūrimojo tipo mašininio mokymo algoritmai.

Įėjimas

Išėjimas

X1	**X2**	
0,22	23	
0,35	45	
0,87	76	**?**
0,99	99	
0,67	45	
1,09	89	
…	…	
21	94	**?**
3	19	

1 pav. *Neprižiūrimojo mašininio mokymo duomenų iliustracija*

Klasterizavimas yra vienas iš labiausiai paplitusių duomenų analizės metodų, naudojamų gilesniam suvokimui (intuicijai) apie duomenų struktūrą gauti.

Tai gali būti apibrėžiama kaip duomenų susikirtimo į tam tikrus pogrupius uždavinys, tuos duomenis sugrupuojant taip, jog visi duomenys, esantys viename pogrupyje (klasteryje), būtų panašūs, tačiau lyginant su duomenimis kituose pogrupiuose labai skirtingi.

Tikslas – pagal tam tikrą panašumo įvertį (pvz., Euklido atstumą) surasti homogeniškus poaibius, kuriuose duomenys tame pačiame klasteryje yra panašūs. Kadangi toks klasterizavimas yra neprižiūrimojo tipo, nėra galimybės patikrinti.

***K*-vidurkių metodas** – neprižiūrimojo tipo duomenų panašumu gristas algoritmas, kuris duomenis bando susiskirstyti į *K* nepersidengiančių klasterių.

Nustatyti, gerai ar ne atliktas sugrupavimas, būtų galima, jeigu turėtume atsakymus, deja, šiuose uždaviniuose tokios informacijos nėra. K-vidurkių metodas labiau skirtas panašumui tarp duomenų nustatyti ir pagal tai pasidaryti tam tikras išvadas. Tam tikrais atvejais duomenys labai aiškiai skiriasi pagal konkrečias savybes, o kartais jie būna per daug panašūs ir toks sugrupavimas gali būti betikslis.

Įsivaizduokime, kad turime kažkokius nežinomus objektus (2 pav. (a)) ir norime juos suskirstyti į kelis klasterius (klases).

1 klausimas. Ar galima surūšiuoti šiuos objektus?

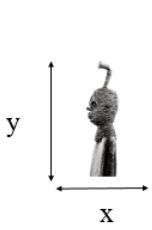
2 klausimas. Pagal kokias savybes, pagal kokius požymius juos galime grupuoti?

Vertinant vien išoriškai galima pastebėti, jog įmanomas grupavimas į tris klasterius pagal spalvą (2 pav. (c)) arba dydį, į du klasterius pagal pasisukimą (2 pav. (b)) ir pan. Galbūt yra ir savybių, kurios paveikslėlyje nėra matomos, pavyzdžiui, svoris, amžius (metais) ir pan.

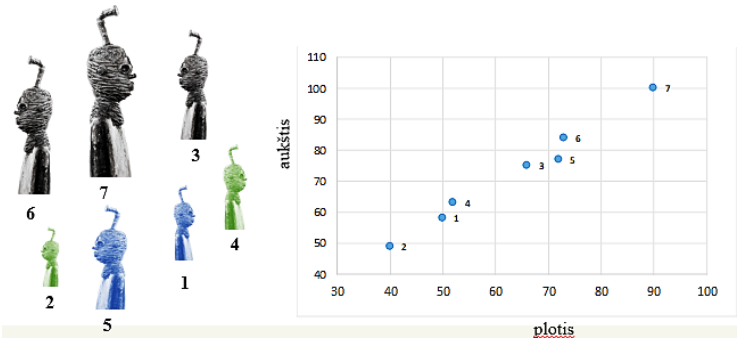
(a) (b) (c)

2 pav. *Nežinomi objektai (a) ir jų suklasterizavimas pagal pasisukimo kampą (b) į du klasterius ir spalvą (c) į tris klasterius.*

Jeigu pasirinkome, pagal kokius požymius bandysime suklasterizuoti objektus, tai reikia išgauti tų požymių skaitines reikšmes. Pateiktame pavyzdyje galima klasifikuoti objektus pagal dydį, todėl fiksuojamos *x*, *y* reikšmės, kurios žymi atitinkamai plotį ir aukštį bei atvaizduojamos grafiškai diagramoje (3 pav.).



	plotis	aukštis
i	**x**	**y**
1	50	58
2	40	49
3	66	75
4	52	63
5	72	77
6	73	84
7	90	100



3 pav. *Objektų dydžio atvaizdavimas X, Y ašyse*

Toliau visas procesas yra vykdomas remiantis K-vidurkių algoritmu, kuris susideda iš tokių 6 žingsnių:

1 žingsnis: pasirinkti klasterių skaičių K;

2 žingsnis: pasirinkti centroidus – duomenų taškus, kurių skaičius lygus K skaičiui. Šių centroido taškų parinkimas gali būti vykdomas atsitiktinai arba remiantis tam tikra taisykle;

3 žingsnis: taškui $i, i = \overline{1,n}$ paskaičiuoti atstumus iki kiekvieno centroido;

4 žingsnis: taškui *i* priskiri klasterį $m, m = \overline{1,K}$, iki kurio atstumas yra mažiausias;

5 žingsnis: perskaičiuoti klasterio m koordinates;

6 žingsnis: kartoti 3–5 žingsnius iki tol, kol nebekis taškams priskirti klasteriai bei centroidų koordinatės.

Pirmas žingsnis norint panaudoti *K*-vidurkių metodą, tai iš anksto nustatyti klasterių skaičių. *K* yra bet koks sveikas skaičius, kuris yra daugiau nei 1. Kitas žingsnis yra centroidų parinkimas. Yra skirtingų centroidų parenkamų būdų. Vienas iš tokių būdų tai labiausių nutolusių taškų parinkimas, pavyzdžiui, 3 paveikslėlyje taškai 2 ir 7.

Šios centroidų reikšmės yra užfiksuojamos ir toliau kiekvienam taškui paskaičiuojamas Euklido atstumas (gali būti iš kitas atstumo paskaičiavimo metodas). Kadangi 2 ir 7 taškai yra pradiniai centroidai, tai jiems iš karto priskiriamas klasteris, nes atstumas šiam taškui iki pačio savęs būtų nulinis.

0 iteracija: Visiems taškams išskyrus centroidus klasteriai yra nežinomi.

i	X	y	k
1	50	58	?
2	40	49	1
3	66	75	?
4	52	63	?
5	72	77	?
6	73	84	?
7	90	100	2

Klasteris	(x,y)
1	(45, 53,5)
2	(90,100)

Visiems kitiems taškams skaičiuojamas atstumas remiantis Euklido formule:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

1 iteracija: pirmajam taškui mažesnis atstumas yra iki pirmojo klasterio, todėl priskiriamas pirmasis klasteris.

Euklido atstumas su 1-uoju klasteriu	$\sqrt{(40-50)^2 + (49-58)^2} = 13{,}45$
Euklido atstumas su 2-uoju klasteriu	$\sqrt{(90-50)^2 + (100-58)^2} = 58{,}0$

2 iteracija: antrajam elementui nėra tikslo tikrinti, iki kurio klasterio atstumas yra mažesnis, nes jis pats priskirtas kaip pirmojo klasterio centroidas ir todėl atstumas iki pirmojo centroido bus 0.

Euklido atstumas su 1-uoju klasteriu	$\sqrt{(40-40)^2 + (49-49)^2} = 0$

3 iteracija: turime jau du taškus iš pirmojo klasterio. Centroidas yra vidurio taškas tarp to klasterio elementų, todėl reikia perskaičiuoti centroido koordinates, pastumiant jį per vidurį tarp šių dviejų taškų.

i	X	y	k
1	50	58	1
2	40	49	1
3	66	75	?
4	52	63	?
5	72	77	?

$$\frac{40+50}{2}, \frac{58+49}{2} = (45,\ 53{,}5)$$

Klasteris	(x,y)
1	(45, 53,5)
2	(90,100)

6	73	84	?
7	90	100	2

4 iteracija: trečiasis elementas pagal Euklido atstumą taip pat priskiriamas 1 klasteriui.

Euklido atstumas su 1-uoju klasteriu	$\sqrt{(45-66)^2+(53.5-75)^2}=30.05$
Euklido atstumas su 2-uoju klasteriu	$\sqrt{(90-66)^2+(100-75)^2}=34.65$

5 iteracija: pirmojo centroido koordinatės turi būti vėl perskaičiuojamos vertinant jau tris taškus.

i	X	y	k
1	50	58	1
2	40	49	1
3	66	75	1
4	52	63	?
5	72	77	?
6	73	84	?
7	90	100	2

$$\frac{40+50+66}{3},$$

$$\frac{58+49+75}{3}$$

$$=(56,\ 60.66)$$

Klasteris	(x,y)
1	(52, 60.66)
2	(90,100)

6 iteracija:

Euklido atstumas su 1-uoju klasteriu	$\sqrt{(52-52)^2+(60{,}66-63)^2}=2{,}33$
Euklido atstumas su 2-uoju klasteriu	$\sqrt{(90-52)^2+(100-63)^2}=53{,}03$

7 iteracija: pirmojo centroido koordinačių perskaičiavimas.

i	X	y	k
1	50	58	1
2	40	49	1
3	66	75	1
4	52	63	1
5	72	77	?
6	73	84	?
7	90	100	2

$$\frac{40+50+66+52}{4},$$

$$\frac{58+49+75+63}{4}$$

$$=(52,\ 61.5)$$

Klasteris	(x,y)
1	(52, 61,5)
2	(90,100)

8 iteracija:

Euklido atstumas su 1-uoju klasteriu	$\sqrt{(52-72)^2+(61.5-77)^2}=25{,}45$
Euklido atstumas su 2-uoju klasteriu	$\sqrt{(90-72)^2+(100-77)^2}=29{,}20$

9 iteracija:

i	x	y	k
1	50	58	1
2	40	49	1
3	66	75	1
4	52	63	1
5	72	77	1
6	73	84	?
7	90	100	2

$$\frac{50+40+66+52+72}{5},$$

$$\frac{58+49+75+63+77}{5}$$

$$=(56,\ 64.4)$$

Klasteris	(x, y)
1	(56, 64,4)
2	(90,100)

10 iteracija:

Euklido atstumas su 1-uoju klasteriu	$\sqrt{(56-73)^2+(64{,}4-84)^2}=25{,}94$
Euklido atstumas su 2-uoju klasteriu	$\sqrt{(90-73)^2+(100-84)^2}=23{,}34$

11 iteracija: antrojo centroido koordinačių perskaičiavimas.

I	X	y	k
1	50	58	1
2	40	49	1
3	66	75	1
4	52	63	1
5	72	77	1
6	73	84	2
7	90	100	2

$$\frac{73+90}{2},$$

$$\frac{84+100}{2}$$

$$=(81{,}5,\ 92)$$

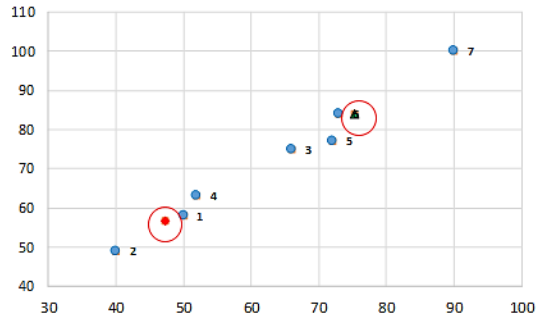
Klasteris	(x, y)
1	(56, 64,4)
2	(81,5, 92)

≥ **12 iteracija**. Visų taškų atstumų perskaičiavimas iki pakitusių centroidų koordinačių. Tada vėl perskaičiuojami centroidai ir vėl vykdomas patikrinimas. Šis procesas vyksta iki tol, kol niekas nebekinta, tai yra centroidų reikšmės ir duomenims priskirti klasteriai.

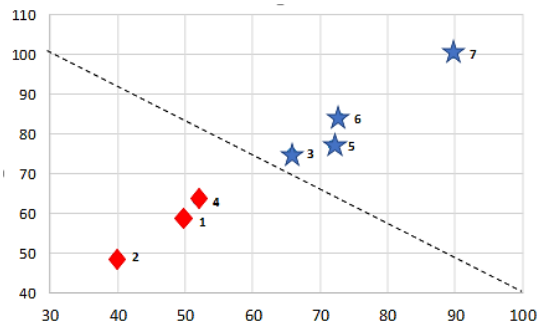
i		x	y	k
1		50	58	**1**
2		40	49	**1**
3		66	75	**2**
4		52	63	**1**
5		72	77	**2**
6		73	84	**2**
7		90	100	**2**

Klasteris	(x, y)
1	(47,33, 56,66)
2	(75,25, 84)

Iš pradinės pozicijos centroidai pasistumia į reikiamas vietas (4 pav.), ir galime teigti, kad uždavinys išspręstas naudojant K-vidurkių metodą ir duomenys išskaidyti į du klasterius (klases) (5 pav.).



4 pav. *Galutinės dviejų centroidų pozicijos*



5 pav. *Duomenų suklasterizavimo į dvi klases grafinė interpretacija*