Zbirni izveštaj sa rešenjima domaćih zadataka iz prepoznavanja oblika (13E054PO)

Studentkinja:
Aleksandra Kojčinović 2018/0453 OS

Mentor:
Željko Đurović
Natalija Đorđević

Univerzitet u Beogradu – Elektrotehnički fakultet
septembar 2022. godine

Sadržaj

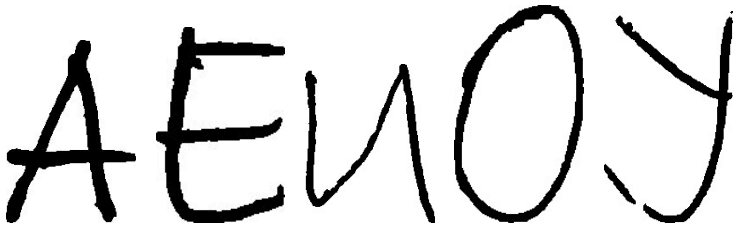
1 Zadatak 1

1.1 Tekst zadatka

Za bazu rukom pisanih samoglasnika, koja je dostupna na sajtu predmeta* projektovati inovativni sistem za prepoznavanje svih pet samoglasnika zasnovan na testiranju hipoteza. Rezultate prikazati u obliku matrice konfuzije. Izveštaj treba da sadrži kratki opis projektovanog sistema, obrazložen izbor obeležja, kao i karakteristične primere pravilno i nepravilno klasifikovanih slova.

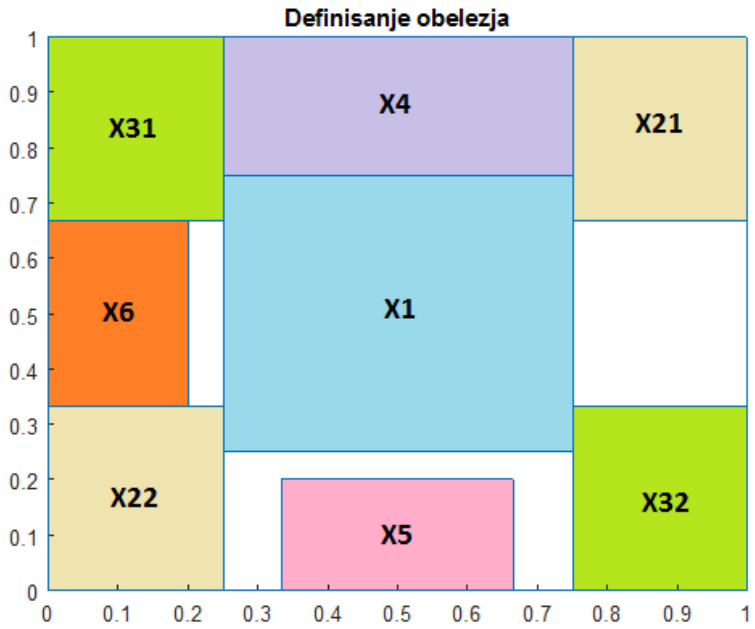
1.2 Rešenje

Za izradu zadatka potrebno je uraditi predobradu slike, zarad lakše ekstrakcije obeležja. Prvo je slika filtrirana Gausovim filtrom, kako bi se tamni regioni proširili, a zatim je odrađena popravka kontrasta metodom adaptizacije histograma. Nakon toga urađena je binarizacija slike tako da su svi pikseli čija je vrednost veća od 85% maksimuma proglašeni za bele, a svi ostali za crne. Takva slika je dalje podvrgnuta uklanjanju crnih okvira eliminacijom svih redova piksela u kojima ima manje od 78% belih piksela. Posle ove obrade potencijalno su ostali crni pikseli po ivicama slike koji mogu praviti problem pri kasnijoj obradi, tako da je nad slikom urađena erozija koja sve crne poiksele koji se nalaze u susedstvu nekih belih piksela proglašava takođe za bele. Nakon toga, uklonjeni su i beli regioni oko samih slova tako što su uklonjeni redovi piksela ukoliko imaju manje od 1.2% crnih piksela, ili njihovi susedni redovi ka centru slike imaju manje od toga. Konačno, zbog toga što je prethodno primenjena erozija i neke ivice su "nagrižene", u poslednjem koraku je primenjena dilatacija, te su svi beli pikseli koji se nalaze u crnom susedstvu proglašeni za crne. Takva slika je sačuvana pod nazivom originalne slike uz prefiks *"iseceno_ "*.



Slika 1: Prikaz slova nakon obrade

Nakon obrade slike, izvršena je ekstrakcija sledećih 6 obeležja:



Slika 2: Obeležja za klasifikaciju slova

Obeležje X_1 razdvaja slovo O od svih ostalih slova, ali usled raznolikosti unutar baze, to nije dovoljan kriterijum klasifikacije pa je potrebno dodatno definisati $X_2 = X_{21} - X22$, koje razlikuje neka A za koje je to obeležje negativno od nekih E ili I za koje je blisko nuli, $X_3 = X_{31} - X_{32}$, koje radi sličnu stvar za neka U. Pored toga, obeležje X_4 razlikuje A, E od I, U, dok obeležje X_5 razlikuje U, E od A, I. Na osnovu tih 5 obeležja dobija se klasifikacija O i U bez greške, dok se neki oblici E i I prepoznaju kao U pa je potrebno dodatno uvesti obeležje X_6 koje ta dva slova razlikuje od U.
Na osnovu matrice konfuzije ispostavlja se da je greška klasifikacije korišćenjem ovako definisanih obeležja 10%.

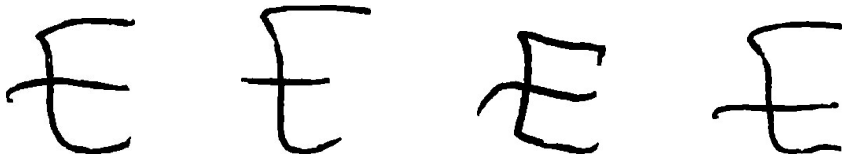
```
Matrica konfuzije:

Mk =

    22     0     1     0     1
     3    17     0     0     4
     0     0    21     0     3
     0     0     0    24     0
     0     0     0     0    24

Ukupna greska klasifikacije: 0.1
```

Slika 3: Matrice konfuzije za klasifikaciju slova

U nastavku su prikazana pogrešno klasifikovana slova.

Slika 4: A: prepoznato kao I (levo), prepoznato kao U (desno)

Slika 5: E prepoznato kao A



Slika 6: E prepoznato kao U

Slika 7: I prepoznato kao U

2 Zadatak 2

2.1 Tekst zadatka

Generisati po $N = 500$ odbiraka iz dveju dvodimenzionih bimodalnih klasa:

$$\Omega_1 : P_{11} \times N(M_{11}, \Sigma_{11}) + P_{12} \times N(M_{12}, \Sigma_{12})$$

$$\Omega_2 : P_{21} \times N(M_{21}, \Sigma_{21}) + P_{22} \times N(M_{22}, \Sigma_{22})$$

Parametre klasa samostalno izabrati.
a) Na dijagramu prikazati odbirke.
b) Generisati geometrijsko mesto tačaka sa konstantnom vrednošću funkcija gustina verovatnoće pa ih prikazati na dijagramu u prostoru oblika (za tri karakteristične vrednosti d^2 izabrane u skladu sa formulisanim klasama).
c) Projektovati Bajesov klasifikator minimalne greške i na dijagramu, zajedno sa odbircima, skicirati klasifikacionu liniju, pa proceniti verovatnoću greške.
d) Ponoviti prethodnu tačku za Neuman-Pearson-ov klasifikator. Obrazložiti izbor $\epsilon_2 = \epsilon_0$
e) Za klase oblika generisanih u prethodnim tačkama, projektovati Wald-ov sekvencijalni test pa skicirati zavisnost broja potrebnih odbiraka od usvojene verovatnoće grešaka prvog, odnosno drugog tipa.

2.2 Rešenje

a) Generisani su odbirci bimodalnih klasa sa sledećim parametrima:

$$P_{11} = 0.6;\ M_{11} = [1; 1];\ \Sigma_{11} = [4,\ 1.1; 1.1,\ 2];$$

$$P_{12} = 0.4;\ M_{12} = [6; 4];\ \Sigma_{12} = [3, 0.8; -0.8,\ 1.5];$$

$$P_{21} = 0.55;\ M_{21} = [9; -1];\ \Sigma_{21} = [2,\ 1.1; 1.1,\ 4];$$

$$P_{22} = 0.45;\ M_{22} = [2; -2];\ \Sigma_{22} = [3,\ 0.8; 0.8,\ 0.5];$$

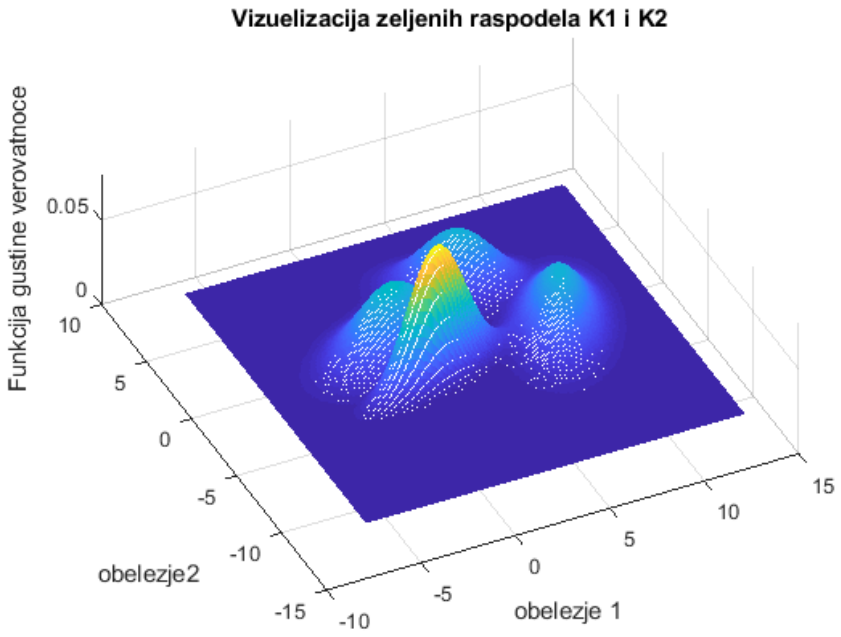
Pošto je generisan jednak broj odbiraka iz obe klase, apriorna verovatnoća pojava klasa je jednaka i iznosi 0.5. Generisanje podataka je izvršeno primenom transformacije bojenja na slučajne vektore Gausove raspodele jedinične varijanse i nultog matematičkog očekivanja. Matrica transformacije je oblika

$$T = \Phi \times \lambda^{1/2}.$$

gde je Φ matrica sopstvenih vektora kovarijacione matrice Σ, a λ matrica sopstvenih vrednosti dobijenih rešavanjem jednačine:

$$det|\Sigma - \lambda \times I| = 0.$$

Funkcija T pomnožena je vrednošću Gausove slučajne promenljive, čime je slučajna promenljiva prevedena iz domena jedinične kovarijacione matrice u domen željene matrice Σ, a zatim joj je dodat željeni vektor srednjih vrednosti. Prikaz funkcije gustine verovatnoće za ovako definisane klase dat je na slici 8.

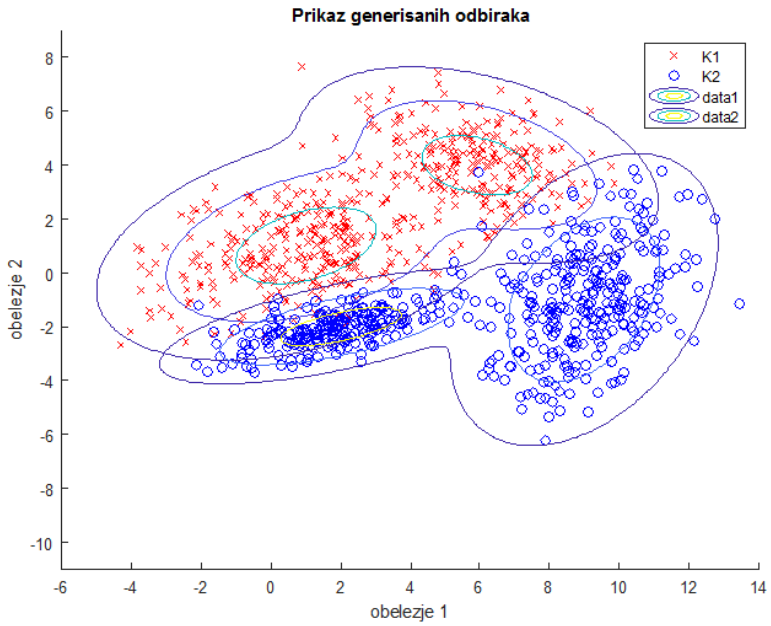


Slika 8: 3D prikaz funkcija gustina verovatnoće klasa

b) Geometrijsko mesto tačaka konstantne vrednosti funkcije gustine verovatnoće generisanih odbiraka naziva se d^2 kriva i predstavlja statističko odstojanje od centra određene klase. Računa se kao:

$$d^2(x) = (x - M)^T \Sigma_x^{-1} (x - M).$$

Generisani odbirci obe klase kao i d^2 krive za vrednosti $d^2 = [1,\ 4,\ 9]$ priloženi su na slici 9.



Slika 9: Odbirci generisanih klasa

c) Bajesov klasifikator minimalne verovatnoće greške se zasniva na testiranju hipoteza i povlačenju klasifikacione linije tako da se ostvari minimalna moguća verovatnoća greške klasifikacije u slučaju postojanja samo dve klase. Za ovu procedutu neophodno je poznavati parametre podataka nad kojima se vrši analiza, kao i raspodelu kojom su podaci opisani. U skladu sa tim, računa se aposteriorna verovatnoća pojave određene klase, ukoliko je dobijena konkretna opservacija i može se zapisati na sledeći način:

$$q_i(X) = Pr(\omega_i/X) = \frac{P_i \times f(X/\omega_i)}{f(X)}$$

gde je P_i apriorna verovatnoća pojave klase i, $f(X/\omega_i)$ funkcija gustine verovatnoće za odbirke iz klase i, a $f(X)$ funkcija gustine verovatnoće oblika X koja se može napisati u formi totalne verovatnoće:

$$f(X) = P_1 f_1(X) + P_2 f_2(X) + ... + P_L f_L(X)$$

za $f_i(X) = f(X/\omega_i)$
Dalje se vrši provera koja od aposteriornih verovatnoća ima veću vrednost za konkretno merenje i to merenje se dodeljuje klasi kojoj odgovara veća vrednost funkcije. Opisani postupak ima sledeću formu:

$$q_1(X) > q_2(X) \Rightarrow X \in \omega_1$$
$$q_1(X) < q_2(X) \Rightarrow X \in \omega_2$$

dalje se može pisati

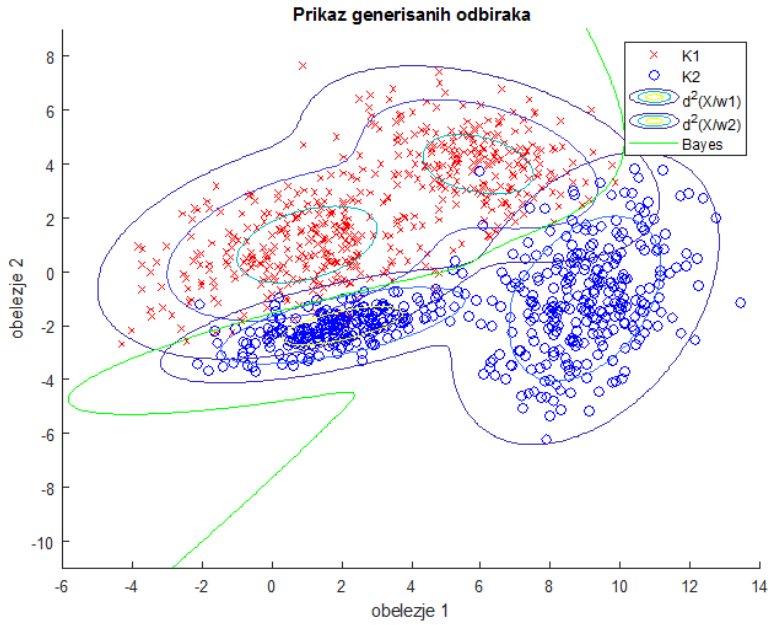
$$P_1 f_1(X) > P_2 f_2(X) \Rightarrow X \in \omega_1$$
$$\frac{f_1(X)}{f_2(X)} > \frac{P_2}{P_1} \Rightarrow X \in \omega_1.$$

Uobičajeno se na levu stranu izraza primeni funkcija negativnog prirodnog logaritma i dobija konačna forma:

$$-ln(\frac{f_1(X)}{f_2(X)}) < -ln(\frac{P_2}{P_1}) \Rightarrow X \in \omega_1$$

što se kraće zapisuje kao:

$$h(X) < T \Rightarrow X \in \omega_1$$
$$h(X) > T \Rightarrow X \in \omega_2$$



Slika 10: Izgled Bajesove klasifikacione krive

Ukupna verovatnoća pogrešne klasifikacije:

$$\epsilon = P_1\epsilon_1 + P_2\epsilon_2$$

gde je ϵ_1 verovatnoća greške prvog tipa i predstavlja verovatnoću da se odbirak iz prve klase klasifikuje kao da je iz druge, a ϵ_2 verovatnoća greške drugog tipa, pri klasifikaciji odbiraka druge klase.

$$\epsilon_1 = \int_{L2} f_1(X)dx, \quad \epsilon_2 = \int_{L1} f_2(X)dx$$

Integral se u realnim primenama najčešće aproksimira trapeznim metodom računanja površine ispod krive ili u 2D slučaju računanjem zapremine ispod funkcije gustine verovatnoće. Za konkretan slučaj, izračunata je teorijska vrednost verovatnoće greške na pomenut način, kao i konkretna greška pri klasifikaciji generisanjem matrice konfuzije dimenzija 2×2, gde se na glavnim dijagonalama nalazi broj ispravno klasifikovanih odbiraka, a na sporednoj broj pogrešno klasifikovanih odbiraka koji zaista pripadaju onoj klasi u kojoj se koloni nalaze.

```
c) Bajesov test:
Matrica konfuzije:

cm =

   487    13
    13   487

Procena verovatnoce greske na osnovu matrice konfuzije:
Greska I tipa: 0.026
Greska II tipa: 0.026
Ukupna greska: 0.026
-------------------------------------
Teorijska procena verovatnoce greske:
Verovatnoca greske I tipa: 0.042915
Verovatnoca greske II tipa: 0.022708
Ukupna verovatnoca greske: 0.032811
-------------------------------------
```

Slika 11: Računanje verovatnoće greške

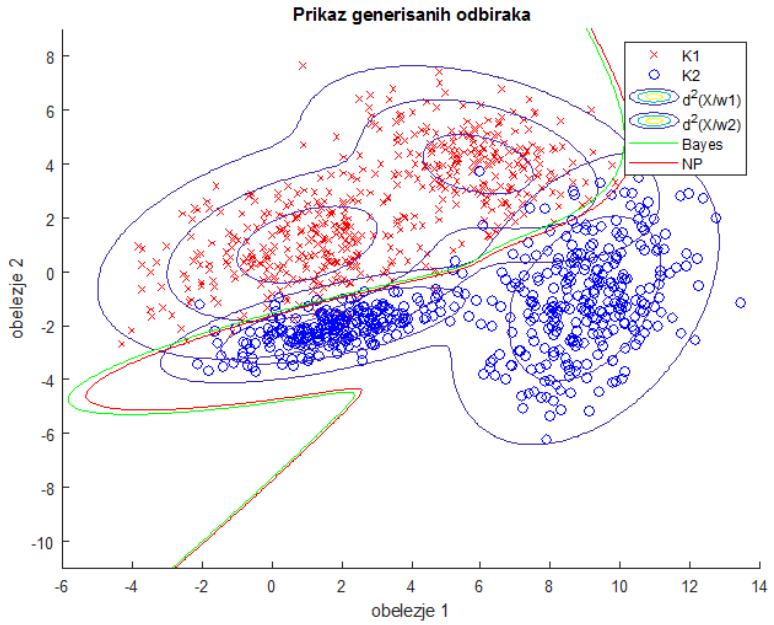
d) Neuman - Pearsonov klasifikator je još jedna metoda klasifikacije na bazi testiranja hipoteza, a zasniva se na fiksiranju vrednosti jednog tipa greške na vrednost ϵ_0 i minimizaciji drugog tipa greške. Kriterijumska funkcija koja se minimizira ima oblik:

$$J = \epsilon_1 + \mu(\epsilon_2 - \epsilon_0)$$

gde je μ Lagranžeov multiplikator. Test se svodi na sledeći oblik:

$$\frac{f_q(X)}{f_2(X)} > \mu \rightarrow -ln(\frac{f_1(X)}{f_2(X)} < -ln(\mu)) \rightarrow h(X) < T \Rightarrow X \in \omega_1$$

$$\frac{f_q(X)}{f_2(X)} < \mu \rightarrow -ln(\frac{f_1(X)}{f_2(X)} > -ln(\mu)) \rightarrow h(X) > T \Rightarrow X \in \omega_2$$

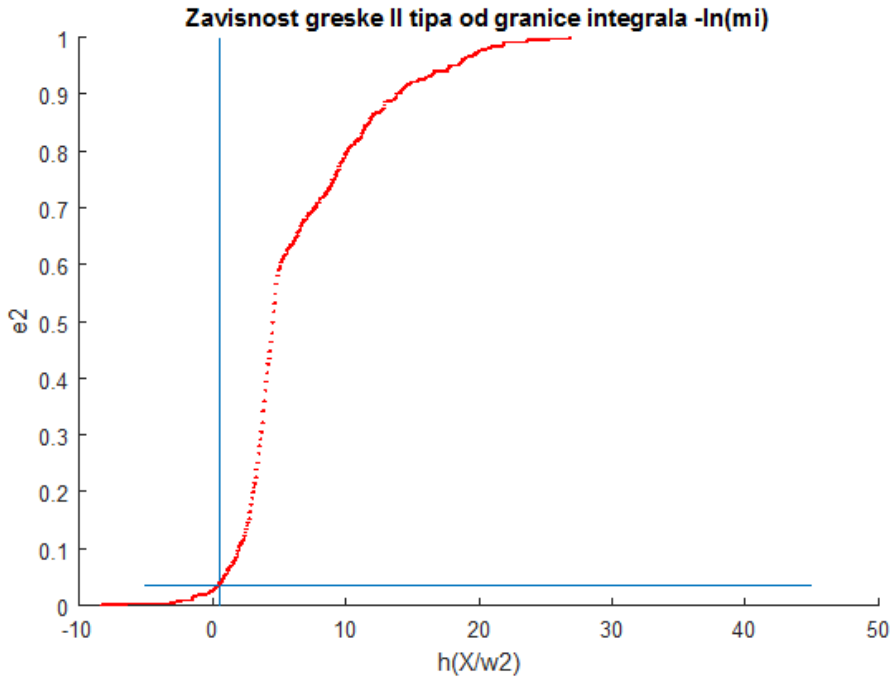


Slika 12: Prikaz NP klasifikacione krive

Parametar μ se za poznato ϵ_0 može dobiti iz definicionog izraza za grešku drugog tipa, koji se može napisati u domenu diskriminacione funkcije $h(X)$ kao:

$$\epsilon_2 = \int_{-\infty}^{-ln(\mu)} f_h(h/\omega_2)dh$$

gde je $f_h(h/\omega_2)$ uslovna gustina verovatnoće vrednsoti diskriminacione funkcije za odbirke druge klase, koja je zapravo jedna slučajna promenljiva. Funkcija gustine verovatnoće funkcije $h(X)$ se dobija metodom procene histograma, a dalje se numeričkim rešavanjem gornjeg integrala i traženja vrednosti μ za koje je on jednak ϵ_0.



Slika 13: Odabir praga μ za željenu verovatnoću greške

Prateći literaturu sa predavanja, odabrana je vrednost ϵ_0 koja je jednaka Bajesovoj proceni ukupne verovatnoće greške iz prethodne tačke, jer je to minimalna teorijski ostvariva greška. Nakon iscrtavanja klasifikacione linije, računaju se teorijska vrednost verovatnoće greške i matrica konfuzije kao u prethodnoj tački.

```
d) NP test:
Konstantna greska II tipa(false-alarm):0.032811
e0=0.034
Prag T=0.46771 mi=0.62643
Matrica konfuzije:

cm =

   490    17
    10   483

Procena verovatnoce greske na osnovu matrice konfuzije:
Greska I tipa: 0.02
Greska II tipa: 0.034
Ukupna greska: 0.027
-----------------------------------
Teorijska procena verovatnoce greske:
Verovatnoca greske I tipa: 0.03514
Verovatnoca greske II tipa: 0.032527
Ukupna verovatnoca greske: 0.033833
```

Slika 14: Računanje verovatnoće greške

d) U slučaju da su opservacije dostupne sekvencijalno, a ne sve odjednom, pribegava se sekvencijalnom testiranju hipoteza, gde se za odbirke pristigle do trenutka m računa združena funkcija gustine verovatnoće (posmatrajući ih kao nezavisne slučajne promenljive) za obe klase, traži se količnik te dve vrednosti i na njega primenjuje negativan logaritam.

$$S_m = -ln(\frac{f_1(X_1, X_2...X_m)}{f_2(X_1, X_2...X_m)}) = -ln(\frac{\prod_{i=1}^{m} f_1(X_i)}{\prod_{i=1}^{m} f_2(X_i)}) = \sum_{i=1}^{m} h(X_i)$$

U zavisnosti od vrednosti funkcije S_m donosi se odluka kojoj klasi pripadaju opservirani odbirci i procedura se završava ili se uzima naredni odbirak i ponovo procenjuje funkcija. Opis algoritma dat je u nastavku:

$$S_m \geq a \Rightarrow X \in \omega_1$$

$$b < S_m < a \Rightarrow uzeti\ odbirak\ m+1$$

$$S_m \leq b \Rightarrow X \in \omega_2$$

Parametri a i b utiču na verovatnoće greške prvog i drugog tipa, ali i na potreban broj odbiraka za donošenje odluke. Izbor parametara a i b predložio je naučnik Wald, a zasniva se na računanju količnika verodostojnosti pri pristizanju m-tog merenja i poređenjem sa pragom.

$$\frac{f_1(X_1...X_m)}{f_2(X_1...X_m)} \leq A \Rightarrow X \in \omega_1$$

$$\frac{f_1(X_1...X_m)}{f_2(X_1...X_m)} \geq B \Rightarrow X \in \omega_2$$

To se može napisati u domenu verovatnoća greške:

$$1 - \epsilon_1 \geq A\epsilon_2 \Rightarrow A \leq \frac{1-\epsilon_1}{\epsilon_2}$$

$$\epsilon_1 \leq B(1-\epsilon_2) \Rightarrow B \geq \frac{\epsilon_1}{1-\epsilon_2}$$

odakle se parametri a i b za usvojeni znak jednakosti svode na:

$$a = -lnA = -ln(\frac{1-\epsilon_1}{\epsilon_2})$$

$$b = -lnB = -ln(\frac{\epsilon_1}{1-\epsilon_2})$$

Tako se za poznate parametre A i B mogu izraziti verovatnoće greške oba tipa:

$$\epsilon_1 \cong \frac{B(A-1)}{A-B}$$

$$\epsilon_2 \cong \frac{1-B}{A-B}$$

Za ovako definisan Wald-ov sekvencijalni test važe sledeće osobine:
1. Poslednji izraz za vezu ϵ_1 i ϵ_2 sa parametrima A i B važi i ukoliko odbirci merenja nisu nezavisni i jednako raspodeljeni.
2. Wald-ov sekvencijalni test se završava sa verovatnoćom 1, tačnije događaj u kome broj odbiraka potrebnih za donošenje odluke neograničeno raste, a S_m ne dostiže granice ima verovatnoću 0.
3. Wald-ov sekvencijalni test minimizira broj potrebnih odbiraka za donošenje odluke za poznato ϵ_1 i ϵ_2.
Srednji broj potrebnih merenja može se izraziti iz sledećih izraza:

$$E\{S/\omega_1\} = E\{\sum_{i=1}^{m} h(X_i)\} = m\eta_1$$

$$E\{S/\omega_2\} = E\{\sum_{i=1}^{m} h(X_i)\} = m\eta_2$$

gde je $\eta_i = E\{h(X/\omega_i)\}$
Za funkciju S_m mogući su ishodi:

$$S_m = a,\ Pr = (1-\epsilon_1)\ za\ X \in \omega_1$$

$$S_m = a,\ Pr = \epsilon_2\ za\ X \in \omega_2$$
$$S_m = b,\ Pr = \epsilon_1\ za\ X \in \omega_1$$
$$S_m = b,\ Pr = (1-\epsilon_2)\ za\ X \in \omega_2$$

Odakle je:

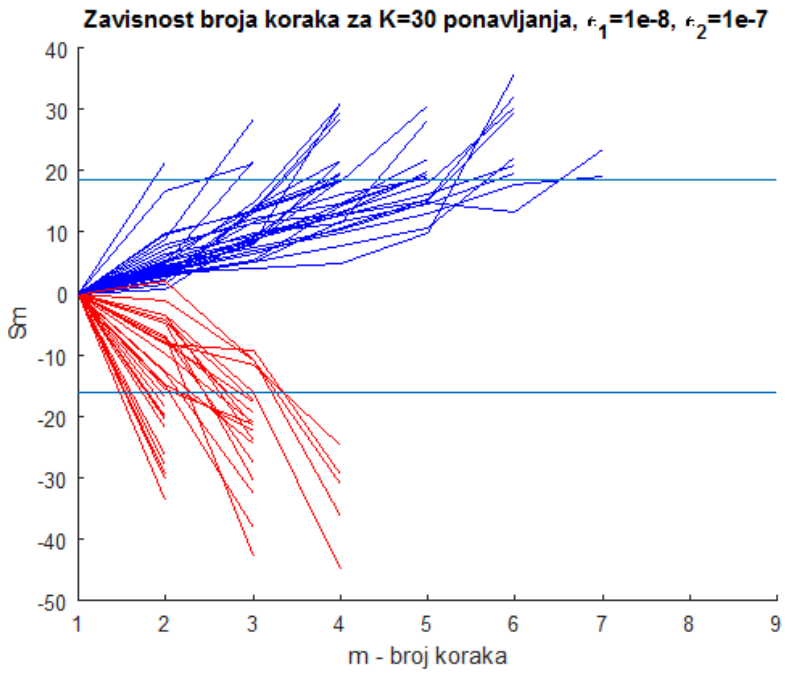
$$E\{S/\omega_1\} = a(1-\epsilon_1) + b\epsilon_1$$
$$E\{S/\omega_1\} = a\epsilon_2 + b(1-\epsilon_2)$$

Kombinovanjem sa gornjim izrazima, dobija se:

$$E\{m/\omega_1\} = \frac{a(1-\epsilon_1) + b\epsilon_1}{\eta_1}$$

$$E\{m/\omega_2\} = \frac{a\epsilon_2 + b(1-\epsilon_2)}{\eta_2}$$

što predstavlja srednji broj potrebnih merenja da bi se donela odluka u slučaju klase 1, odnosno klase 2.
Uticaj vrednosti funkcije S_m na brzinu donošenja odluke dat je na slici 15. Analizom zavisnosti broja koraka od jednog tipa greške, dok je verovatnoća drugog tipa greške fiksirana, primećuje se da je broj potrebnih odbiraka pri klasifikaciji oblika iz prve klase linearno zavisan od vrednosti verovatnoće greške drugog tipa, dok za promenu vrednosti verovatnoće greške prvog tipa, broj odbirka ima skoro konstantnu vrednost. Isto važi i u slučaju klasifikacije oblika druge klase, gde je zavisnost od verovatnoće greške prvog tipa opadajuća, dok je zavisnost od verovatnoće greške drugog tipa konstantna.



Slika 15: Zavisnost kriterijuma od broja odbiraka

3 Zadatak 3

3.1 Tekst zadatka

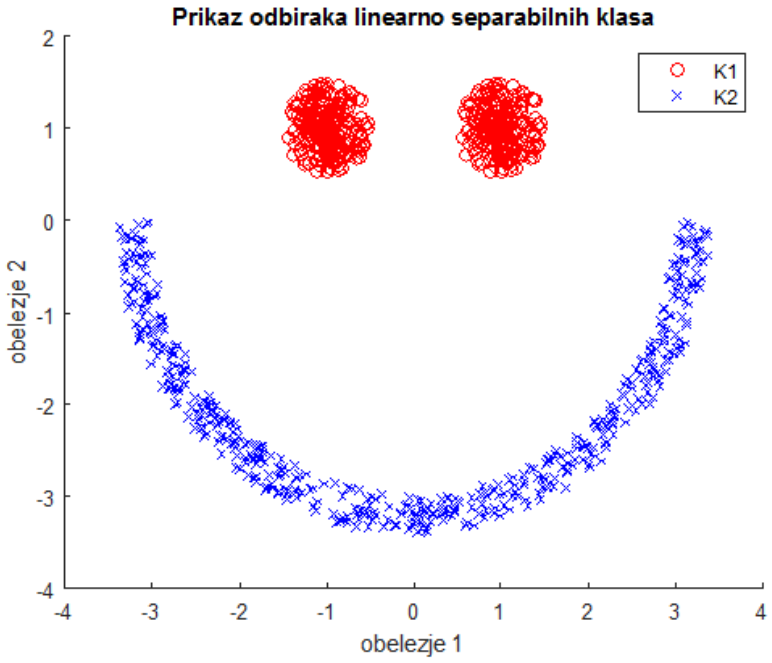
1. Generisati dve klase dvodimenzionalnih oblika. Izabrati funkciju gustine verovatnoće oblika tako da klase budu linearno separabilne.
a) Za tako generisane oblike izvšiti projektovanje linearnog klasifikatora jednom od tri iterativne procedure.
b) Ponoviti prethodni postupak korišćenjem metode željenog izlaza. Analizirati uticaj elemenata u matrici željenih izlaza na konačnu formu linearnog klasifikatora.
2. Generisati dve klase dvodimenzionalnih oblika koje jesu seperabilne, ali ne linearno pa isprojektovati kvadratni klasifikator metodom po želji.

3.2 Rešenje

1. Generisano je po $N = 500$ odbiraka dve linearno separabilne klase kao na slici 16.
a) Projektovan je linearni klasifikator, jedan od metoda parametarske klasifikacije zasnovan na proceni statističkih parametara iz uzorka.
Linearni klasifikator može se izraziti u formi:

$$h(X) = V^T X + v_0$$

gde V predstavlja pravac na koji se projektuju realizacije vektora X, a v_0 granica između oblika dveju klasa na tom pravcu. Ključni zadatak u projektovanju je određivanje takvog pravca V na kome su projekcije odbiraka dveju klasa separabilne.
Ovaj klasifikator ne zahteva poznavanje uslovnih funkcija gustine verovatnoće, ali je pogodan samo za klasifikaciju linearno separabilnih oblika čija se separabilnost ogleda u rastojanju matematičkih očekivanja. Takođe, daje optimalne rezultate i u slučaju da vektor X nije Gausovski raspodeljen, jer se može smatrati da sa povećanjem dimenzionalnosti vektora X, realizacije



Slika 16: Generisani odbirci linearno separabilnih klasa

slučajne promenljive $h(X)$, koja predstavlja zbir n slučajnih promenljivih, teže Gausovoj raspodeli. Pri projektovanju linearnog klasifikatora uzimaju se u obzir sledeći parametri:

$$\eta_1 = E\{h(X)/\omega_1\} = E\{V^T X + v_0/\omega_1\} = V^T M_1 + v_0$$

$$\eta_2 = E\{h(X)/\omega_2\} = E\{V^T X + v_0/\omega_2\} = V^T M_2 + v_0$$

$$\sigma_1^2 = var\{h(X)/\omega_1\} = var\{V^T X + v_0/\omega_1\} = V^T \Sigma_1 V$$

$$\sigma_2^2 = var\{h(X)/\omega_2\} = var\{V^T X + v_0/\omega_2\} = V^T \Sigma_2 V$$

na osnovu kojih se formira kriterijumska funkcija koja obezbeđuje maksimalno rastojanje između matematičkih očekivanja dveju klasa, u odnosu na njihova rasipanja. Za uopšteni oblik kriterijumske funkcije $f(\eta_1, \eta_2, \sigma_1^2, \sigma_2^2)$, optimalan rezultat se dobija iz jednačina:

$$\frac{\partial f}{\partial V} = \frac{\partial f}{\partial \eta_1}\frac{\partial \eta_1}{\partial V} + \frac{\partial f}{\partial \eta_2}\frac{\partial \eta_2}{\partial V} + \frac{\partial f}{\partial \sigma_1^2}\frac{\partial \sigma_1^2}{\partial V} + \frac{\partial f}{\partial \sigma_2^2}\frac{\partial \sigma_2^2}{\partial V} = 0$$

$$\frac{\partial f}{\partial v_0} = \frac{\partial f}{\partial \eta_1}\frac{\partial \eta_1}{\partial v_0} + \frac{\partial f}{\partial \eta_2}\frac{\partial \eta_2}{\partial v_0} + \frac{\partial f}{\partial \sigma_1^2}\frac{\partial \sigma_1^2}{\partial v_0} + \frac{\partial f}{\partial \sigma_2^2}\frac{\partial \sigma_2^2}{\partial v_0} = 0$$

uz uslove:

$$\frac{\partial \eta_i}{\partial V} = M_i, \ \frac{\partial \sigma_i^2}{\partial V} = 2\Sigma_i V, \ \frac{\partial \eta_i}{\partial v_0} = 1, \ \frac{\partial \sigma_i^2}{\partial V} = 0$$

Daljim rešavanjem napisanih jednačina po V, dobija se optimalan pravac koji minimizira funkciju f:

$$V = (s\Sigma_1 + (1-s)\Sigma_2)^{-1}(M_2 - M_1)$$

gde je

$$s = \frac{\frac{\partial f}{\partial \sigma_1^2}}{\frac{\partial f}{\partial \sigma_1^2} + \frac{\partial f}{\partial \sigma_2^2}}$$

Parametar v_0 dobija se iz

$$\frac{\partial f}{\partial \eta_1} + \frac{\partial f}{\partial \eta_2} = 0.$$

Dodatno, ukoliko je raspodela kriterijumske funkcije normalna (ako je X normalno raspodeljen vektor ili mu je dimenzija dovoljno velika da važi centralna granična teorema), moguće je odrediti V i v_0 tako da minimiziraju verovatnoću greške. Na osnovu pravila klasifikacije

$$V^T X + v_0 < 0 \Rightarrow \ X \in \omega_1$$

$$V^T X + v_0 > 0 \Rightarrow \ X \in \omega_2$$

može se definisati verovatnoća greške

$$\epsilon = P_1\epsilon_1 + P_2\epsilon_2 = P1 \int_0^{\infty} f(h/\omega_1)dh + P2 \int_{-\infty}^{0} f(h/\omega_2)dh$$

$$= P_1 \int_0^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{(h-\eta_1)^2}{\sigma_1^2}} dh + P_2 \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\frac{(h-\eta_2)^2}{\sigma_2^2}} dh$$

$$= P_1 \int_{-\frac{\eta_1}{\sigma_1}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\zeta^2}{2}} d\zeta + P_1 \int_{-\infty}^{-\frac{\eta_2}{\sigma_2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\zeta^2}{2}} d\zeta$$

Sada je ϵ kriterijum koji treba optimizovati. Dodatno, za diferenciranje integralnog izraza po promenljivoj koja se nalazi u granicama integrala koristi se:

$$\frac{d}{d\alpha} \int_{f(\alpha)}^{g(\alpha)} h(X)dx = h(g(\alpha))\frac{dg(\alpha)}{d\alpha} - h(f(\alpha))\frac{df(\alpha)}{d\alpha}$$

a iz ranije izvedenih jednačina, optimizacija se svodi na:

$$\frac{\partial \epsilon}{\partial \eta_1} = \frac{P_1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}(\frac{\eta_1}{\sigma_1})^2}$$

$$\frac{\partial \epsilon}{\partial \eta_2} = -\frac{P_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}(\frac{\eta_2}{\sigma_2})^2}$$

$$\frac{\partial \epsilon}{\partial \sigma_1^2} = -\frac{P_1}{\sqrt{2\pi}}(-\frac{\eta_1}{2\sigma_1^3}) e^{-\frac{1}{2}(\frac{\eta_1}{\sigma_1})^2}$$

$$\frac{\partial \epsilon}{\partial \sigma_2^2} = \frac{P_2}{\sqrt{2\pi}}(-\frac{\eta_2}{2\sigma_2^3}) e^{-\frac{1}{2}(\frac{\eta_2}{\sigma_2})^2}$$

Iz uslova

$$\frac{\partial \epsilon}{\partial \eta_1} + \frac{\partial \epsilon}{\partial \eta_2} = 0 \ \Rightarrow \ \frac{P_1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}(\frac{\eta_1}{\sigma_1})^2} = \frac{P_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}(\frac{\eta_2}{\sigma_2})^2}$$

sledi da v_0 treba da omogući

$$f(h/\omega_1)/f(h/\omega_2) = P_2/P_1$$

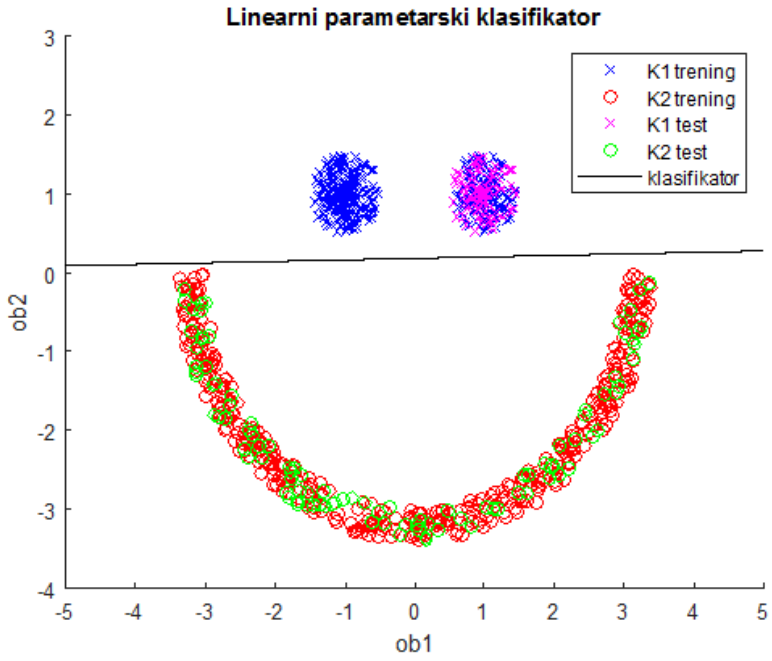
Finalno, dobijaju se relacije:

$$s = \frac{-\eta_1/\sigma_1^2}{-\eta_1/\sigma_1^2 + \eta_2/\sigma_2^2} \ \in \ [0,1]$$

$$V = (s\Sigma_1 + (1-s)\Sigma_2)^{-1}(M_2 - M_1)$$

$$v_0 = \frac{s\sigma_1^2 V^T M_2 + (1-s)\sigma_2^2 V^T M_1}{s\sigma_1^2 + (1-s)\sigma_2^2}$$

Pošto su određene promenjive međusobno zavisne, ne može se pronaći eksplicitno rešenje nego se pribegava iterativnim procedurama rešavanja. Pri izradi ovog zadatka, korišćena je treća iterativna procedura (hold-out metod) za rešavanje koja se svodi na sledeće korake:
1. Podaci se podele na obučavajući i testirajući skup, recimo tako da po 80% iz svake od klasa pripadne obučavajućem, a preostalih 20% testirajućem skupu.
2. Nad obučavajućim skupom se odredi srednja vrednost M_i i kovarijaciona matrica Σ_i za svaku od klasa.
3. Iterativno se menja parametar s u opsegu $[0, 1]$ sa određenim korakom
4. Za konkretno s se izračuna V
5. Na osnovu dobijenog V generiše se projekcija $Y_i^k = V^T X_i^k,\ i = 1, ..., N_k, k = 1...L$, gde X predstavlja vektor obeležja iz testirajućeg skupa.
6. Vrši se iteracija po v_0 od $v_{min} = -max(max(Y^{(1)}), max(Y^{(2)}))$ do $v_{max} = -min(min(Y^{(1)}), min(Y^{(2)}))$ sa korakom $\Delta v_0 = \frac{Y_i + Y_{i+1}}{2}$ i računa broj pogrešno klasaifikovanih odbiraka N_ϵ za takav odabir V i v_0
7. Kao optimalno v_0 usvaja se ono koje dovodi do minimalnog N_ϵ za jednu iteraciju po s. Takvih $N_\epsilon(s)$ ima koliko i promenljivih s.
8. Za optimalno s usvaja se ono koje odgovara minimumu $N_\epsilon(s)$
Rezultati su prikazani na slici 17.



Slika 17: Rezultati klasifikacije

b) Ukoliko se uslov linearnog klasifikatora napiše u formi:

$$-V^T X - v_0 > 0 \ \Rightarrow X \in \omega_1$$

$$+V^T X + v_0 > 0 \ \Rightarrow X \in \omega_2$$

mogu se definisati matrice

$$W^T = [v_0\ V^T];\ Z = [-1; -X] \rightarrow W^T Z > 0 \Rightarrow X \in \omega_1$$

$$W^T = [v_0\ V^T];\ Z = [+1; +X] \rightarrow W^T Z > 0 \Rightarrow X \in \omega_2$$

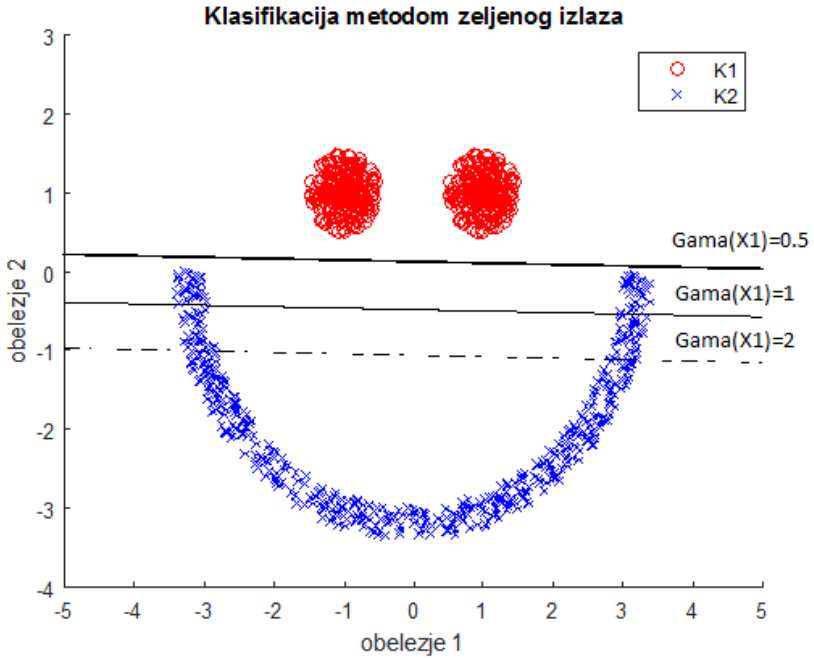
a svi oblici Z se grupišu u matricu oblika $U = [Z_1 ... Z_{N_1+N_2}]$. Dalje se definiše matrica Γ u kojoj se nalaze željene vrednosti sa desne strane nejednakosti i može se pisati izraz:

$$U^T W = \Gamma.$$

Sada su od interesa parametri V i v_0 iz matrice W, pa se za nalaženje W primenjuje metod najmanjih kvadrata, jer U^T nije kvadratna matrica. Finalno se rešava jednačina:

$$W = (UU^T)^{-1} U\Gamma$$

Rezultati ovakve analize dati su u nastavku za različite vrednosti u matrici željenih izlaza Γ. Ukoliko se u prvi deo matrice (onaj koji se odnosi na klasu 1) upišu veće vrednosti, klasifikaciona linija se pomera bliže oblicima druge klase, praveći manju grešku pri klasifikaciji oblika iz prve klase. Sa druge strane, ukoliko se u drugi deo matrice Γ upišu veće vrednosti, pravi se manja greška pri klasifikaciji oblika iz druge klase.



Slika 18: Klasifikacija metodom željenih izlaza

2. Generisano je po $N = 500$ odbiraka dve nelinearno separabilne klase kao na slici 19. Projektovan je kvadratni klasifikator forme:

$$h(X) = X^T Q X + V^T X + v_0 < 0 \Rightarrow X \in \omega_1$$

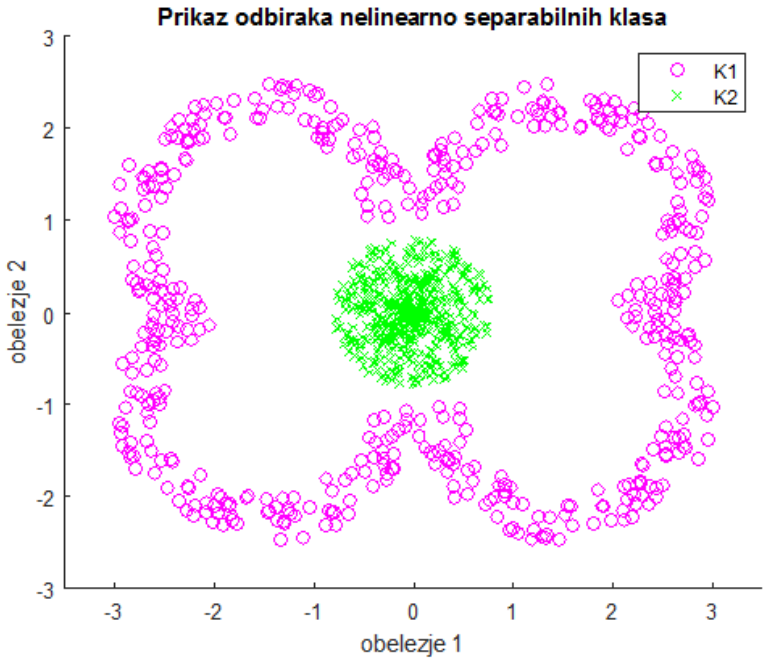
$$h(X) = X^T Q X + V^T X + v_0 > 0 \Rightarrow X \in \omega_2$$

Ukoliko se diskriminaciona funkcija napiše u obliku

$$h(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} x_i x_j + \sum_{i=1}^{n} v_i x_i + v_0$$

i koeficijenti se grupišu u matricu K, a svi oblici u matricu D

$$K^T = [];\ D = []$$



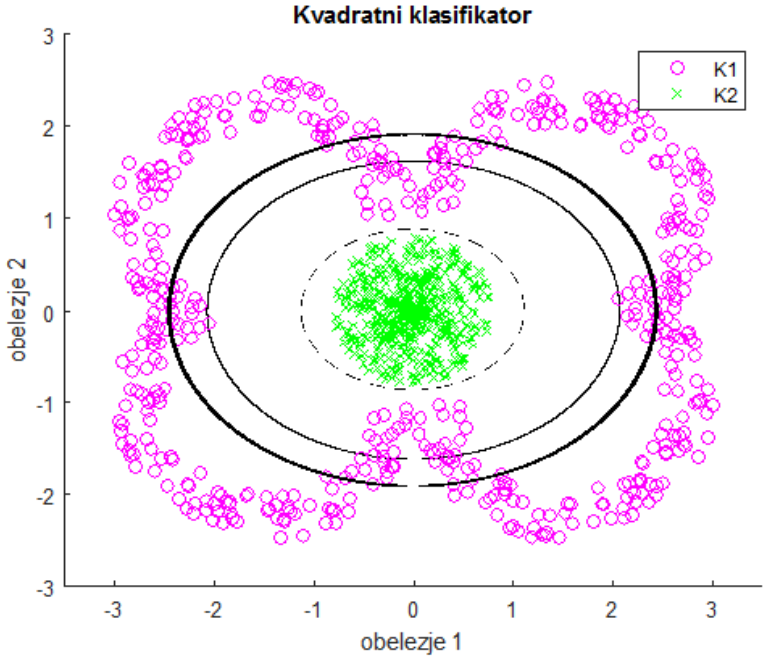
Slika 19: Odbirci nelinearno separabilnih klasa

problem se prividno svodi na linearnu formu

$$K^T D < 0 \Rightarrow X \in \omega_1$$

$$K^T D > 0 \Rightarrow X \in \omega_2$$

i ponovo se može definisati vektor željenih izlaza Γ. U zavisnosti od vrednosti unutar tog vektora, pomera se i klasifikaciona kriva. U konkretnom slučaju, sa povećanjem vrednosti koje se odnose na prvu klasu, kružnica se sužava.



Slika 20: Rezultati kvadratne klasifikacije, $\Gamma(K1) = \{0.5; 1; 4\}$

4 Zadatak 4

4.1 Tekst zadatka

1. Generisati po $N = 500$ dvodimenzionih odbiraka iz četiri klase koje će biti linearno separabilne. Preporuka je da to budu Gausovski raspodeljeni dvodimenzioni oblici.
Izabrati jednu od metoda za klasterizaciju (c mean metod, metod kvadratne dekompozicije) i primeniti je na formirane uzorke klasa. Izvršiti analizu osetljivosti izabranog algoritma na početnu klasterizaciju kao i srednji broj potrebnih iteracija. Takođe izvršiti analize slučaja kada se apriorno ne poznaje broj klasa.
2. Na odbircima iz prethodne tačke izabrati jednu od metoda klasterizacije (metod maksimalne verodostojnosti ili metod grana i granica) i primeniti je na formirane uzorke klasa. Izvršiti analizu osetljivosti izabranog algoritma na početnu klasterizaciju kao i srednji broj potrebnih iteracija. Takođe izvršiti analize slučaja kada se apriorno ne poznaje broj klasa.
3. Generisati po $N = 500$ dvodimenzionih odbiraka iz dve klase koje su nelinearno separabilne. Izabrati jednu od metoda za klasterizaciju koje su primenjive za nelinearno separabilne klase (metod kvadratne dekompozicije ili metod maksimalne verodostojnosti) i ponoviti analizu iz prethodnih tačaka.

4.2 Rešenje

1. Generisano je po $N = 500$ odbiraka iz četiri klase, Gausove raspodele parametara:

$$M_1 = [1; -2];\ \Sigma_1 = [0.4,\ 0.3; 0.3,\ 0.6]$$

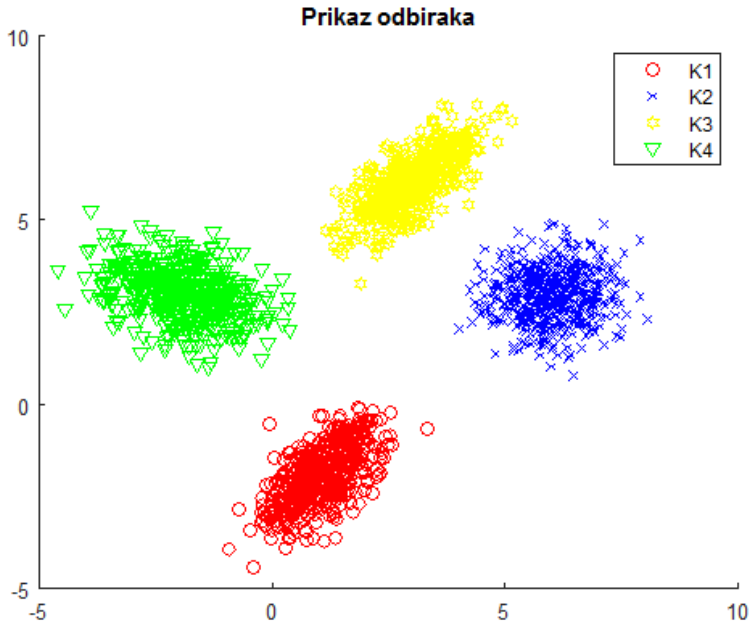
$$M_2 = [6; 3];\ \Sigma_2 = [0.5,\ 0; 0,\ 0.5]$$

$$M_3 = [3; 6];\ \Sigma_3 = [0.5,\ 0.4; 0.4,\ 0.6]$$

$$M_4 = [-2; 3];\ \Sigma_4 = [0.8,\ -0.2; -0.2,\ 0.5]$$

Verovatnoća pojava klasa je jednaka i iznosi $P_i = 0.25,\ i = 1, 2, 3, 4$.
Primenjena je transformacija bojenja kao u zadatku 2, i dobijeni su odbirci klasa prikazani na

slici 21.



Slika 21: Generisani odbirci linearno separabilnih klasa

Izabran je c-mean metod klasterizacije, koji se zasniva na klasterizaciji pojedinačnih odbiraka prema najbližoj srednjoj vrednosti klastera, ne uzimajući u obzir kovarijacione matrice klasa. Kriterijumska funkcije prema čijem priraštaju se vrši reklasifikacija odbiraka ima oblik:

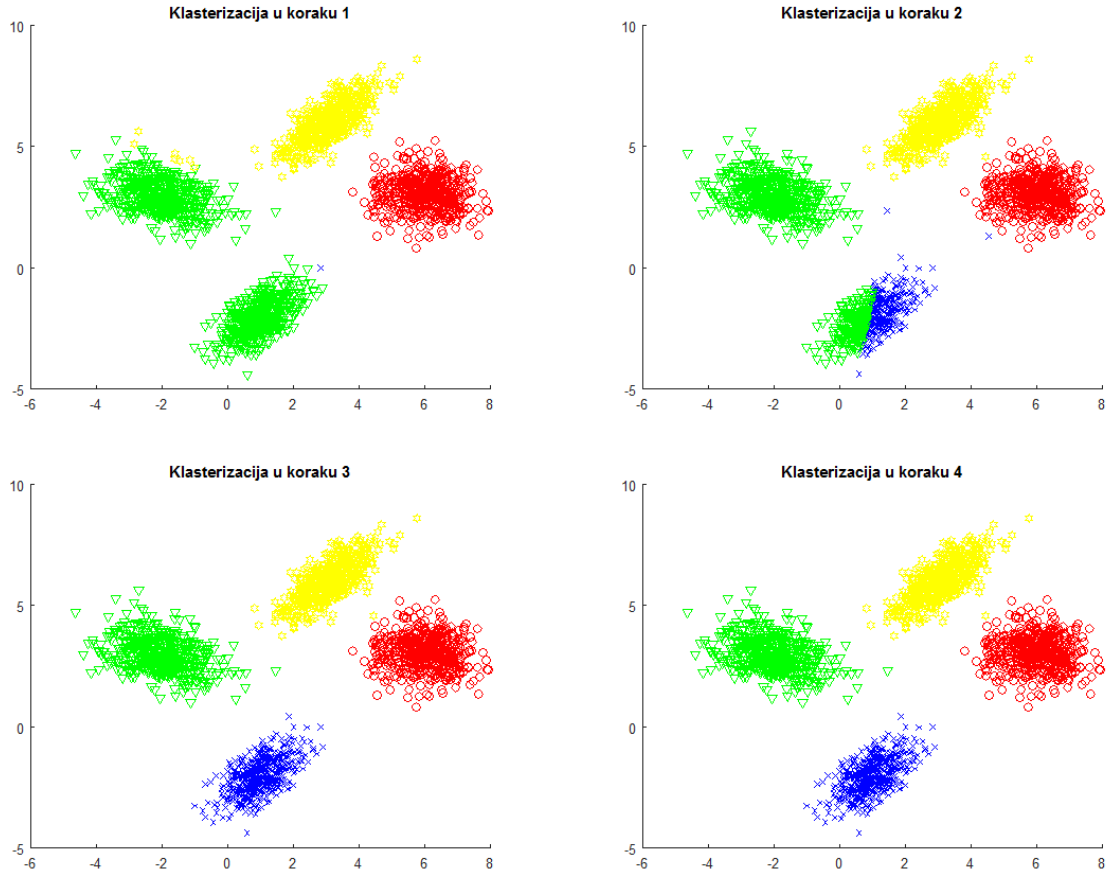
$$J = tr(S_m^{-1} S_W)$$

gde je $S_W = \sum_{i=1}^{L} P_i \Sigma_i$ matrica unutarklasnog rasejanja, a $S_m = S_B + S_W$ miksovana matrica, za $S_B = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T$ matricu međuklasnog rasejanja, gde je $M_0 = \sum_{i=1}^{L} P_i M_i$ združeni vektor matematičkog očekivanja za sve klase.
Pri reklasifikaciji odbirka X_i iz klase k_i u klasu j u l-toj iteraciji, dolazi do priraštaja kriterijuma:

$$\Delta J(i, j, l) = \frac{1}{N}(||X_i - M_j(l)||^2 - ||X_i - M_{K_i}(l)||^2)$$

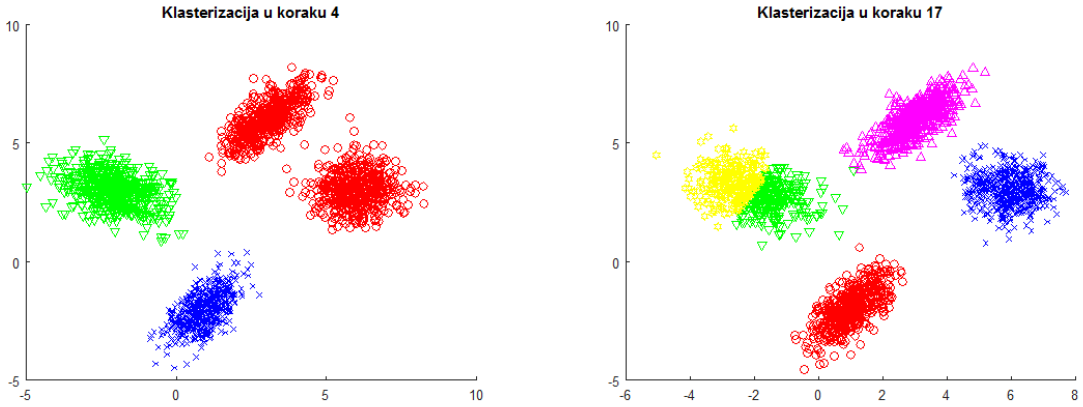
i ukoliko je priraštaj minimalan za neko $M_t(l)$, odbirak X_i se dodeljuje klasi t u l-toj iteraciji.
Procedura se svodi na sledeći algoritam:
1. Odredi se početna klasterizacija $\Omega(0)$
2. Izračunaju se $M_i(l),\ i = 1...L$
3. Računaju se odstojanja pojedinačnih odbiraka u l-toj iteraciji i odbirci se reklasifikuju prema najbližoj srednjoj vrednosti.
4. Ukoliko u l-tom koraku nije došlo do reklasifikacije, algoritam se završava. U suprotnom, generiše se korak $l = l + 1$ i vraća na korak 2.

Izvršena je analiza za slučaj poznatog broja klasa, $L = 4$ za stohastičku početnu klasterizaciju i dobijeni su rezultati prikazani na slici 22, a zatim je procedura pokrenuta za 100 ponavljanja i zaključeno da je srednji broj iteracija potrebnih za klasterizaciju ovako raspodeljenih odbiraka jednak 4.
U slučaju da broj klasa nije apriorno poznat, mogući su sledeći ishodi:
1. Broj klasa je manji od realnog (recimo L=3) $\Rightarrow$ podaci se svrstaju u 3 klase i to tako da jedna



Slika 22: Iteracije C-mean klasterizacije

od njih predstavlja uniju dva skupa podataka (slika 23), a broj iteracija na 100 ponavljanja je 7.
2. Broj klasa je veći od realnog (recimo L=5) ⇒ podaci se svrstaju u 5 klasa i to tako da je jedna grupa podataka podeljena na dva klastera (slika 23), a broj iteracija na 100 ponavljanja je 17.
C-mean metod ima osobine:
1. Neophodno je apriorno poznavanje broja klastera L.
2. Početna klasterizacija se može odabrati stohastički.
3. Klasteri su podeljeni deo po deo linearnim pravama (2D slučaj) koje predstavljaju simetrale duži koja spaja matematička očekivanja.
4. Nije garantovana optimalnost, procedura se može zaustaviti u nekom od lokalnih a ne globalnih minimuma.
5. Nije garantovana stabilnost, algoritam može ući u beskonačnu petlju reklasterizacije ista dva odbirka.



Slika 23: L=3 (levo), L=5 (desno)

2. Za rešavanje ove tačke odabran je metod klasterizacije maksimalne verodostojnosti (maximum-likelihood, ML) koji maksimizira verovatnoću da je određeni odbirak zaista iz klase u koju je smešten klasterizacijom. Ovaj metod polazi od pretpostavke da su klase Gausovski raspodeljene, pa se ukupna funkcija gustine verovatnoće iz koje dolaze odbirci može napisati kao suma Gausovih raspodela pojedinačnih klasa (polimodalna Gausova raspodela). Za nezavisna merenja, ta funkicija se dalje može zapisati kao proizvod funkcija gustine verovatnoće za pojedinačna merenja, a kada se na nju primeni negativan logaritam naziva se funkcija verodostojnosti. Dalje se traže arumenti $P_i,\ M_i\ i\ \Sigma_i$ za koje je funkcija verodostojnosti određenog odbirka maksimalna, uzimajući u obzir činjenicu $\sum_{i=1}^{L} P_i = 1$. Kriterijumska funkcija se svodi na:

$$J = \sum_{j=1}^{N} ln(f(X_j)) - \mu(\sum_{i=1}^{L} P_i - 1)$$

gde je $f(X_j) = P_1 f_1(X_j) + ... + P_L f_L(X_j)$,
$f_i(X_j) :\ N(M_i, \Sigma_i)$
pa se može pisati aposteriorna verovatnoća da odbirak X_j pripada klasi ω_i kao

$$q_i(X_j) = \frac{P_i f_i(X_j)}{f(X_j)}$$

traženjem izvoda po $P_i,\ M_i,\ \Sigma_i$ i izjednačavanjem sa nulom dobijaju se argumenti za koje je J maksimalno:

$$P_i = \frac{1}{N} \sum_{j=1}^{N} q_i(X_j)$$

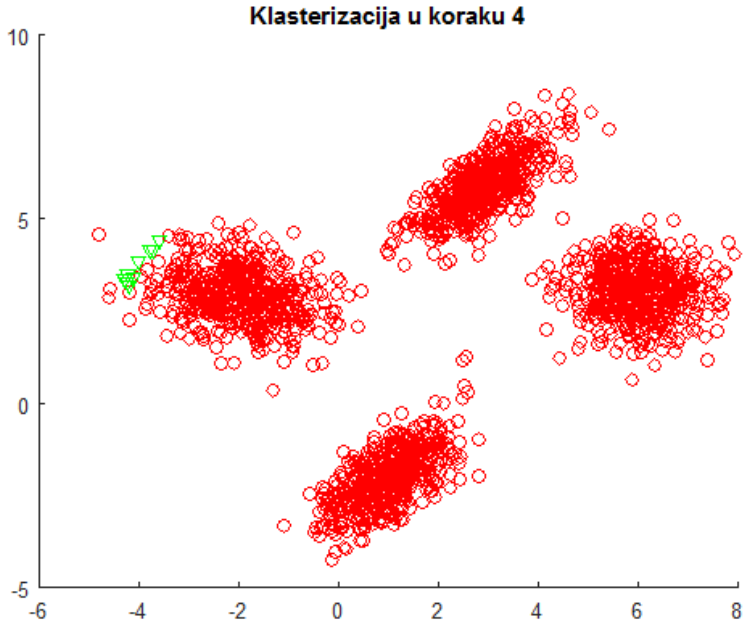
$$M_i = \frac{1}{N_i} \sum_{j=1}^{N} q_i(X_j) X_j,\ N_i = P_i N$$

$$\Sigma_i = \frac{1}{N_i} \sum_{j=1}^{N} q_i(X_j)(X_j - M_i)(X_j - M_i)^T$$

Algoritam se svodi na korake:
1. Odredi se početna klasterizacija $\Omega(0)$
2. Izračunaju se $P_i(l),\ M_i(l),\ \Sigma_i(l),\ i = 1...L$
3. Izračuna se $q_i^{(l)}(X_j),\ i = 1...L,\ j = 1...N$, odbirak X_j se reklasifikuje u neku klasu ω_t za koju je $q_i^{(l)}(X_j)$ maksimalno.

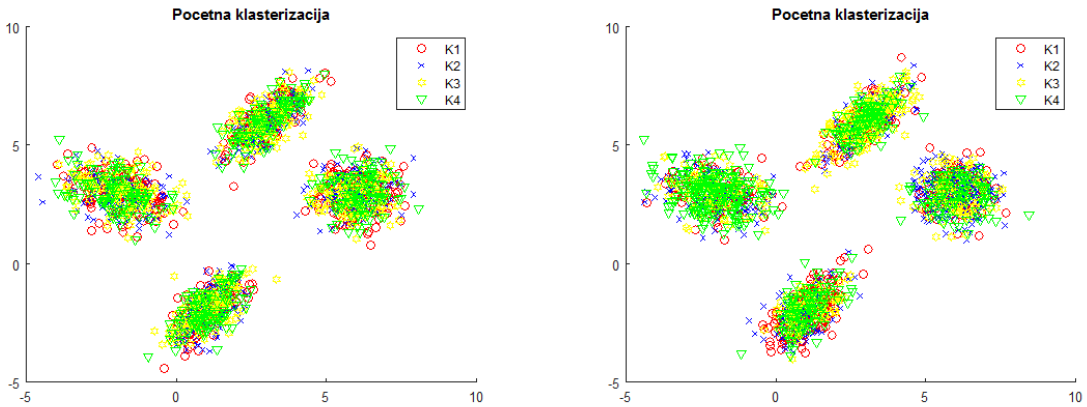
4. Ukoliko u l-toj iteraciji nije došlo do reklasifikacije algoritam se završava, u suprotnom se generiše iteracija $l = l + 1$ i vraća na korak 2.

Prvo je izvršena analiza za poznat broj klastera u slučaju stohastičke početne klasterizacije i primećeno da u tom slučaju dolazi do nestajanja pojedinih klastera usled reklasifikacije svih odbiraka tog klastera u neke druge klastere, po kriterijumu optimalnije. Konkretan primer dat je na slici 24 gde se primećuje opstajanje samo 2 od 4 klase, gde jedna vidno preovlađuje. Ovakav problem može se prevazići uvođenjem određenog predznanja u početnu klasterizaciju (ukoliko je moguće), što je ovde urađeno ubacivanjem po 100 odbiraka svake od klasa u odgovarajuće klastere.

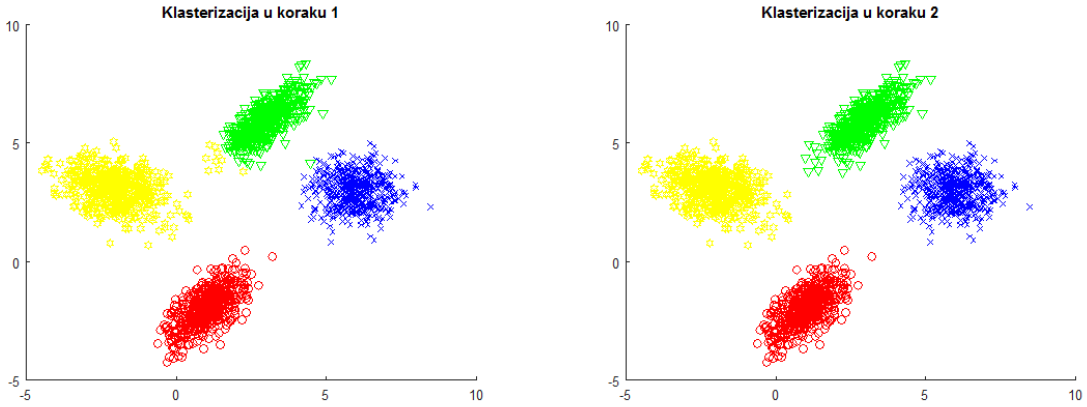


Slika 24: ML u slučaju stohastičke početne klasterizacije

Razlika slučajne i heuristički odabrane početne klasterizacije može se videti na slici 25, a prikaz rada ML algoritma dat je na slici 29. Srednji broj iteracija na 100 ponavljanja je 4.



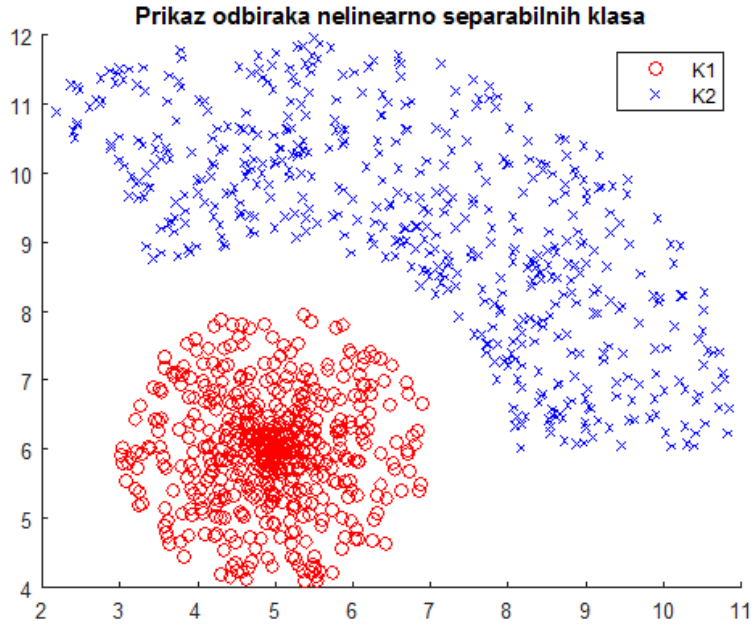
Slika 25: Poređenje izbora početne klasterizacije: slučajna(levo), sa predznanjem(desno)



Slika 26: Iteracije ML klasterizacije

U slučaju nepoznatog broja klasa dolazi do nestajanja klasa kao pri slučajnoj početnoj klasterizaciji, upravo zbog toga. Ukoliko ne postoji predznanje o broju klasa, ne postoji ni predznanje o načinu podele poodataka u nultoj iteraciji, te je zbog toga neophodno znati tačno o kom broju klasa je reč za ovu vrstu klasterizacije.
Zaključak o osobinama ML algoritma:
1. Neophodno je apriorno poznavanje broj klastera L.
2. Početna klasterizacija se ne može izvršiti stohastički.
3. Metod je pogodan za nelinearno separabilne klase.

3. Generisano je po $N = 500$ odbiraka dve nelinearno separabilne klase, od kojih je jedna kružnog oblika sa centrom u $C = (5, 6)$ i poluprečnikom $R = [0, 2\pi]$, a druga oblika trećine prstena debljine $d = 3$ sa centrom u $C = (5, 6)$.



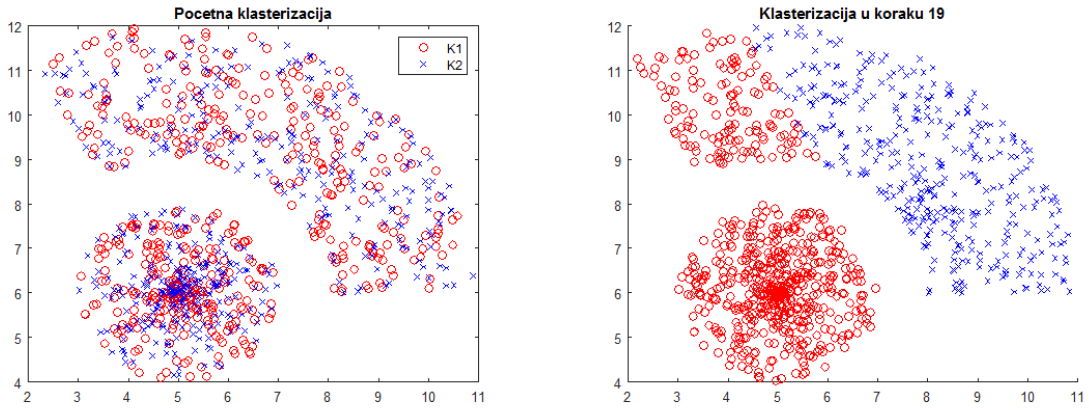
Slika 27: Prikaz odbiraka nelinearno separabilnih klasa

Klaterizacija je izvršena metodom kvadratne dekompozicije, koja za razliku od c-mean klasterizacije

uzima u obzir rasipanje klasa, kao i verovatnoću pojave klasa. Kriterijum se svodi na:

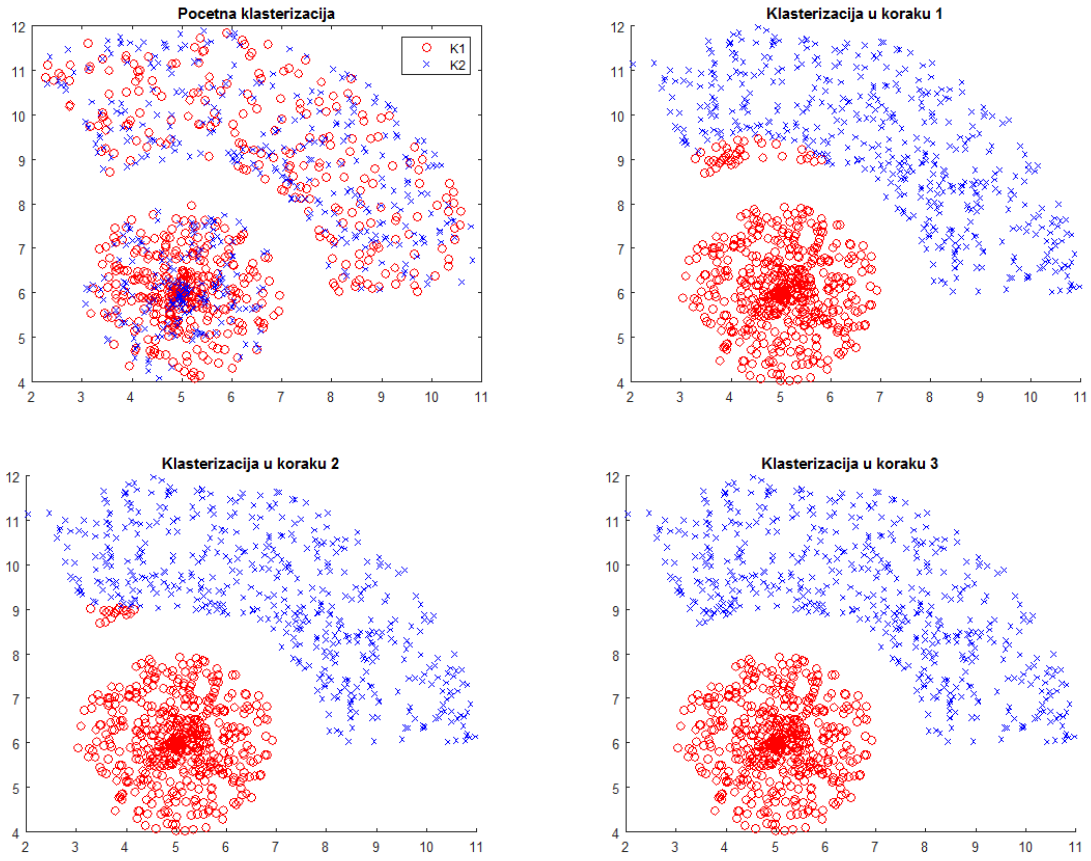
$$J(i,j,l) = \frac{1}{2}(X_i - M_j(l))^T \Sigma_j^{-1}(X_i - M_j(l)) + \frac{1}{2} ln|\Sigma_j(l)| - \frac{1}{2} ln(P_j(L)),\ \ j = 1...L$$

i ukoliko je minimalan za $j = t$, odbirak i se dodeljuje klasi t.
Algoritam se svodi na korake c-mean algoritma, s tim što se u koraku 2 dodatno računaju Σ_j i P_j.
Prvo je razmatran slučaj random početne klasterizacije. Za određena pokretanja algoritma dobijeni su dobri rezultati, ali u nekim slučajevima je zbog stohastičke početne podele klasterizacija rezultovala pogrešnom podelom (slika 28).



Slika 28: Prikaz loše klasterizacije (desno) u slučaju random početne podele (levo)

Srednji broj iteracija u ovom slučaju, na 100 ponavljanja, iznosi 19.
Zatim je urađena popravka smeštanjem po 100 odbiraka iz svake od klasa u odgovarajuće klastere i dobijen korektan rezultat u manjem broju koraka.
Srednji broj iteracija na 100 ponavljanja u ovom slučaju je 4.
U slučaju nepoznatog broja klasa dolazi do sličnog problema kao kod ML algoritma, zbog neophodnog predznanja.
Osobine algoritma:
1. Neophodno je apriorno poznavanje broja klastera L.
2. Početna klasterizacija se ne može izvršiti stohastički.
3. Granice između klastera su deo po deo kvadratne krive.
4. Nije garantovana optimalnost.
5. Nije garantovana stabilnost.



Slika 29: Prikaz dobre klasterizacije u slučaju početne podele sa predznanjem (slika 1)