Podaci za Data Mining

Glava _. Sadržaj

- Skupovi podataka
- Atributi i tipovi atributa
- Tipovi skupova podataka
- Kvalitet podataka
- Mjerenje sličnosti u skupu podataka
- Analiza na osnovu veza u podacima

Skupovi podataka

- Skup podataka se sastoji od objekata (entiteta, slogova, tačaka itd.)
- Svaki objekat se opisuje skupom atributa (promjenljivih, dimenzija, svojstava itd.)

Atributi

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objekti

Atributi

- Atribut je svojstvo ili karakteristka objekta koje može da ima različite vrijednosti za različite objekte ili trenutke vremena
- Mjera je funkcija koja atributu dodjeljuje numeričku ili simboličku vrijednost
  - Jednom atributu može biti dodijeljeno više različitih vrijednosti
  - Različiti atributi mogu da imaju vrijednosti iz istog skupa

Tipovi atributa

- Podjela na osnovu svojstava vrijednosti koje se dodjeljuju atributima
- Nominalni ili imenički (engl. nominal)
  - boja očiju
- Redni (engl. ordinal)
  - ocjena na ispitu
- Intervalni (engl. interval)
  - period vremena
- Količnički (engl. ratio)
  - dužina

Svojstva vrijednosti atributa

- Svojstva ili operacije koje određuju podjelu
  - Sličnost =, !=
  - Uređenje <, >, <=, >=
  - Sabiranje i oduzimanje +, -
  - Množenje i dijeljenje *, /

  - Nominalni: sličnost
  - Redni: sličnost i uređenje
  - Intervalni: sličnost i uređenje i sabiranje i oduzimanje
  - Količnički: sličnost i uređenje i sabiranje i oduzimanje i množenje i dijeljenje

Tip atributa	Opis	Primjeri	
Nominalni	Vrijednosti atributa su različita imena, tj. nominalni atributi obezbjeđuju samo dovoljno informacija za razlikovanje jednog objekta od drugog (=, ≠)	Boja očiju, pol, JMBG	
Redni	Na osnovu vrijednosti atributa moguće je sortirati objekte. (<, >)	visina iz skupa {*nizak, prosječan, visok*}, ocjena, adresa	
Intervalni	Postoji jedinica mjere, razlika između vrijednosti je značajna. (+, -)	Datum, temperatura u stepenima Celzijusove skale	
Količnički	Razlika i odnos između vrijednosti je značajan. (*, /)	Dužina, temperatura u stepenima Kelvinove skale	

Atributi	Transformacija koja ne mijenja značenje atributa (S. Stevens)	Komentar
Nominalni	Svaka 1-1 funkcija	Ponovna dodjela brojeva indeksa ne mijenja značenje
Redni	Primjena monotone funkcije *new_value = f(old_value)*	Skup vrijednosti {1, 2, 3} može sa bude predstavljen skupom {0.5, 1, 10}.
Intervalni	*new_value =a * old_value + b* gdje su *a* i *b* konstante	Prevođenje iz Celzijusove u Farenhajtovu temperaturnu skalu
Količnički	*new_value = a * old_value* gdje je *a* konstanta	Mjerenje dužine metrima ili stopama.

Tipovi atributa prema broju vrijednosti

- Diskretni atributi
  - Konačan ili prebrojiv skup vrijednosti
  - Često se predstavljaju cijelim brojevima
  - Primjer: binarni atribut
  - Obično su to nominalni i redni atributi
- Kontinualni atributi
  - Vrijednosti su iz skupa realnih brojeva
  - Često se predstavljaju pokretnim zarezom (float tip)
  - Primjer: temperatura, visina, težina
  - Obično su to intervalni i količnički atributi

Asimetrični atrbuti

- Samo su važne ne nula vrijednosti
- Binarni asimetrični atributi naročito važni za asocijativnu analizu
- Mogu da budu i asimetrični diskretni ili asimetrični kontinualni

Tipovi skupova podataka

- Record data
  - Matrice podataka
  - Document data
  - Transakcione BP
- Graph-based data
  - Hemijski molekuli
  - WWW
- Oredred data
  - Temporal (sequential) data
  - Sequence data

Karakteristike skupova podataka

- Dimenzionalnost
  - Broj atributa koje posjeduju objekti
  - Curse of Dimensionality
- Rijetkost
  - Npr. manje od 1% ne nula vrijednosti
- Stepen generalizacije
  - Otkrivanje šablona zavisi od stepena generalizacije na kojem su predstavljeni podaci

Record data

- Skup podataka se sastoji od slogova, a svaki slog od fiksiranog broja polja (atributa)
- Čuvaju se u običnim datotekama ili relacionim BP

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Matrice podataka

- Svaki objekat se opisuje jednakim skupom numeričkih atributa i posmatra se kao vektor u više-dimenzionalnom prostoru gdje dimenzija predstavlja atribut
- Skup podataka se predstavlja matricom sa m vrsta i n kolona: jedna vrsta za jedan objekat, jedna kolona za jedan atribut

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Rijetke matrice podataka

- Atributi su istog tipa i asimetrični su

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

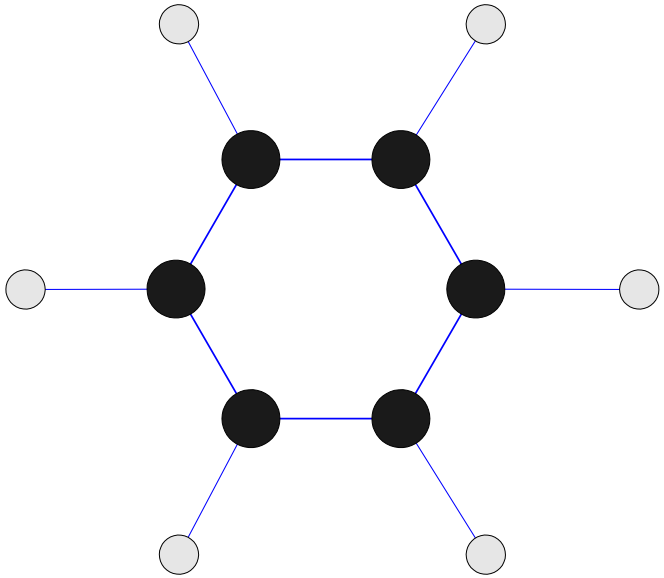
Transakcione BP

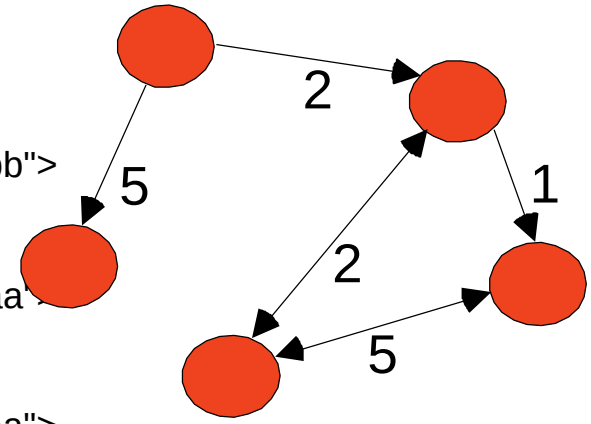
- Slog predstavlja transakciju
  - Svaka transakcija sadrži skup objekata (item-a)
  - market basket data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph data

- Dva tipa
  - Grafom se predstavljaju relacije između objekata
  - Sami objekti su grafovi





Data Mining

Graph Partitioning

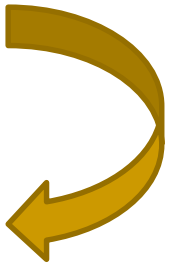
Parallel Solution of Sparse Linear System of Equations

<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

Ordered data

- Temporal data

Vrijeme	Kupac	Proizvodi
T1	C1	A, B
T2	C3	A, C
T2	C1	C, D
T3	C2	A, D
T4	C2	E
T5	C1	A, E



Kupac	Vrijeme + Proizvodi
C1	(t1 : A, B) (t2 : C, D) (t5 : A, E)
C2	(t3 : A, D) (t4 : E)
C3	(t2 : A, C)

Ordered data (2)

- Sequence data
  - Nema vremenske odrednice
  - Položaj u uređenoj sekvenci

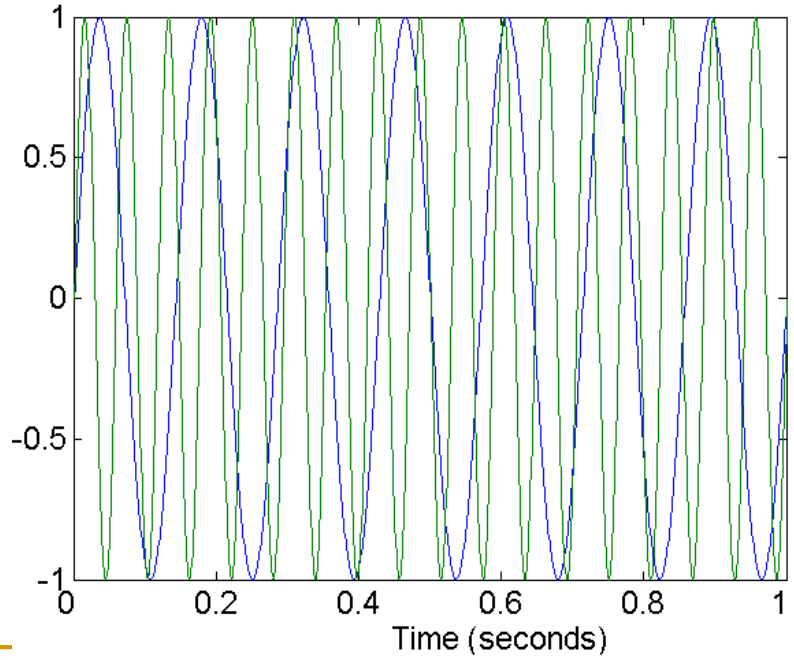
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

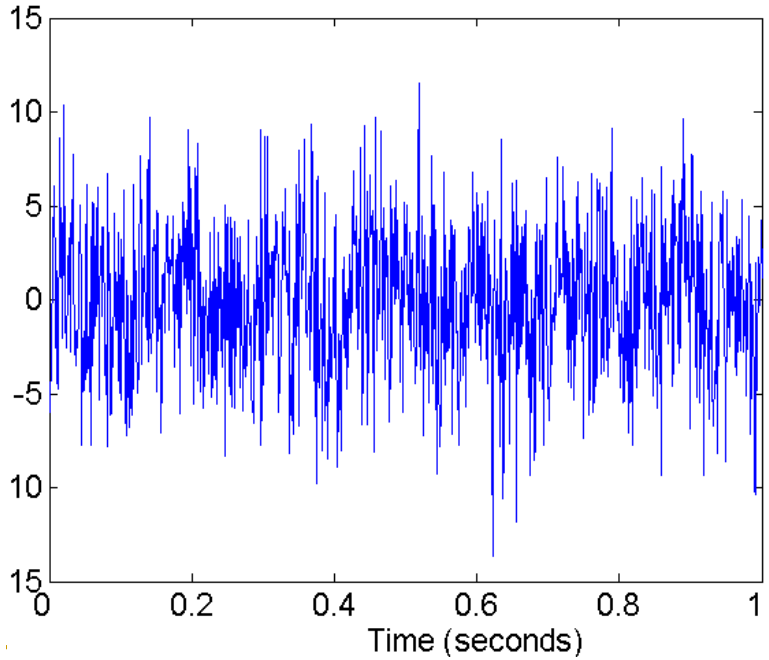
Kvalitet podataka

- Problemi vezani za kvalitet podataka
  - Greške u mjerenju
    - šum u podacima
  - Greške u sakupljanju podataka
    - izuzeci (outliers)
    - nedostajuće i nekonzistentne vrijednosti
    - duplikati
- Otkrivanje i ispravljanje problema kvaliteta je čišćenje podataka (data cleaning)

Šum u podacima

- Šum je slučajna greška mjerenja podataka
  - Primjer: ljudski glas preko telefonske žice ili sniježenje slike



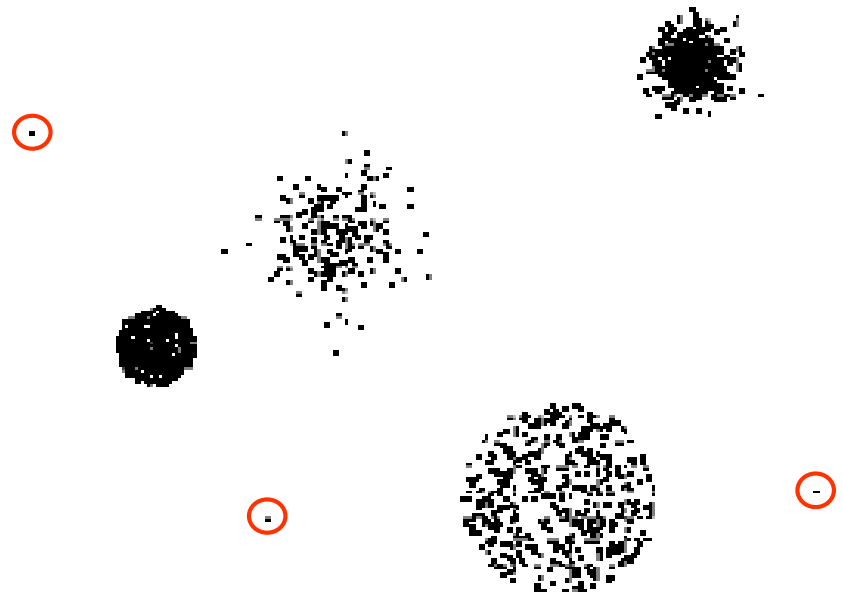


Bias i preciznost u skupu podataka

- Pretpostavka je da je izvršeno više mjerenja jednog objekta i srednja (prosječna) vrijednost se uzima kao stvarna mjera
- Bias = abs(mean – stvarna vrijednost)
- Preciznost = uzoračka disperzija
- Primjer: testira se vaga tako što se 5 puta mjeri 1g. Rezultati mjerenja su {1.015, 0.99, 1.013, 1.001, 0.986}, pa je
  - mean = 1.001
  - bias = 0.001
  - preciznost = 0.013

Izuzeci

- Izuzeci su
  - Objekti koji su po karakteristikama jako različiti od ostalih objekata u skupu podataka
  - Vrijednosti atributa koje su neočekivane za taj atribut u skupu podataka



Nedostajuće vrijednosti

- Razlozi
  - Vrijednosti nijesu unijete
  - Vrijednosti nijesu odgovarajuće za sve objekte
- Strategije
  - Eliminacija objekata
  - Eliminacija atributa
  - Ignorisanje nedostajućih vrijednosti tokom analize
  - Procjena nedostajućih vrijednosti

Duplikati

- Skupovi podataka mogu da sadrže duplikate ili “skoro” duplikate
- Detekcija duplikata
  - Ako u skupu podataka postoje dva objekta koja u stvari predstavljaju jedan, vrijednosti odgovarajućih atributa mogu da se razlikuju (nekonzistentnost)
  - Slični objekti nijesu uvijek duplikati

Mjerenje sličnosti u skupu podataka

- Sličnost objekata u skupu podataka je važna za veliki broj algoritama
- Sličnost
  - Broj koji pokazuje koliko su slična dva objekta, često iz segmenta [0,1]
- Različitost (ili rastojanje)
  - Broj koji pokazuje koliko su različita dva objekta
  - Minimalna različitost je 0, gornja granica često 1 ili beskonačno
- Proximity: sličnost ili različitost

Transformacije

- Transforacijama se
  - mjera sličnosti prevodi u mjeru različitosti
  - proximity mjera prevodi u željeni interval
- Transformacija različitosti u sličnost ma kojom monotono opadajućom funkcijom
  - s=-d ili s=1/(d+1) ili s=exp(-d)
- Transformacija u [0,1]: s1=(s-min_s)/(max_s-min_s)

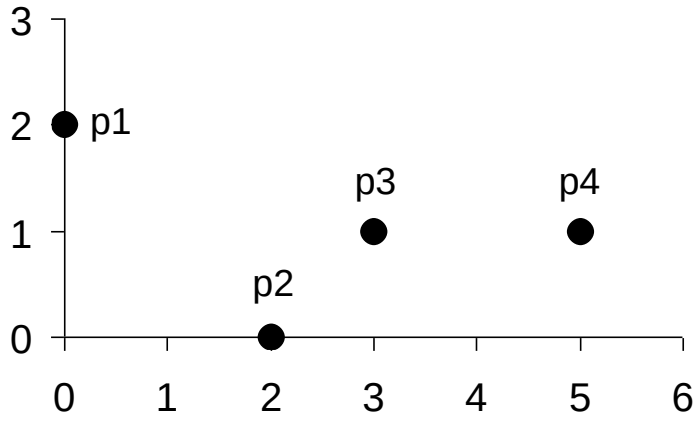
Sličnost među objektima sa jednim atributom

- *p* i *q* su vrijednosti atributa za dva objekta

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Rastojanje

- Euklidsko rastojanje $$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Rastojanje (2)

- Minkowski rastojanje $$dist = (\sum_{k=1}^{n} | p_k - q_k |^r)^{\frac{1}{r}}$$
- r=1, city block distance. Primjer je Hamingovo rastojanje između binarnih vektora
- r=2, Euklidsko rastojanje
- $r \to \infty$, maksimalno rastojanje između atributa dva objekta
- City block distance, Euklidovo i maksimalno rastojanje su definisani za sve vrijednosti n

Rastojanje (3)

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Svojstva rastojanja

- Rastojanje je mjera različitosti koja zadovoljava uslove (zove se i metrika)
  1. $d(p, q) \geq 0$ za sve p i q
  2. $d(p, q) = 0$ samo ako je $p = q$
  3. $d(p, q) = d(q, p)$ za sve p i q
  4. $d(p, r) \leq d(p, q) + d(q, r)$ za sve p, q i r
- Mjere za različitost koje nijesu metrike
  - Razlika skupova

Sličnost

- Uobičajena svojstva mjere sličnosti s(x, y) su
  1. $s(p, q) = 1$ (poklapanje) samo ako je $p = q$
  2. $s(p, q) = s(q, p)$ za sve p i q

Sličnost za binarne vektore

- Objekti x i y imaju n binarnih atributa
- Sličnost se računa na osnovu
  - M_{01} = broj atributa koji su 0 za x a 1 za y
  - M_{10} = broj atributa koji su 1 za x a 0 za y
  - M_{00} = broj atributa koji su 0 za x i 0 za y
  - M_{11} = broj atributa koji su 1 za x i 1 za y
- Simple matching coefficient (SMC) = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
- Jaccard coefficient = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

Sličnost za binarne atribute, primjer

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

Kosinus kao mjera sličnosti

- Karakteristike
  - Ingnoriše 0-0 sličnost
  - Primjenljiva ne samo za binarne atribute
  - Najčešće se koristi za sličnost dokumenata
- Ako su x i y dokument predstavljeni vektorima, tada je $\cos(x,y)=x*y/||x||\ ||y||$, gdje je x*y skalarni proizvod a $||x||^2=x*x$

Kosinus kao mjera sličnosti (2)

- Primjer

  d_1 = **3 2 0 5 0 0 0 2 0 0**

  d_2 = **1 0 0 0 0 0 0 1 0 2**

  $d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

  $||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

  $||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

  $\cos(d_1, d_2) = .3150$

Prošireni Jaccard koeficijent

- Može da se primijeni na sličnost dokumenata,
- U slučaju binarnih atributa svodi se na Jaccard koeficijent
- Naziva se i Tanimoto koeficijent

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q},$$

Korelacija

- Korelacija između dva objekta je linearna mjera veza između njihovih artributa (binarnih ili kontinualnih)
- Korelacija je iz [-1,1], ako je 1 (-1) onda je savršena pozitivna (negativna) linearna veza između x i y, ako je 0 nema linearne veze

Correlation(x, y) = cov(x, y) / std(x) * std(y)

Računanje sličnosti ako su atributi različitih tipova

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Računanje sličnosti sa težinama

- Težina atributa je njegov značaj za određivanje sličnosti
  - Težine su između 0 i 1, a njihova suma je 1

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$

Gustina kao mjera sličnosti

- Density-based klasterizacija prepoznaje klastere kao regione sa velikom gustinom
- Euklidska gustina za objekat je broj objekata u krugu zadatog poluprečnika