

Demonstracija koncepta kroz jednostavne modele podataka

Metode i tehnike analize poslovnih
podataka

Prof. dr Zita Bošnjak

Primer za k-means algoritam: Insurance Fraud Detection

Olson, D., Shi, Y. (2007), Introduction to
Business Data Mining, McGraw-Hill

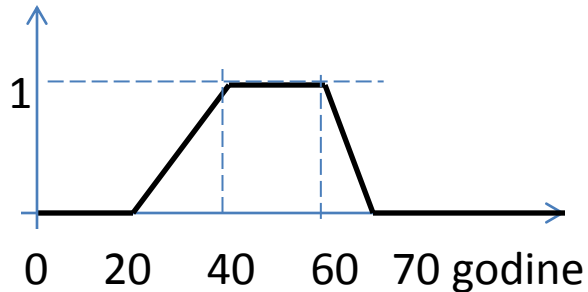
TRAINING DATA SET

Claimant		Claim		Prior		Outcome
Age	Gender	Amount	Tickets	Claims	Attorney	
52	Male	2000	0	1	Jones	ok
38	Male	1800	0	0	None	ok
21	Female	5600	1	2	Smith	Fraudulent
36	Female	3800	0	1	None	ok
19	Male	600	2	2	Adams	ok
41	Male	4200	1	2	Smith	Fraudulent
38	Male	2700	0	0	None	ok
33	Female	2500	0	1	None	Fraudulent
18	Female	1300	0	0	None	ok
26	Male	2600	2	0	None	ok

Zbog \neq raspona vrednosti, podatke NORMALIZUJEMO!

NORMALIZACIJA podataka (SKALIRANJE na [0,1])

- Godine:



Stara vr.	Nova vr.
< 20	0.0
20 - 40	$(\text{godine} - 20)/20$
40 - 60	1.0
60 - 70	$(70 - \text{godine})/10$
70 -	0.0

- Pol: Female -0; Male - 1

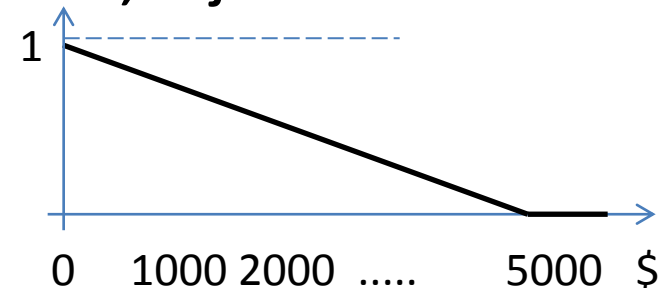
- Claim amount: nova = $\max \{1 - \text{stara}/5000, 0\}$

- Tickets: $0 \rightarrow 1$; $1 \rightarrow 0.6$; $2 \rightarrow 0$

- Prior claims: $0 \rightarrow 1$; $1 \rightarrow 0.5$; $\geq 2 \rightarrow 0$

- Attorney: None $\rightarrow 1$, ostali $\rightarrow 0$

- Outcome: OK $\rightarrow 0$; Fraudulent $\rightarrow 1$



Primer za k-means algoritam: Insurance Fraud Detection

Olson, D., Shi, Y. (2007), Introduction to
Business Data Mining, McGraw-Hill

Case	Claimant		Claim		Prior		Outcome
	Age	Gender	Amount	Tickets	Claims	Attorney	
1	1	1	0.6	1	0.5	0	0
2	0.9	1	0.64	1	1	1	0
3	0.05	0	0	0.6	0	0	1
4	0.8	0	0.24	1	0.5	1	0
5	0	1	0.88	0	0	0	0
6	1	1	0.16	0.6	0	0	1
7	0.9	1	0.46	1	1	1	0
8	0.65	0	0.5	1	0.5	1	1
9	0	0	0.74	1	1	1	0
10	0.3	1	0.48	0	1	1	0

Algoritam k-centara

1. korak: Odabiramo broj klastera: 2 (za potrebe primera)
2. korak: Odabiramo inicijalne centre: C1 i C3

Case	Claimant Age	Gender	Claim Amount	Tickets	Prior Claims	Attorney	Outcome
1	1	1	0.6	1	0.5	0	0
2	0.9	1	0.64	1	1	1	0
3	0.05	0	0	0.6	0	0	1
4	0.8	0	0.24	1	0.5	1	0
5	0	1	0.88	0	0	0	0
6	1	1	0.16	0.6	0	0	1
7	0.9	1	0.46	1	1	1	0
8	0.65	0	0.5	1	0.5	1	1
9	0	0	0.74	1	1	1	0
10	0.3	1	0.48	0	1	1	0

Algoritam k-centara

3. korak: Primere za obučavanje pridružujemo klasterima

– Euklidska distanca:

$$D1: (0.9-1)^2 + (1-1)^2 + (0.64-0.6)^2 + (1-1)^2 + (1-0.5)^2 + (1-0)^2 + (0-0)^2 = 1.2616$$

$$D2: (0.9-0.05)^2 + (1-0)^2 + (0.64-0)^2 + (1-0.6)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 = 4.2921$$

Case	Claimant		Claim		Prior		Outcome
	Age	Gender	Amount	Tickets	Claims	Attorney	
1	1	1	0.6	1	0.5	0	0
2	0.9	1	0.64	1	1	1	0
3	0.05	0	0	0.6	0	0	1
4	0.8	0	0.24	1	0.5	1	0
5	0	1	0.88	0	0	0	0
6	1	1	0.16	0.6	0	0	1
7	0.9	1	0.46	1	1	1	0
8	0.65	0	0.5	1	0.5	1	1
9	0	0	0.74	1	1	1	0
10	0.3	1	0.48	0	1	1	0

Algoritam k-centara

Case	Claimant		Claim		Prior		Outcome
	Age	Gender	Amount	Tickets	Claims	Attorney	
1	1	1	0.6	1	0.5	0	0
2	0.9	1	0.64	1	1	1	0
3	0.05	0	0	0.6	0	0	1
4	0.8	0	0.24	1	0.5	1	0
5	0	1	0.88	0	0	0	0
6	1	1	0.16	0.6	0	0	1
7	0.9	1	0.46	1	1	1	0
8	0.65	0	0.5	1	0.5	1	1
9	0	0	0.74	1	1	1	0
10	0.3	1	0.48	0	1	1	0

4. Izračunavamo nove vrednosti centara na osnovu pridruženih entiteta:

C1, C2, C6, C7, C10: $\text{Age} = (1 + 0.9 + 1 + 0.9 + 0.3) / 5 = 0.82$

C3, C4, C5, C8, C9: $\text{Age} = (0.05 + 0.8 + 0 + 0.65 + 0) / 5 = 0.3$

... po analogiji za ostale attribute.

Algoritam k-centara

5. korak: Iterativno ponavljamo korake 3-5, sve dok se ne ponovi pridruživanje iz prethodnog koraka (centri klastera ostaju nepromenjeni u narednoj iteraciji).

- Prednost algoritma: razumljivost/jednostavnost
- Deskriptivna, a ne prediktivna analiza - finalni klasteri ne moraju odgovarati ni jednom izlazu!

	Claimant Age	Gender	Claim Amount	Tickets	Prior Claims	Attorney	Outcome
Klaster 1:	0.82	1	0.468	0.72	0.7	0.6	0.2
Klaster 2:	0.3	0.2	0.472	0.72	0.4	0.6	0.4

Algoritam k-centara

Šta smo naučili kroz deskripciju?

Klaster 1:	0.82	1	0.468	0.72	0.7	0.6	0.2
Case	Claimant Age	Gender	Claim Amount	Tickets	Prior Claims	Attorney	Outcome
1	1	1	0		0.5		0
2	0.9	1	0.4		1		0
6	1	1	0.5	0.5	0		1
7	0.9	1	0.5		1		0
10	0.3	1	0.3		1		0
Klaster 2:	0.3	0.2	0.472	0.72	0.4	0.6	0.4
Case	Claimant Age	Gender	Claim Amount	Tickets	Prior Claims	Attorney	Outcome
3	0.05	0	0	0.6	0	0	1
4	0.8	0	0.24	1	0.5	1	0
5	0	1	0.88	0	0	0	0
8	0.65	0	0.5	1	0.5	1	1
9	0	0	0.74	1	1	1	0

Godine čine malu razliku: u K1 su nešto stariji nego u K2.

Pol je značajan: K1 je čisto muški, a većinu u K2 čine žene.

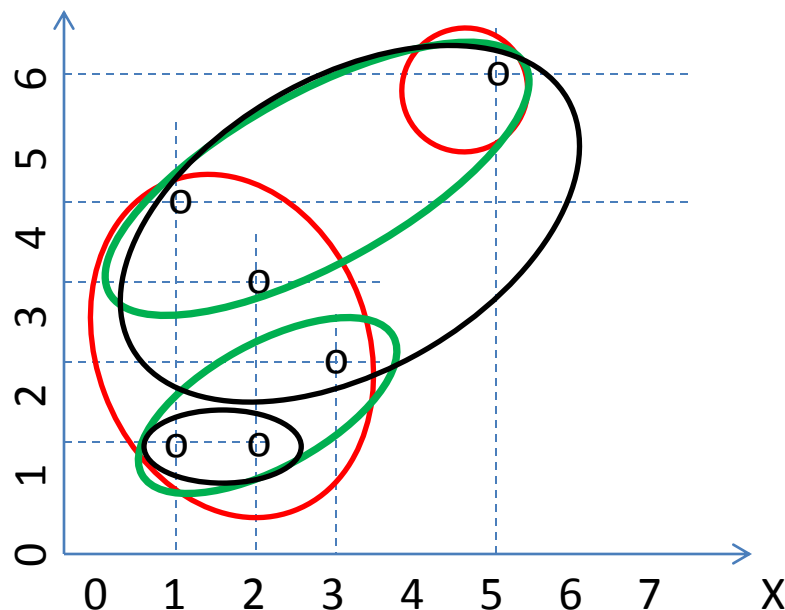
Iznos štete, br. kazni, sudija su beznačajni - isti za K1 i K2.

Prethodni zahtevi čine malu razliku.

K1 ima manje FRAUD od K2, ali je PREMALI UZORAK!!!!

Nije zagarantovan rezultat - primer

Entitet	X	Y
e1	1	1.5
e2	1	4.5
e3	2	1.5
e4	2	3.5
e5	3	2.5
e6	5	6



REZULTATI	CENTRI KLASTERA	ČLANOVI	KVADRAT GREŠKE
I	$k_1 = (2.67, 4.67)$ $k_2 = (2, 1.83)$	e2, e4, e6 e1, e3, e5	14.50
II	$k_1 = (1.5, 1.5)$ $k_2 = (2.75, 4.125)$	e1, e3 e2, e4, e5, e6	15.94
III	$k_1 = (1.8, 2.7)$ $k_2 = (5, 6)$	e1, e2, e3, e4, e5 e6	9.60

Nedostaci

- Algoritam daje najbolje rezultate ukoliko su klasteri približno iste veličine.
- Nemamo način da odredimo koji atributi su značajni za formiranje klastera. Uključivanje irelevantnih atributa utiče da rezultat ne bude optimalan.
- Odgovornost za objašnjavanje generisanih klastera je na nama!

Kako odrediti optimalan broj klastera?

- Videti prezentaciju:

Algoritam k-centara u DATAENGINE alatu.ppt