Osnovna svojstva podataka

Milan M.Milosavljević

Osnovni pojmovi o podacima

- Skup objekata i njihovih atributa
- Atributi su svojstvo ili karakteristika objekta
  - Primer: temperatura, boja auta, veličina ekrana, itd.
  - Atributi su poznati i kao promenljive, polja, osobine, karakteristike, ...
- Skup atributa opisuje objekat
  - Objekat je takođe poznat i kao slog, tačka, slučaj, primer, entitet, instanca, ...

Atributi- obeležja (eng. Features)

Objekti -primeri

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

Tipovi atributa

- Vrednosti atributa su brojevi ili simboli koji su im pridruženi
- Osobine i operacije (nad brojevima) koje se najčešće koriste radi određivanja tipa atributa su:
  - Različitost: $=$ i $\neq$
  - Uređenje: $<$, $\leq$, $>$ i $\geq$
  - Aditivnost: $+$ i $-$
  - Multiplikativnost: $*$ i $/$

Tipovi atributa

- Prema ovim osobinama mogu se definisati
  - Imenski atributi - različitost
    - Primer: JMBG, boja očiju, poštanski broj, radno mesto
  - Redni atributi - različitost i uređenje
    - Primer: rangiranje, godine studija, poređenja (dobar, loš, zao), ...
  - Intervalni atributi – različitost, uređenje i aditivnost
    - Primer: dan u nedelji, datum, temperatura (u stepenima Celzijusa)
  - Razmerni atributi – sve četiri osobine
    - Primer: temperatura u Kelvinima, dužina, vreme, ...

Tip atributa	Opis	Primeri	Operacije
Imenski (eng. Nominal)	Vrednost imenskog atributa su upravo različita imena, tj. imenski atributi pružaju samo mogućnost razlikovanja jednog od drugog objekta (=, ≠)	poštanski kodovi, identifikacije zaposlenih, boja očiju, pol (muški, ženski)	način, entropija, korelacija kontingenata, χ^2 test
Redni (eng. Ordinal)	Vrednosti rednih atributa pružaju dovoljno informacija za uređenje objekata (<, >)	tvrdoća minerala, stepeni, redni brojevi zgrada u ulici	procenat, korelacija ranga, izvršavanje testova oznake testova
Intervalni (eng. Interval)	Za intervalne atribute , ima smisla razlika između vrednosti, tj. postoji jedinica mere takvih atributa (+, -)	datumi u kalendaru, temepratura u stepenima Celizijusa	srednja vrednost, standardna devijacija
Razmerni (eng. Ratio)	Kod razmernih atributa ima smisla i proizvod i količnik (*, /) tih atributa	temepratura u Kelvinima, količina novca, godine, masa, dužina	geometrijska sredina, harmonijska sredina, procenat varijacije

Vrsta atributa	Transformacija	Komentar
Imenski	Bilo koja permutacija vrednosti, tj. preslikavanje 1-1	Ako svi zaposleni dobiju nove identifikacije to neće doneti bilo kakve razlike
Redni	Promena vrednosti koja čuva uređenje: *new_value = f(old_value)* *gde je f* monotona funkcija	Atribut koji sadrži poređenje *dobar, bolji , najbolji* podjednako dobro je predstavljen vrednostima {1, 2, 3} ili { 0.5, 1, 10}.
Intervalni	*new_value =a * old_value + b* gde su a i b konstante	Celizijusova i Farenhatjova temperaturna skala se razlikuju u veličini stepena i u tome gde je nula
Razmerni	*new_value = a * old_value*	Dužina može da se meri u metrima ili stopama.

Diskretni i kontinuirani atributi

- Diskretni atributi
  - Imaju konačan ili prebrojivo beskonačan skup vrednosti
  - Primer: poštanski brojevi, računi, skup reči u nekom dokumentu
  - Često se prikazuju kao celobrojne promenljive
  - Binarni atributi su specijalan slučaj diskretnih atriubuta

- Kontinuirani (neprekidni) atribututi
  - Skup vrednosti ovih atributa čine realni brojevi
  - Primer: temepratura, visina, težina, pritisak, brzina
  - Realne vrednosti mogu da se mere i predstavljaju samo preko konačnog broja cifara
  - Uobičajen način predstavljanja je u obliku realnih brojeva u pokretnom zarezu

Asimetrični atributi

- Jedino se prisustvo ne-nula vrednosti smatra značajnim
  - Na primer, neka je objekat student čiji su atributi informacija da li je student slušao neki od kurseva koji se drže na univerzitetu.
  - Za konkretnog studenta vrednost atributa 1 znači da je on slušao kurs pridružen tom atributu, a 0 da nije slušao
  - Najveći broj vrednosti će biti 0
  - Efikasnije je koncentrisati se na ne-nula vrednosti (ako to ne uradimo, a studenti se porede npr. po kursevima koje nisu uzeli tada će svi studenti biti vrlo slični jer je broj mogućih kurseva velik)

- Binarni atributi kod kojih su bitne ne-nula vrednosti se zovu asimetrični binarni atributu.

Tipovi skupova podataka

- Data set – skup podataka
- Slogovi
  - Matrica podataka
  - Podaci u dokumentima
  - Transakcioni podaci
- Grafovi
  - World Wide Web
  - Molekulane strukture
- Podaci sa poretkom (eng. Ordered)
  - Prostorni podaci
  - Vremenski (zavisni) podaci
  - Redosledni podaci
  - Genetički redosledni podaci

Značajne karakteristike struktuiranih podataka

- Dimenzionalnost
  - Broj atributa koje poseduje objekat iz skupa podataka
  - Prokletstvo dimenzionalnosti – teškoće pri analizi podataka sa velikim brojem dimenzija
  - Primenjuje se dimenzionalna redukcija
- Proređenost
  - Broji se samo prisustvo. Npr. asimetrični atributi
  - Prednost zbog ušteda prostora i vremena
- Rezolucija
  - Obrasci zavise od skaliranja. Npr. različite razmere pri predstavljanju površine Zemlje

Slogovni podaci

- Podaci se sastoje od skupa slogova od kojih se svaki sastoji od fiksnog skupa atributa

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

Matrica podataka

- Ako objekti imaju identičan skup fiksiranih numeričkih atributa, tada možemo da ih posmatramo kao da su u pitanju tačke u višedimenzionalnom prostoru u kome svaka dimenzija odgovara jednom od različitih atributa.
- Takvi skupovi podataka se predstavljaju matricama gde su objekti predstavljeni u vrstama a atributi u kolonama.

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

Matrica terma u dokumentima

- Svaki dokument postaje vektor
  - moguće reči u dokumentima su termi koji se navode kao komponente vektora,
  - vrednost svake komponente je broj pojavljivanja te reči

	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

Proređene matrice

- Specijalan slučaj matrice podataka u kojoj su atributi istog tipa i asimetrični
- Primer: prethodna matrica terma u dokumentima
- U praksi se čuvaju jedino ne-nula upisi u ovakvoj matrici

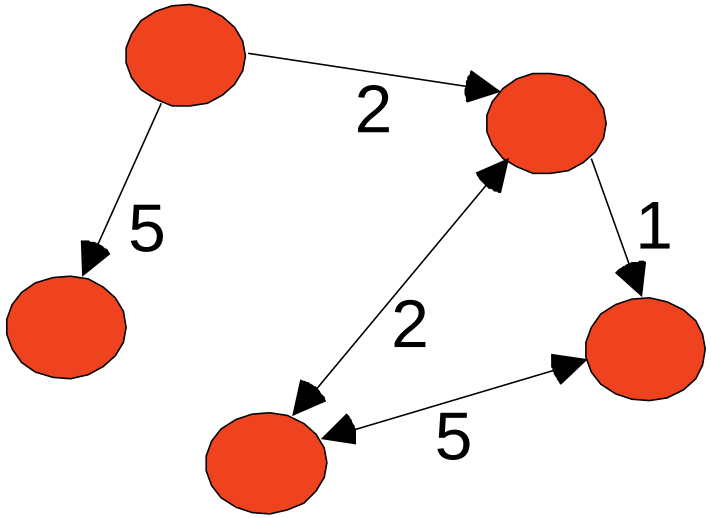
Transakcioni podaci

- Specijalni tip slogovnih podataka za koje važi:
  - svaki slog (transakcija) sadrži skup stavki
  - Na primer, podaci o prodavnici prehrambene robe. Transakciju predstavlja skup proizvoda koji je neki kupac kupio. Stavke su pojedinačni proizvodi.

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

Grafovski podaci

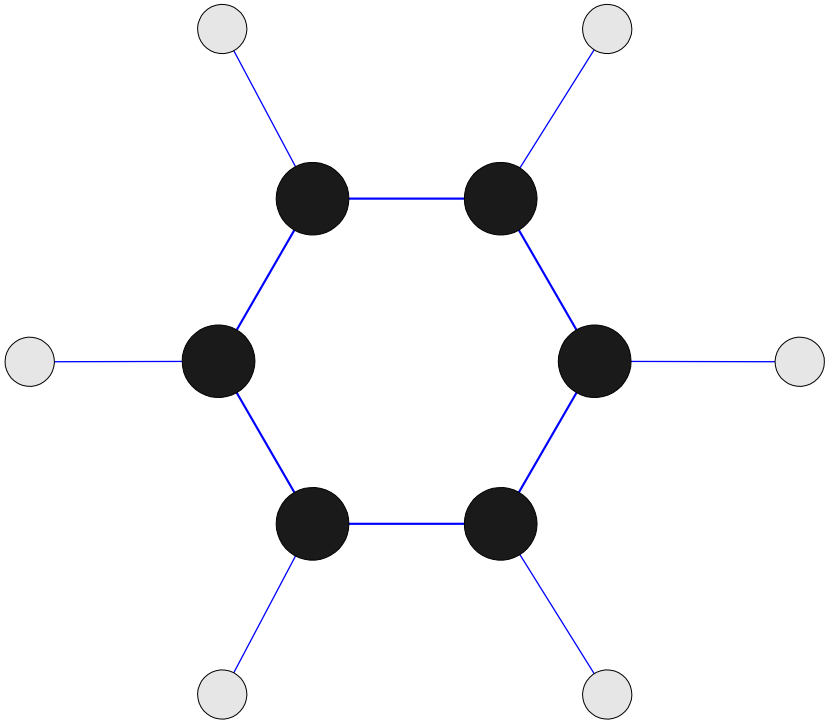
- Primer: Generički graf i HTML veze




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

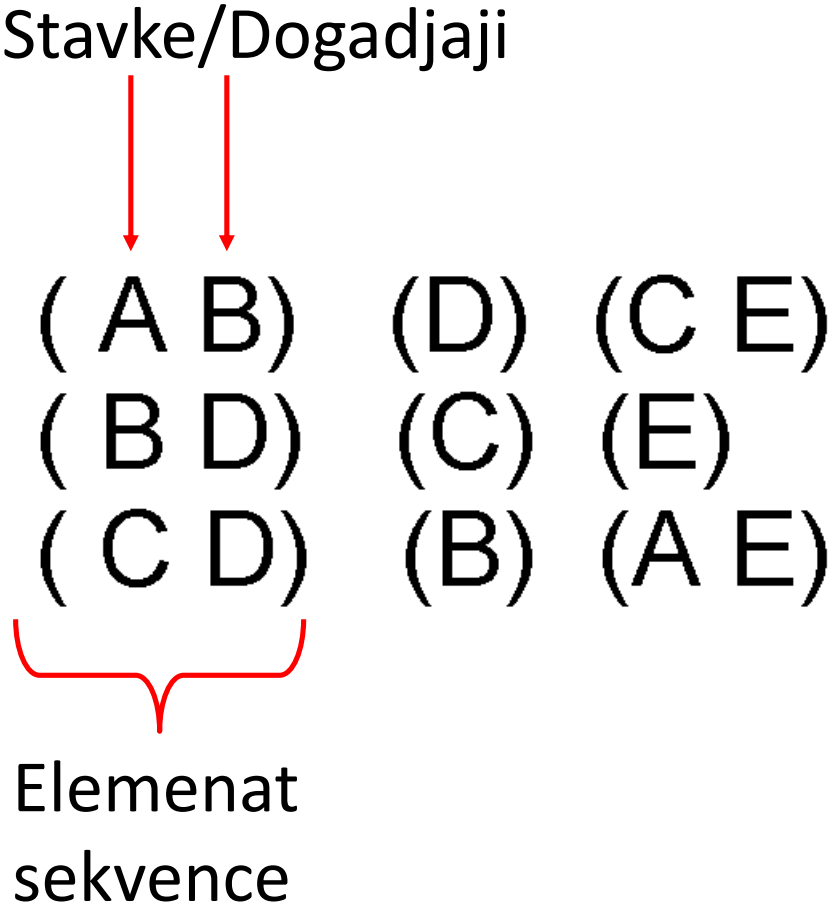
Podaci o hemijskim strukturama

- Molekul benzena: C_6H_6
  - Prikaz strukture (ugljenik - crno, vodonik - sivo)



Sekvencijalni podaci (podaci sa poretkom)

- Nazivaju se i vremenski podaci
- Niz transakcija



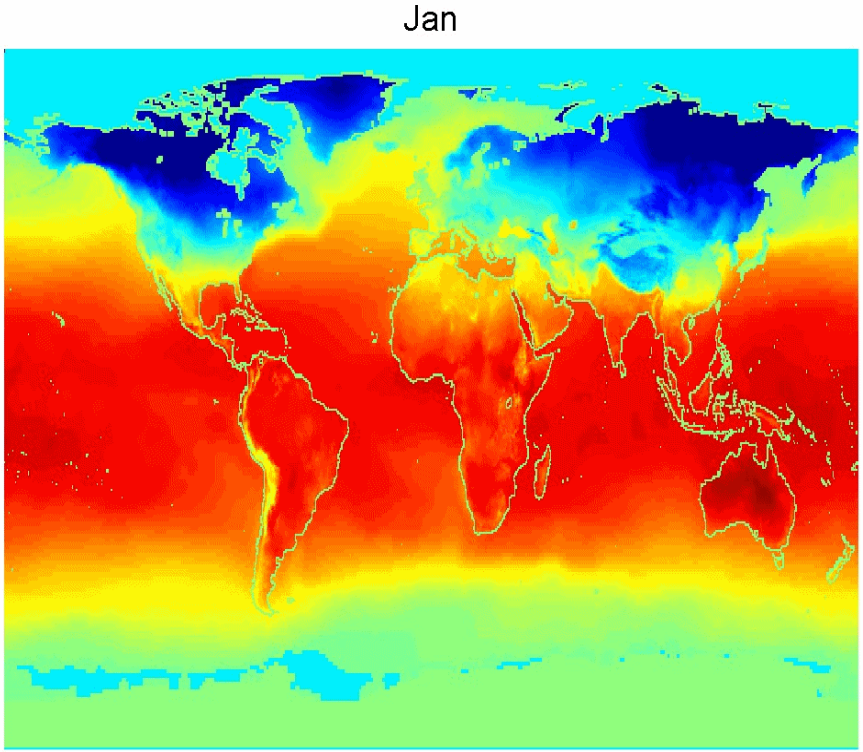
Sekvencijalni podaci (podaci sa poretkom)

- Genomske sekvence

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Sekvencijalni podaci (podaci sa poretkom)

- Prostorno-vremenski podaci
- U prostorne podatke spadaju i podaci vezani za vremenske serije



Prosečne mesečne temeprature kopna i mora

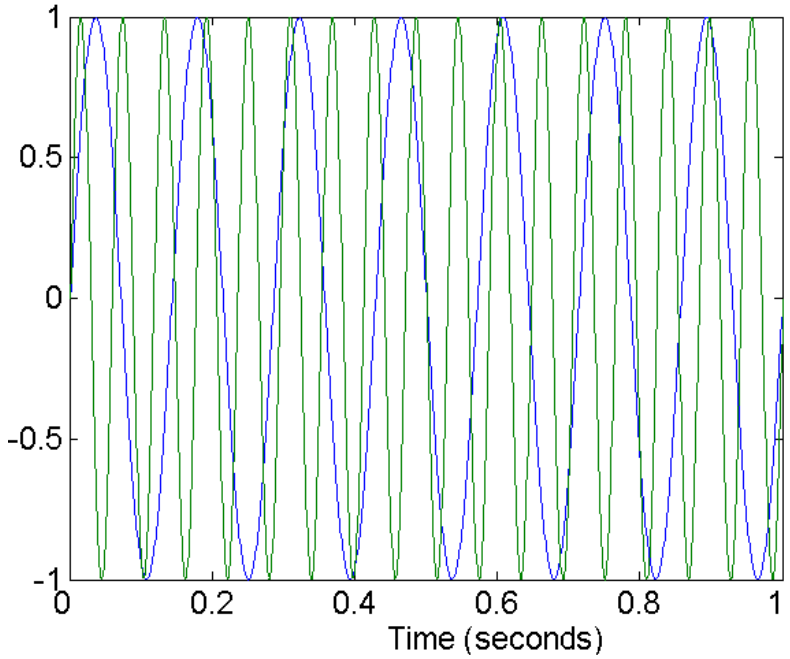
Kvalitet podataka

- Koje su vrste problema pri određivanju kvaliteta podataka?
- Kako odrediti probleme sa podacima?
- Šta raditi sa uočenim problemima?

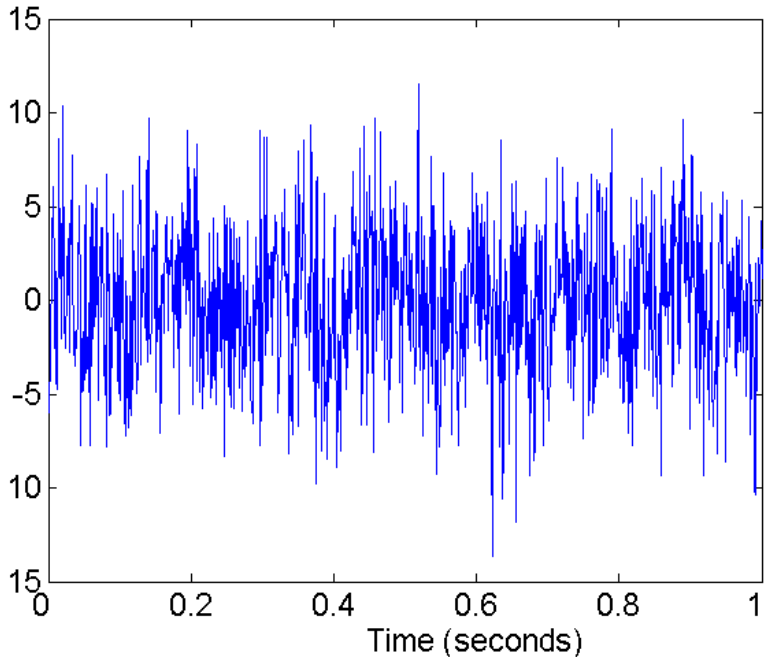
- Primer problema kvaliteta podataka:
  - šum i elementi van granica
  - nedostajuće vrednosti
  - duplirani (multiplicirani) podaci

Šum

- Šum predstavlja modifikaciju originalnih vrednosti
  - Primer: Izobličenje glasa osobe koja govori u mikrofon i pojava 'snega' na ekranu televizora, pogrešno očitani senzori



Signal dve sinusoide



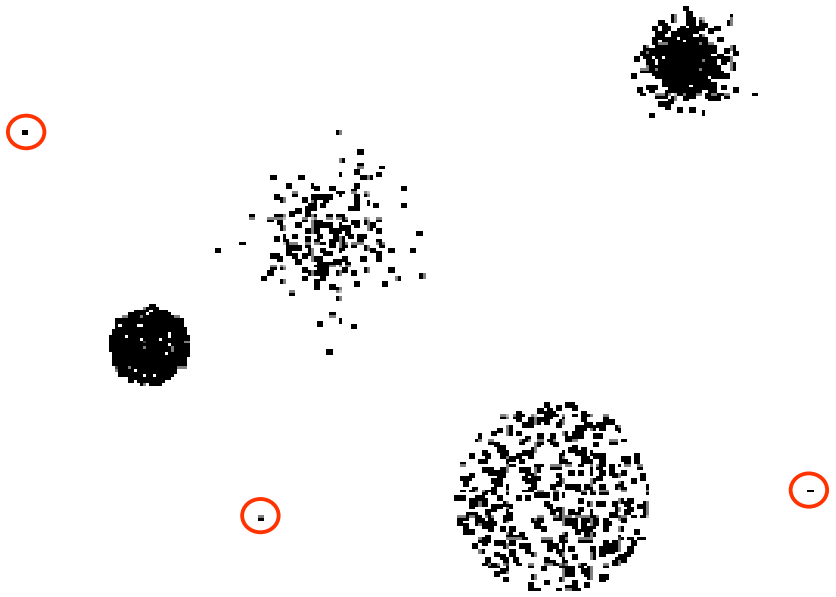
Signal dve sinusoide + šum

Šum

- Eliminacija šuma nije jednostavna
- Robusni algoritmi - daju prihvatljiva rešenja i kada je šum prisutan
- Deterministička izobličenja (npr. zamućenje na istom mestu celog skupa fotografija) se nazivaju *artifacts*

Vrednosti van granica (autlajeri)

- Autlajeri su objekti sa karakteristikama koje su značajno različite od najvećeg broja objekata u skupu podataka



Nedostajuće vrednosti

- Razlozi za pojavu
  - Informacije nisu prikupljene (npr. ljudi odbijaju da prikažu svoju težinu, starost, veličinu plate,...)
  - Atributi nisu primenljivi u svim slučajevima (npr. plata nije primenljiva na decu)

- Rukovanje nedostajućim vrednostima
  - Eliminacija objekata
  - Procena nedostajućih vrednosti
  - Ignorisanje nedostajućih vrednosti pri obradi
  - Zamena sa svim mogućim vrednostima (poređanim težinski prema verovatnoći pojavljivanja)
  - Nekonsistentne vrednosti

Duplirani podaci

- Skupovi podataka mogu da uključe duplikate, ili skoro identične podatke
  - Najčešće se javljaju kod spajanja podataka iz heterogenih izvora
- Primer:
  - Ista osoba sa više elektronskih adresa
- Proces obrade (eliminacije) duplikata se naziva čišćenje podataka

Preprocesiranje podataka

Primenjuje se radi dobijanja podataka koji više odgovaraju potrebama istraživanja podataka

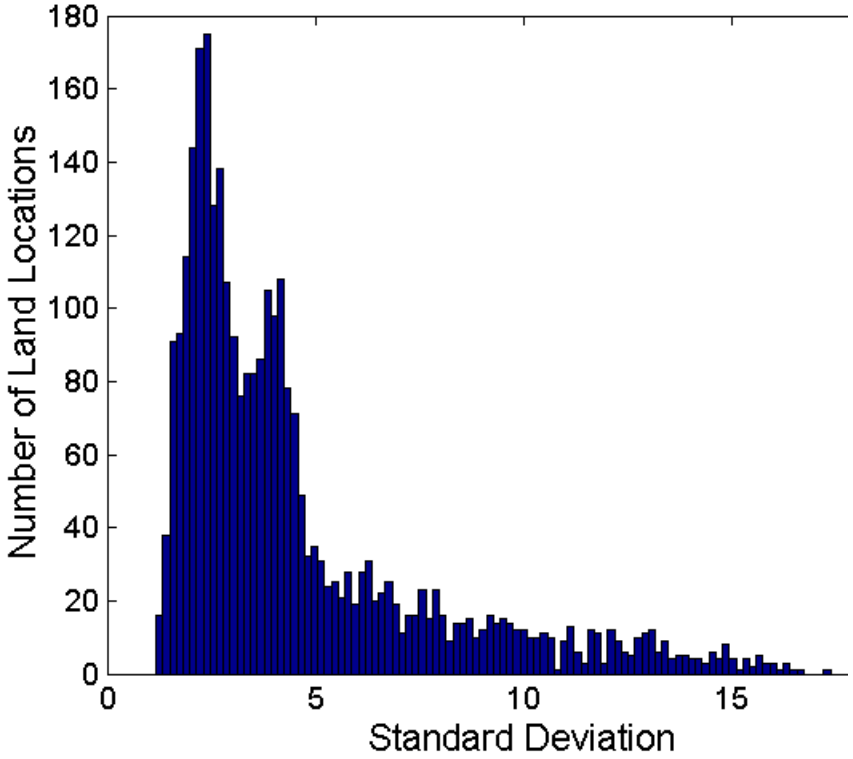
- Agregacija
- Izbor uzoraka (eng. *sampling*)
- Smanjenje dimenzije
- Izbor podskupa atributa
- Formiranje atributa
- Diskretizacija i binarizacija
- Transformacija atributa

Agregacija

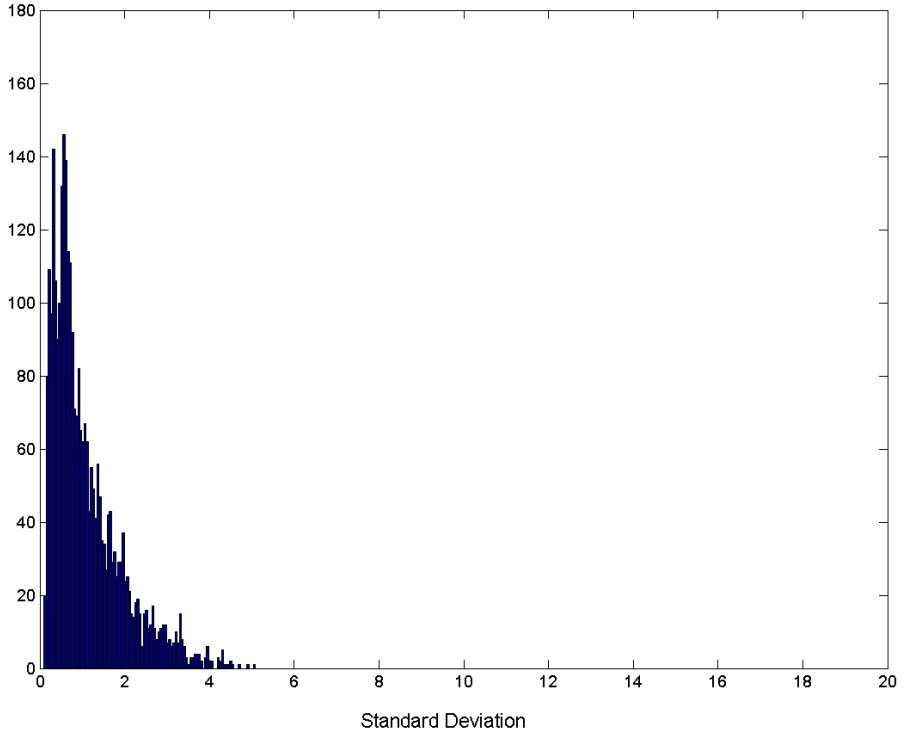
- Kombinovanje dva ili više atributa (ili objekata) u jedan atribut (objekat)
- Svrha
  - Redukcija podataka
    - Smanjivanje broja atributa ili objekata
  - Promena skale
    - Npr. umesto 365 dana dobijamo 12 meseci
  - 'Stabilniji' podaci
    - Agregirani podaci imaju tendenciju da imaju manja odstupanja

Agregacija

Primer: Vrednosti padavina u Australiji



Standardna devijacija prosečnih mesečnih padavina



Standardna devijacija prosečnih godišnjih padavina

Izbor uzoraka

- Izbor uzoraka je glavna tehnika koja se koristi u izdvajanju podataka.
  - Često se koristi kako za preliminarna istraživanja tako i za konačne analize podataka
- Statističari biraju uzorke jer je dobijanje kompletnog skupa podatakak koji su od interesa jako skupo i vremenski zahtevno
- Izbor uzoraka se koristi jer je obrada kompletnog skupa podataka od interesa računarski skupa ili vremenski zahtevna

Izbor uzoraka

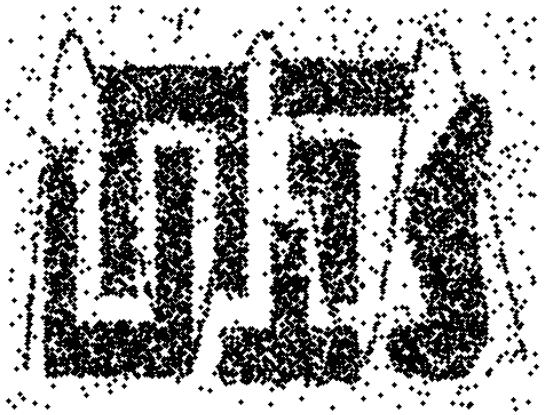
- Ključni principi za efektivan izbor uzoraka su:
  - Korišćenjem uzoraka koji su reprezentativni dobija se skoro isti efekat kao da je rađeno na kompletnom skupu podataka
  - Uzorak je reprezentativan ako ima aproksimativno iste osobine kao i originalni skup podataka

Tipovi uzoraka

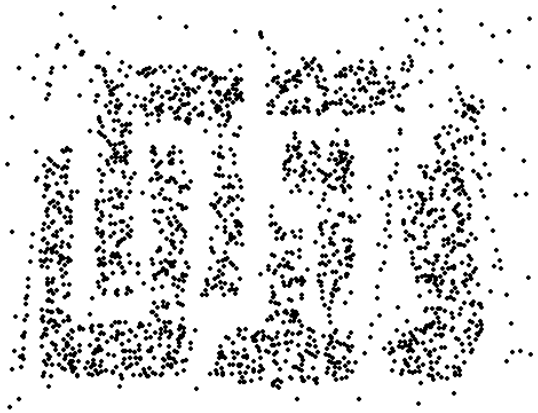
- Jednostavan slučajni uzorak
  - Postoji jednaka verovatnoća za izbor bilo koje slučajne stavke
- Izbor uzoraka bez zamene
  - Svaka stavka koja se bira uklanja se iz populacije
- Izbor uzoraka sa zamenom
  - Objekti se ne uklanjaju iz populacije po izboru u uzorak.
    - Posledica: isti objekat može da bude izabran više puta.
    - Jednostavnije za analizu jer verovatnoća izbora svake stavke ostaje ista u procesu izbora uzorka
- Izbor uzorka po slojevima
  - Podaci se dele u više delova, a zatim se bira jednostavan slučajni uzorak iz svakog od delova

Veličina uzorka

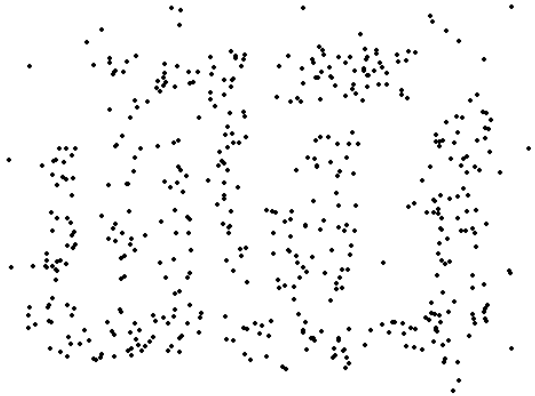
- Veličina uzorka treba da bude dovoljno velika da se ne naruši struktura objekta ili izostave interesantne osobine



8000 tačaka



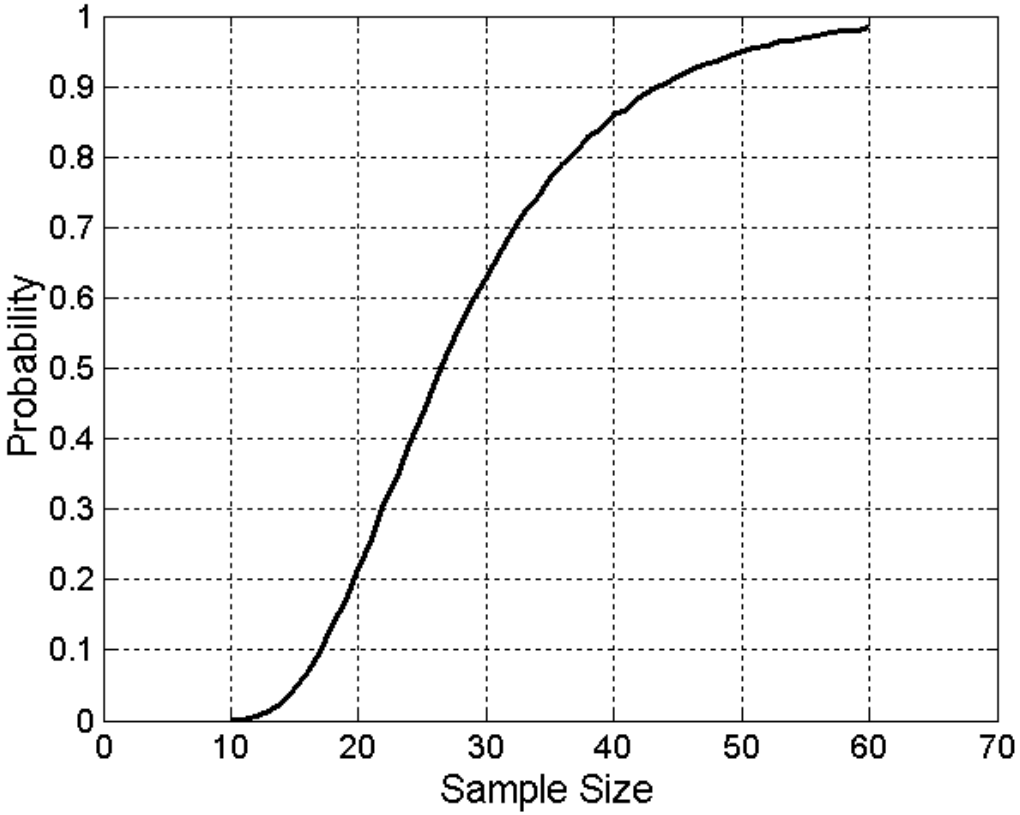
2000 tačaka



500 tačaka

Veličina uzorka

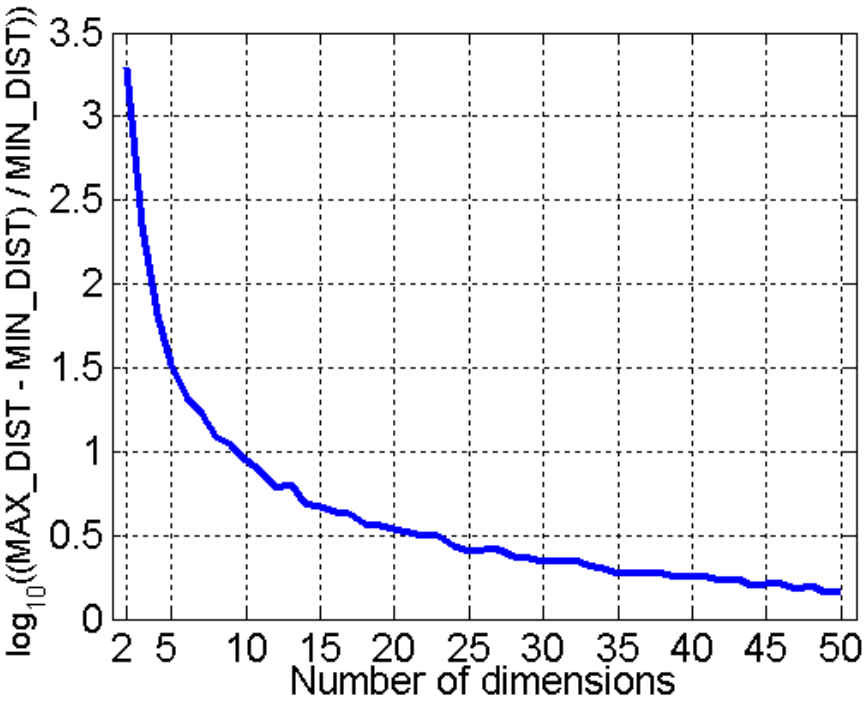
- Koja treba da je veličina uzorka da bi se u njemu našao po jedan objekat iz svake od 10 grupa?



Verovatnoća da uzorak sadrži tačke iz svake od 10 grupa

Prokletstvo dimenzionalnosti

- Kada se dimenzionalnost povećava, podaci postaju sve proređeniji u prostoru koji zauzimaju
- Definicije gustine i rastojanja između tačaka koje su kritične za klasterovanje i otkrivanje elemenata van granica postaju kontra intuitivne



- Slučajno se generiše 500 tačaka
- Računa se razlika maksimuma i minimuma rastojanja parova tačaka

Redukcija dimenzija

- Veliki broj algoritama bolje rade sa podacima manjih dimenzija
- Eliminišu se šum, redundantni podaci, ...
- Potreba za manjim obučavajućim i test skupovima
- Dobija se jednostavniji model
- Lakša vizualizacija

Redukcija dimenzija

- Metode redukcije dimenzija se generalno dele na
- Endogene - cilj je da zadrže što više informacija o skupu podataka kao celina (primer PCA)
- Egzogene – cilj je da zadrže diskriminatornu informaciju unutar datog skupa podatala (primer LDA)
- Tehnike
  - Analiza glavnih komponenata (eng. *Principle Component Analysis*)
  - Dekompozicija singularne vrednosti (eng. *Singular Value Decomposition*)
  - Druge nelinearne tehnike i tehnike sa nadzorom

Izbor podskupa atributa (feature selection)

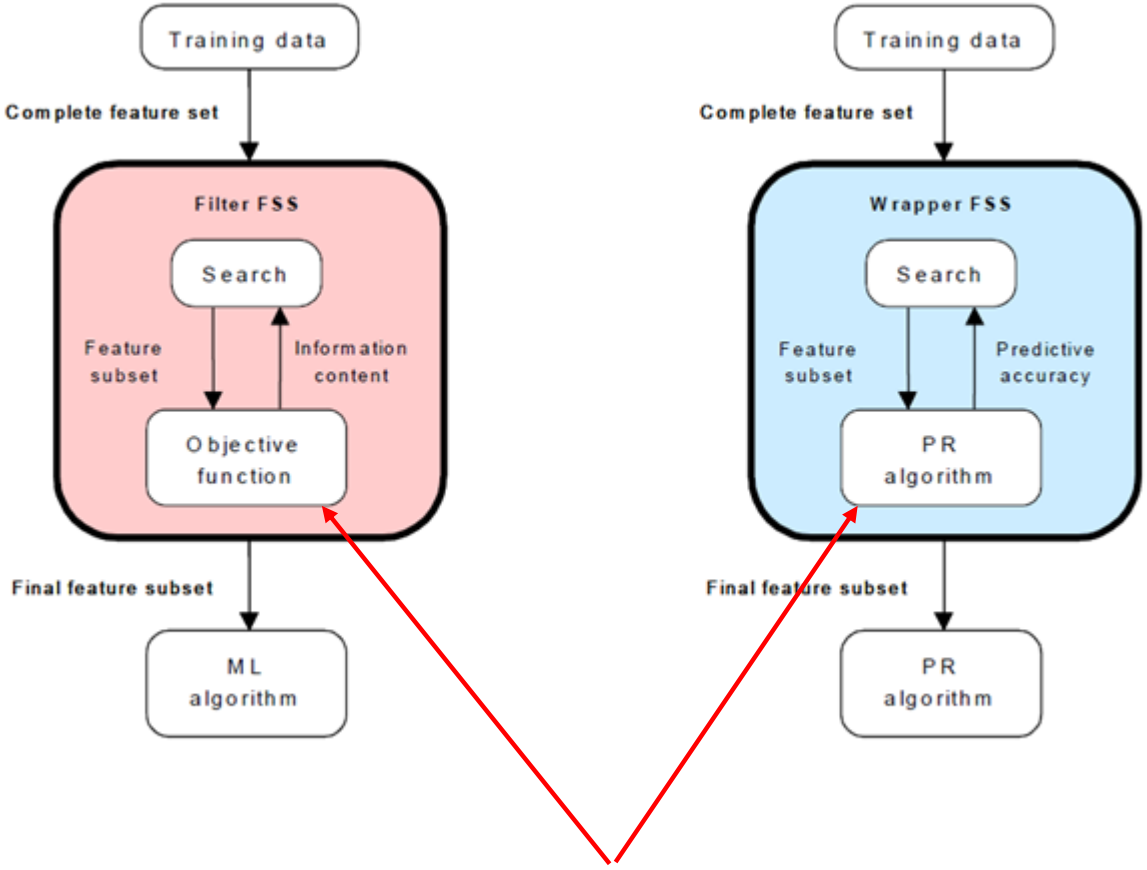
- Redundatni atributi
  - ponavljanje jedne ili svih informacija sadržanih u jednom ili više atributa
  - Primer: cena proizvoda i PDV

- Atributi sa irelevantnim vrednostima
  - sadrže informacije koje nisu korisne za proces IP-a
  - Primer: broj indeksa studenta je irelevantan za predviđanje prosečne ocene studenta

Izbor podskupa atributa (feature selection)

- Metod grube sile (Brute force):
  - Probaju se svi mogući podskupovi atributa kao ulaz u ML algoritam
  - Neprikladan zbog velikog broja podskupova
- Ugradjeni (Embedded) metode:
  - Izbor podskupova atributa je deo algoritma.
- Filterske metode:
  - atributi se biraju pre početka rada ML algoritma nekim pristupom koji je nezavisan od ML procesa
- Metode omotača (Wrapper):
  - Koristi se ML algoritam kao crna kutija koja pronalazi najbolji podskup skupa atributa.
  - Slično primeni grube sile ali se ne uzimaju u obzir baš svi podskupovi

Filters vs. Wrappers



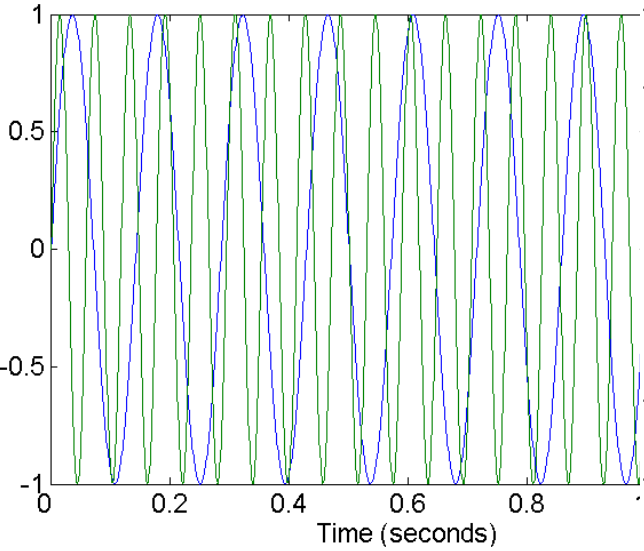
Formalno se razlikuju u kriterijumskoj funkciji

Formiranje obeležja (Feature extraction)

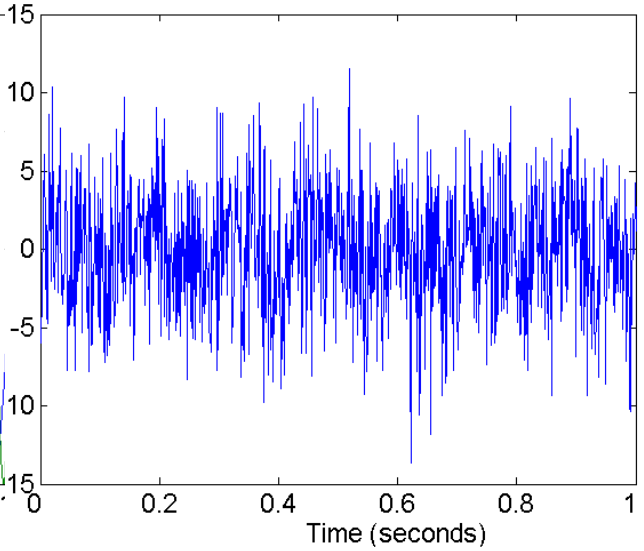
- Formiraju se nova obeležja koja sadrže najvažnije informacije iz skupa podataka na mnogo efikasniji način nego originalna obeležja
- Opšte metodologije:
  - Izdvajanje obeležja
    - zavisi od domena
  - Preslikavanje obeležja u novi prostor
  - Konstrukcija obeležja
    - kombinovanje (starih) obeležja

Preslikavanje atributa u novi prostor

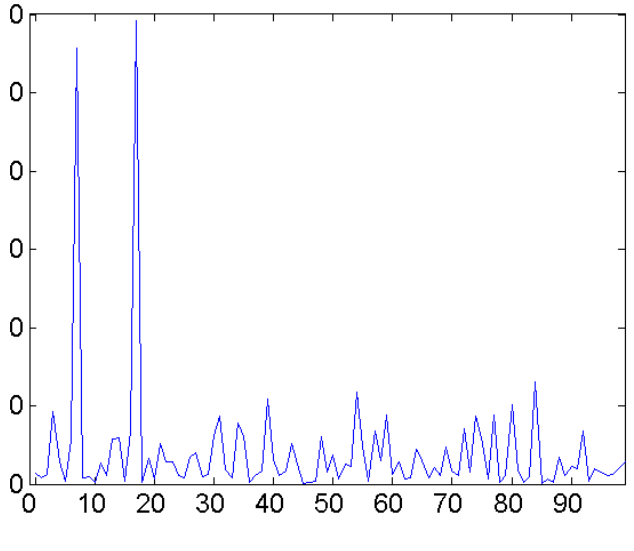
- Furijeove transformacije (Fourier transformation)
- Transformacije talasićima (Wavelet transformation)



Talas sa dve sinusoide



Talas sa dve sinusoide + šum



Frekvencija

Primena Furijeovih transformacija za identifikaciju frekvencija u podacima sa vremenskim serijama

Diskretizacija i binarizacija

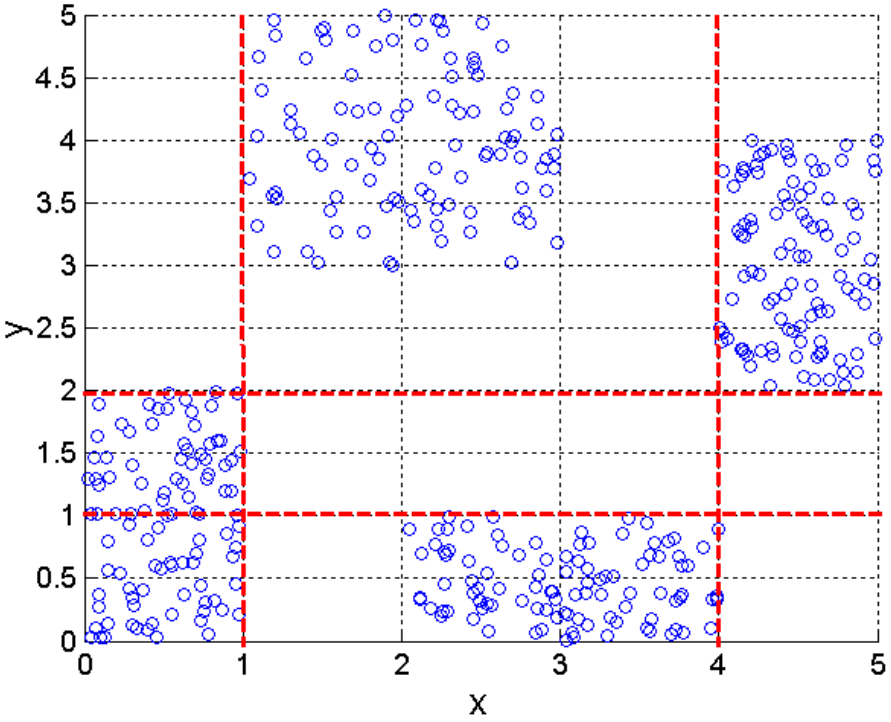
- Transformacija neprekidnih u disktetne atribute - diskretizacija
- Transformacija neprekidnih i diskretnih atributa u binarne - binarizacija
- Jednostavna tehnika binarizacije: ako ima *m* diskretnih vrednosti tada se svakoj dodeljuje jedinstven broj u intervalu [0,*m*-1] i konvertuje svaki od tih brojeva u binarnu vrednost

Diskretizacija i binarizacija

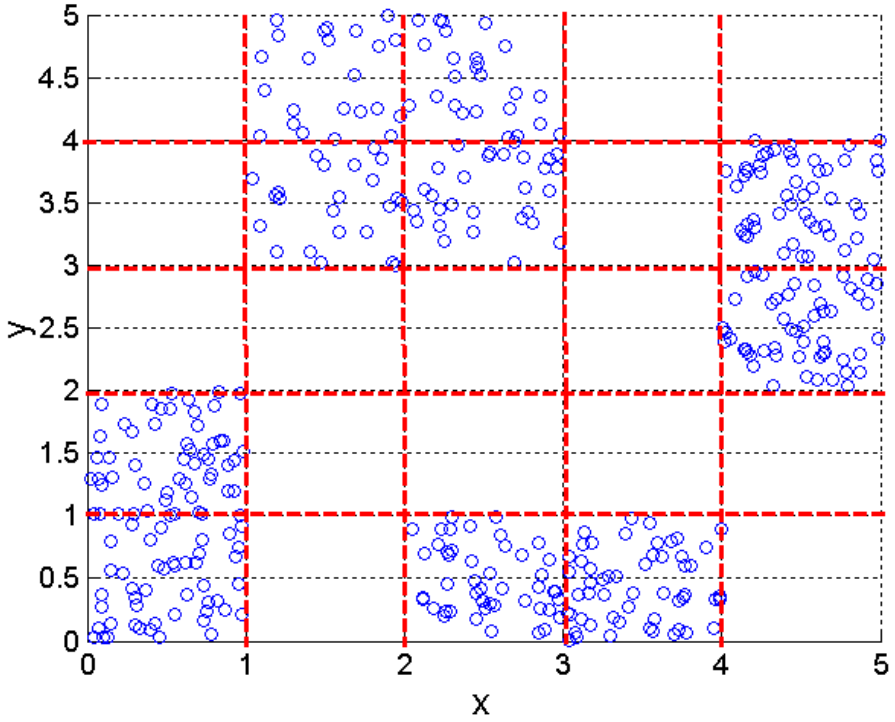
- Diskretizacija se obično primenjuje na atribute koji se koriste u klasifikaciji ili analizi zasnovanoj na pravilima pridruživanja
- Transformacija neprekidnih atributa u diskretne se sastoji iz dve faze:
  - odabrati broj kategorija
  - odrediti kako preslikati vrednosti neprekidnih atributa u te kategorije
- Na kraju prve faze, posle sortiranja, vrednosti neprekidnih atributa se dele u *n* intervala navođenjem *n-1* tačke razdvajanja
- U drugoj fazi sve vrednosti iz jednog intervala se preslikavaju u istu kategoričku vrednost.

Diskretizacija korišćenjem informacija o klasama

- Pristup zasnovan na entropiji

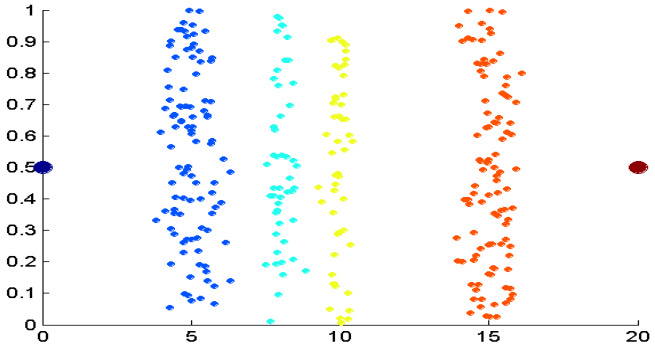


po 3 kategorije za x i y

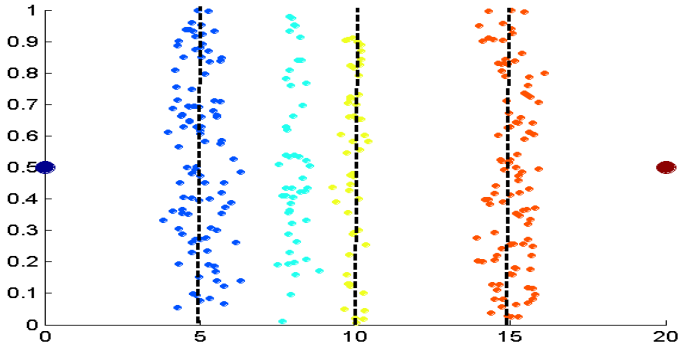


po 5 kategorija za x i y

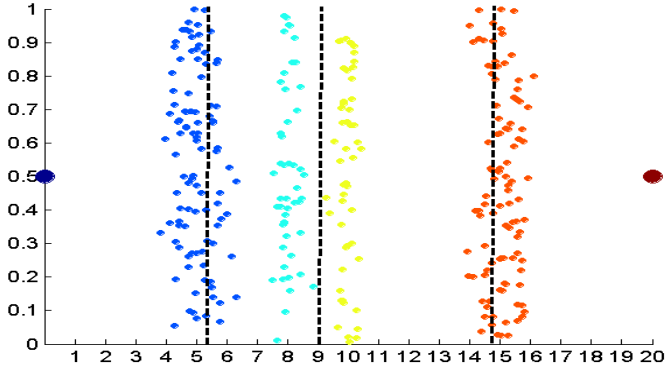
Diskretizacija bez korišćenja informacija o klasama



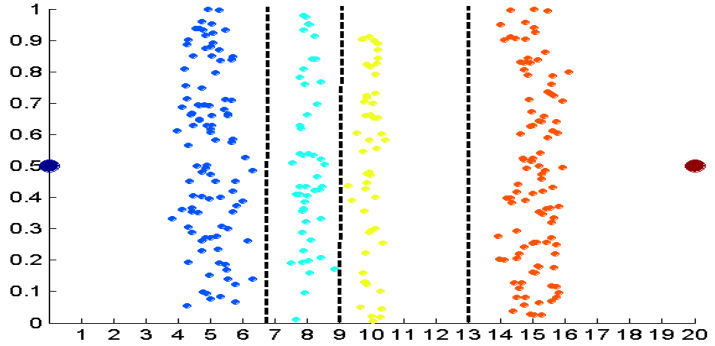
Originalni podaci



Intervali jednake širine



Jednaka frekvencija



K-srednjih

Transformacija atributa

- Transformacija promenljive označava transformaciju koja se primenjuje na sve vrednosti te promenljive.

- Za svaki objekat, transformacija se primenjuje na vrednosti promenljive za taj objekat.

Transformacija atributa – jednostavne funkcije

- Jednostavne funkcije, npr. : $\sqrt{x}$, x^k, log(x), e^x, |x|, 1/x
- U statistici se često koriste $\sqrt{x}$, log(x) i 1/x radi transformacije podataka koji nemaju Gausovu (normalnu) raspodelu u podatke koji imaju tu raspodelu
- U ML procesu ima i drugih razloga. Npr. ako je vrednost promenljive između 1 i 1.000.000.000, primenom log funkcije se dobijaju bolji odnosi kod poređenja (npr. 10^8 sa 10^9 i 10 sa 1000)
- Transformaciju promenljivih treba primenjivati sa oprezom jer može da promeni prirodu podataka (npr. transformacija sa 1/x)

Transformacija atributa - Standardizacija (normalizacija)

- Cilj: kompletan skup vrednosti treba da dobije neku željenu osobinu
- Primer: ako je $\overline{X}$ srednja vrednost (vrednosti) atributa i s_x strandardna devijacija vrednosti tog atributa, tada transformacija $x' = (x - \overline{X})/s_x$ formira novu promenljivu koja ima srednju vrednost 0 i standardnu devijaciju 1.
- Ako se različite promenljive kombinuju ne neki od načina, tada je ovakva transformacija neophodna da bi se izbegla dominacija u izračunavanjima promenljive koja ima veću vrednost

Transformacija atributa - Standardizacija (normalizacija)

- Primer: potrebno je porediti osobe uzimajući u obzir atribute starost i prihod. Ako se ne uzme u obzir različita priroda atributa, razlika u prihodima dve osobe je mnogo veća nego razlika u godinama.
- Sredina i standardna devijacija su jako osetljive na elemente van granica →vrši se modifikacija transformacije.
- Umesto srednje vrednosti se uzima medijana, a standardna devijacija se zamenjuje apsolutnom standardnom devijacijom (x_i je i-ta vrednost promenljive, m broj objekata a μ ili srednja vrednost ili sredina)

$$\sigma A = \sum_{i=1}^{m} |x_i - \mu|$$

Sličnost i različitost

- Sličnost
  - Numerička mera koliko su dva objekta slični
  - Što dva objekta više liče jedan na drugi sličnost im je veća
  - Često se meri vrednostima u intervalu [0,1]
- Različitost
  - Numerička mera koliko su dva objekta različiti
  - Što dva objekta više liče jedan na drugi različitost im je manja
  - Najmanja različitost je često 0; gornja granica varira
  - Kao sinonim koristi se i termin *rastojanje*
- Blizina (eng. proximity) označava ili sličnost ili različitost

Sličnost i različitost - transformacije

- Najčešće se vrši radi
  - konverzije sličnosti u različitost i obratno
  - transformacije mere blizine u interval [0,1]
  - U opštem slučaju transformacija
    - sličnosti u interval [0,1] se vrši izrazom

$$s' = (s - min_s)/(max_s - min_s)$$

    - različitosti u interval [0,1] se vrši izrazom

$$d' = (d - min_d)/(max_d - min_d)$$

    gde su s i d početne vrednosti, s' i d' su nove vrednosti, max i min su najveće odnosno najmanje vrednosti sličnosti i različitosti, respektivno

Sličnost i različitost objekata sa više atributa

- Bliskost objekata sa većim brojem atributa se tipično definiše kao kombinacija bliskosti pojedinačnih atributa.
- Neki načini određivanja sličnosti i različitosti za jednostavne (pojedinačne) atribute su prikazani u narednoj tabeli

Sličnost i različitost za jednostavne atribute

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 $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

p i *q* su vrednosti atributa za dva objekta

Euklidsko rastojanje

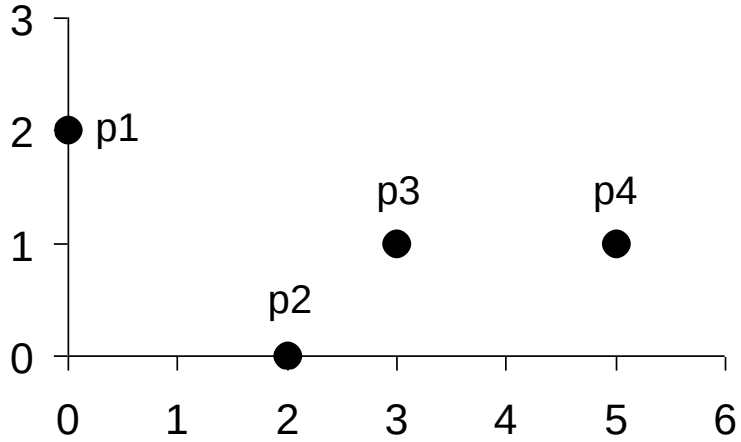
- Euklidsko rastojanje

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

gde su n je broj dimenzija (atributa), p_k i q_k su vrednosti k-tih atributa objekata p i q

- Ako se skale razlikuju neophodno je izvršiti njihovu stadradizaiciju

Euklidsko rastojanje



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

x i y koordinate tačaka

	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

Matrica rastojanja

Rastojanje Minkovskog

- Rastojanje Minkovskog je uopštenje Euklidskog rastojanja

$$dist = (\sum_{k=1}^{n} | p_k - q_k |^r)^{\frac{1}{r}}$$

gde su *r* parametar, *n* broj dimenzija (atributa), a p_k i q_k su vrednosti k-tih atributa objekata *p* and *q*

Rastojanje Minkovskog : Primeri

- r = 1. Gradski blok (City block) (L_1 norma) rastojanje.
  - Najčešći primer ovoga je Hamingovo rastojanje koje predstavlja broj različitih bitova između dva binarna vektora.

- r = 2. Euklidsko rastojanje

- $r \rightarrow \infty$. "supremum" (L_{max} norma, L_{∞} norma) rastojanje.
  - Predstavlja maksimum razlike između odgovarajućih komponenti vektora. Računa se kao *lim* po r u prethodnom izrazu.

Rastojanje Minkovskog : Primeri

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

x i y koordinate tačaka p1-p4

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

Matrice rastojanja

Uobičajene osobine rastojanja

Rastojanja, kao što je npr. Euklidsko, imaju neke dobro poznate osobine:

1. *Pozitivna određenost*
   1. $d(p, q) \geq 0$ za svako p i q
   2. $d(p, q) = 0$ samo ako je $p = q$
2. *Simetrija*

   $d(p, q) = d(q, p)$ for all p and q
1. *Nejednakost trougla*

   $d(p, r) \leq d(p, q) + d(q, r)$ za sve tačke p, q, i r.

   gde je $d(p, q)$ rastojanje (različitost) između tačaka (objekata) p iq

Rastojanje koje zadovoljava ove uslove se naziva metrika

Uobičajene osobine rastojanja

- Ne moraju sve različitosti da zadovoljavaju ovu metriku

- Razlika skupova A i B definisana kao
  d(A,B)=broj(A-B)
  gde broj(X)=broj elemenata skupa X

  Metrika važi ako se definiše
  d(A,B)=broj(A-B) + broj(B-A)

- Mera razlike između dva vremena d(t1,t2)
  Dati definiciju

Uobičajene osobine sličnosti

Sličnost takođe ima dobro poznate osobine:

1. *s(p, q) = 1 ako je p = q* (0≤ s ≤ 1)
2. *s(p, q) = s(q, p)* za svako *p* i *q* (Simetrija)

gde je *s(p, q)* sličnost između tačaka (objekata) *p* i *q*

Ne moraju sve sličnosti da zadovoljavaju ovu metriku.

- Na primer, matrica konfuzije za prepoznavanje slova o kao 0 i 0 kao slova o

Mera sličnosti za binarne podatke

- Neka su p i q binarni vektori. Mera njihove sličnost se obično naziva koeficijenat sličnosti i obično je u [0,1]

- Sličnost se računa pomoću sledećih vrednosti

  M_{01} = broj atributa koji su 0 u p i 1 u q

  M_{10} = broj atributa koji su 1 u p i 0 u q

  M_{00} = broj atributa koji su 0 u p i 0 u q

  M_{11} = broj atributa koji su 1 u p i 1 u q

Mera sličnosti za binarne podatke

- Koeficijent jednostavnog slaganja (eng. Simple Matching Coefficient -SMC)

  SMC = broj složenih / broj atributa

  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

- Džakardovi (Jaccard) koeficijenti

  Koriste se u slučaju asimetričnih atributa

  J = broj parova 11/ broj ne oba-su-nula vrednosti atributa

  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

Porimer: poređenje SMC i J

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 (broj atributa koji su 0 u p i 1 u q)

M_{10} = 1 (broj atributa koji su 1 u p i 0 u q)

M_{00} = 7 (broj atributa koji su 0 u p i 0 u q)

M_{11} = 0 (broj atributa koji su 1 u p i 1 u q)

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

Kosinusna sličnost

- Ako su d_1 i d_2 dva vektora dokumenata, tada važi

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

gde $\bullet$ označava skalarni proivod vektora $||\, d \,||$ je dužina vektora d.

Koristi se kod velikog broja parova tipa ’00’ pri čemu može da barata sa ne-binarnim vektorima (npr. poređenje sličnosti dva dokumenta po rečima koje se javljaju u njima)

Primer:

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 Džakardovi koeficijenti (Koeficijenti Tanimoto-a)

- Varijanta Džakardovih koeficijenata primenljiva na neprekidne i prebrojive atribute

- U slučaju binarnih atributa redukuje se na Džakardove koeficijente

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

Korelacija

Korelacija dva objekta koji imaju binarne ili neprekidne atribute je mera linearnog odnosa između njihovih atributa.

corr(x,y)=kovarijansa(x,y)/

/(standardna devijacija(x) *standardna devijacija(y))

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

tj. kovarinajca(x,y)=

$$s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

Korelacija (nastavak)

- Gde važe uobičajene statističke formule:
  - kovarinajns(x,y) $$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$
  - standardna devijacija(z) $$s_z = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2}$$
  - Srednja vrednost od z $$\bar{z} = \frac{1}{n}\sum_{k=1}^{n} z_k$$

Korelacija (nastavak)

- Za izračunavanje korelacije neophodno je standardizovati objekte x i y i zatim naći njihov skalarni proizvod.

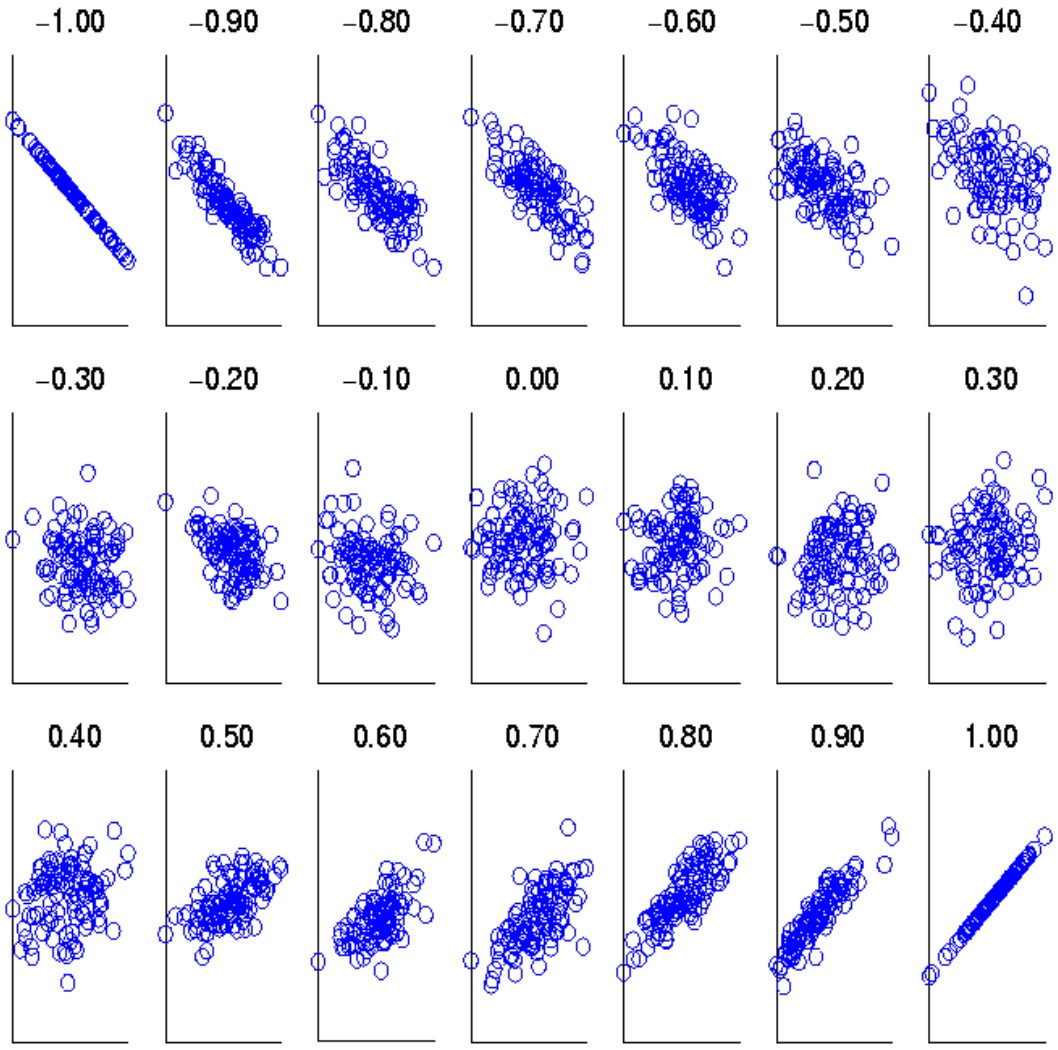
$$x_k^{'} = (x_k - \bar{x}) / std(x)$$

$$y_k^{'} = (y_k - \bar{y}) / std(y)$$

$$corr(x, y) = x^{'} \cdot y^{'}$$

Ako je korelacija =1 (-1) → perfektni pozitivan (negativan) linearni odnos $x_k=ay_k+b$

Vizuelno ocenjivanje korelacije



Rasute tačke pokazuju slličnost od –1 do 1.

Mahanalobisovo rastojanje

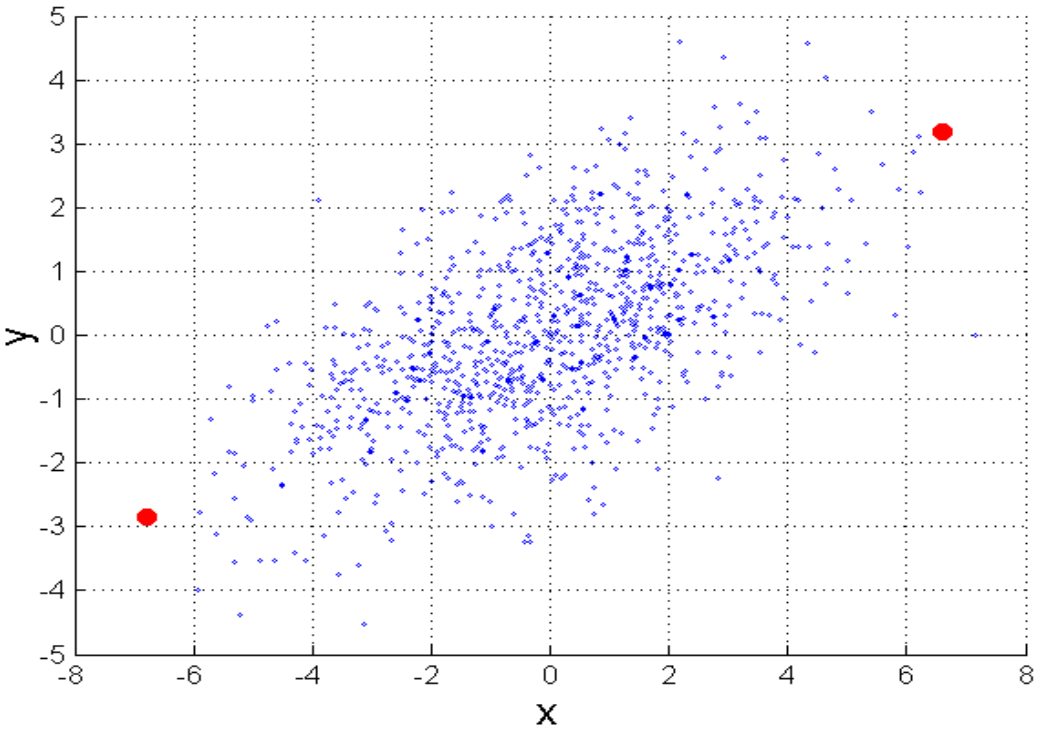
- Predstavlja uopštenje Euklidskog rastojanja koje se koristi kada postoji korelacija nekih atributa, uz eventualni dodatak u razlikama opsega vrednosti atributa.
- Mahanalobisovo rastojanje je korisno kada važi
  - atributi su u korelaciji
  - imaju različite opsege vrednosti (različite varijanse)
  - raspodela podataka je približno normalna (Gausova)

- Mahanalobisovo rastojanje dva objekta (vektora) p i q je

$$mahalanobis(p,q) = (p-q)\sum^{-1}(p-q)^T$$

gde je $\sum^{-1}$ inverzna matrica matrici kovarijansi podataka.

Mahanalobisovo rastojanje

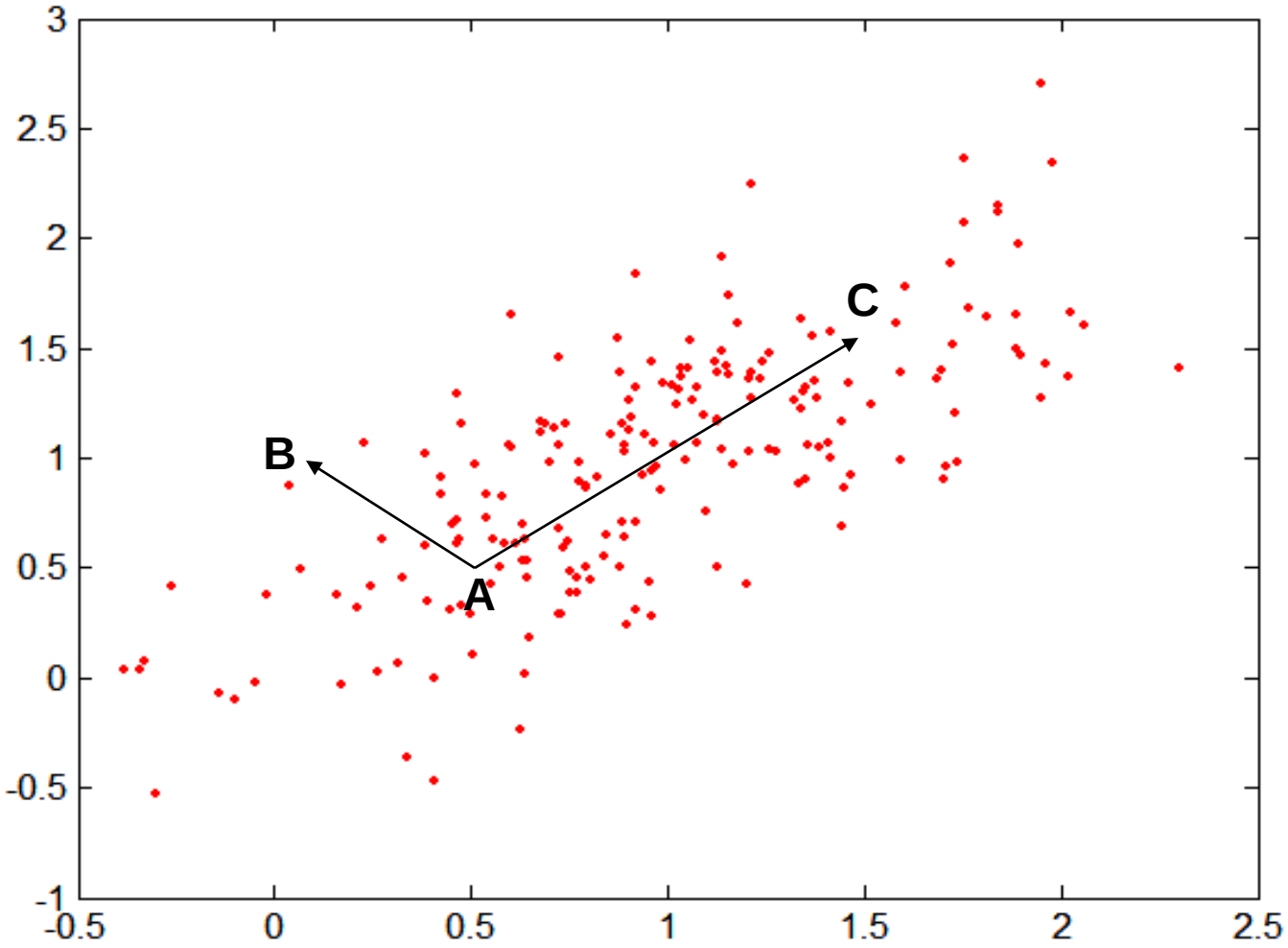


Σ je matrica kovarijansi za ulazni podataka *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

Euklidsko rastojanje crvene tačke je 14.7, a Mahalanobisovo rastojanje je 6

Mahanalobisovo rastojanje



Kovarijaciona matrica

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4