Univerzitet u Beogradu - Elektrotehnički fakultet

Katedra za signale i sisteme

DOMAĆI ZADATAK

Prepoznavanje oblika

Student
Marija Rakonjac 2020/0222

Beograd, *Jul* 2024. godine

SADRŽAJ

Zadatak 1 - Opcija 2

U timu predmeta na MSTeams platformi dostupna je baza slikanih šaka koje pokazuju „papir“, „kamen“ i „makaze“. Projektovati inovativni sistem za prepoznavanje pokazanih znakova zasnovan na testiranju hipoteza.

Slika 1. Prikaz slikanih šaka koje pokazuju kamen, papir i makaze

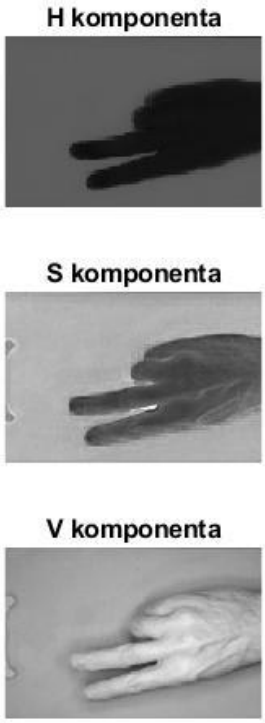
Zadatak 1.a)

Detaljno opisati algoritam za obradu slike i odabir obeležja koji prethodi samoj klasifikaciji. Algoritam treba da bude što robusniji (na različite osvetljaje, položaje šaka, načine pokazivanja znakova itd).

Na početku, iz baze su učitane slike, gde je jedan primer takve slike prikazan na sledećoj slici.

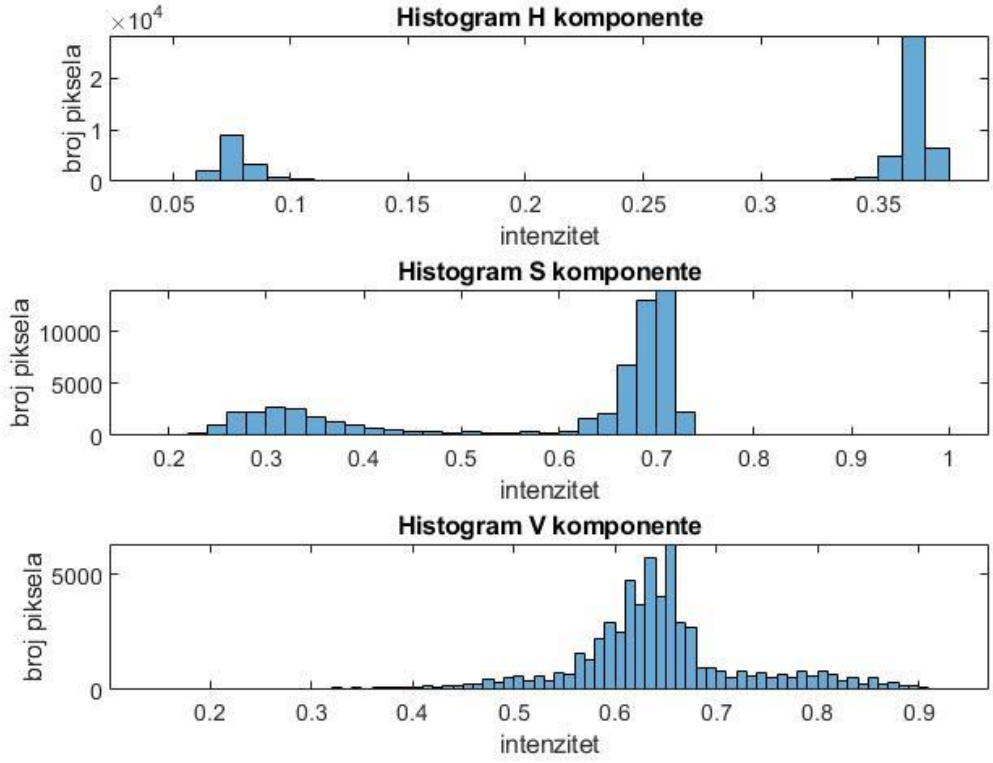
Slika 2. Primer učitane slike

Sa ciljem binarizacije slika je iz RGB sistema prebačena u HSV kolor sistem, što je prikazano na sledećoj fotografiji.



Slika 3. Slika u HSV kolor sistemu

Histogrami svake od komponenti HSV sistema prikazani su na slici. Možemo zaključiti da su intenziteti H komponente lako separabilni, te biramo ovu komponentu za prag binarizacije, za svaku od tri klase (tačan prag koji je odabran iznosi 0.2).

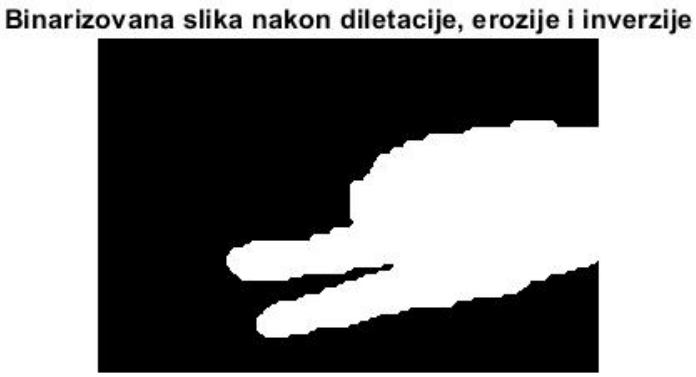


Slika 4. Histogrami komponenti HSV sistema

Nakon binarizacije slika izgleda na sledeći način.

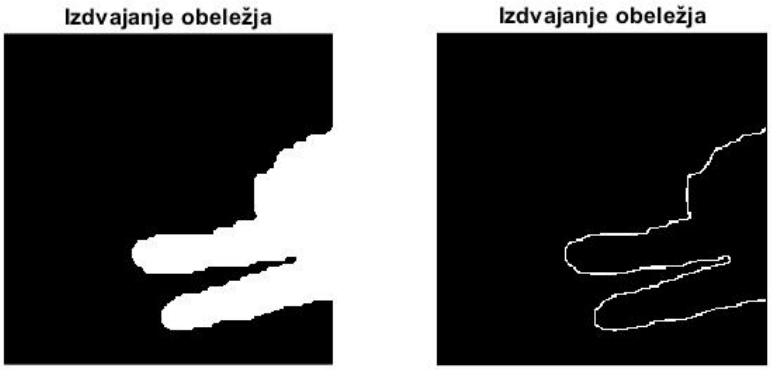
Slika 5. Binarizovana slika

Možemo primetiti blagi crni okvir oko slike, koji odlučujemo da eliminišemo primenom diletacije i erozije. Nakon tog postupka kao i inverzije slike, dobijen je konačni rezultat binarizacije prikazan na sledećoj slici.



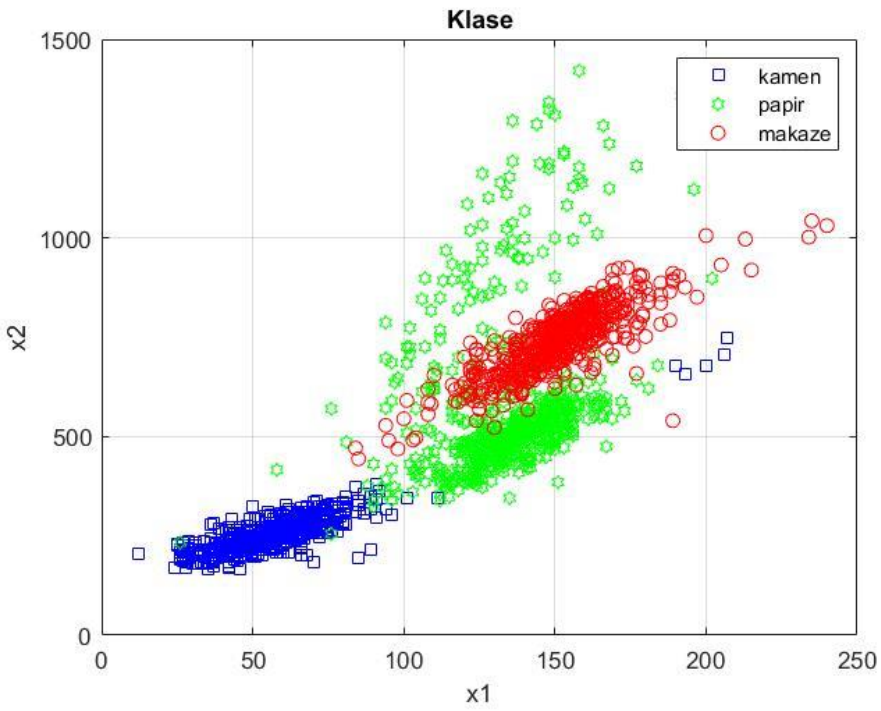
Slika 6. Binarizovana slika nakon diletacije, erozije i inverzije

Izabrana su dva reprezentativna obeležja: distanca između najlevljeg belog piksela i najšireg dela šake, kao i dužina ivica istog dela slike. Ova obeležja prikazana su na sledećoj slici.



Slika 7. Izdvojena obeležja

Obeležja za sve 3 klase prikazana su na sledećoj slici.

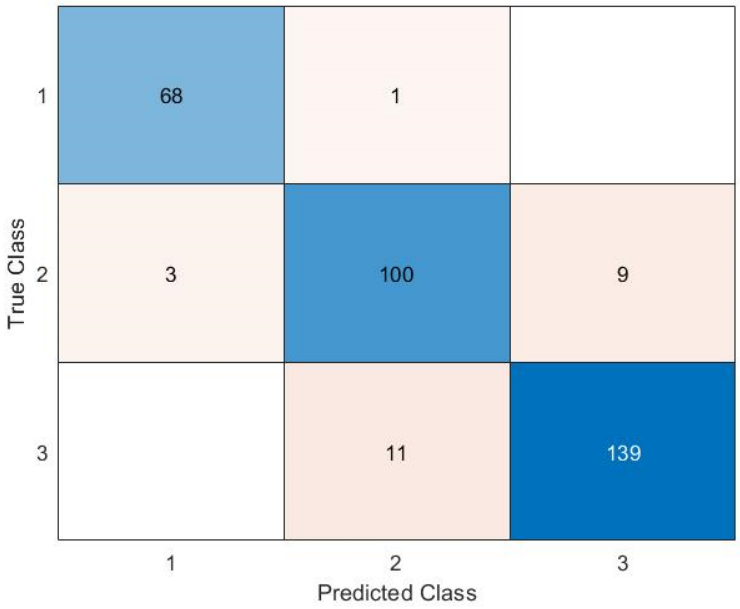


Slika 8. Prikaz klasa uz pomoć odabranih obeležja

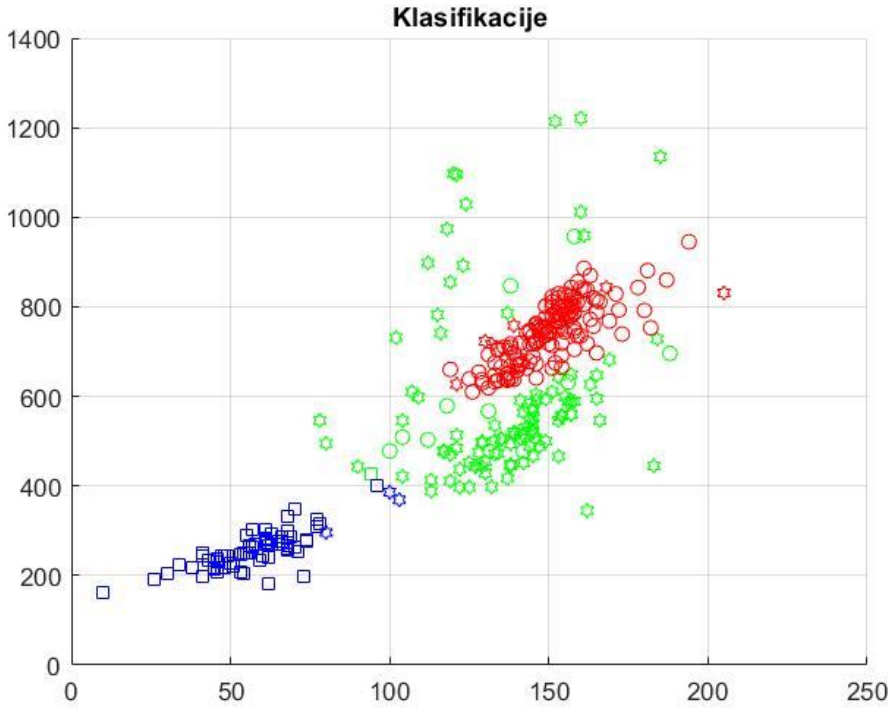
Zadatak 1.b)

Izvršiti podelu na trening i test skup. Rezultate klasifikacije test skupa prikazati u obliku matrice konfuzije.

Svaka od klasa sadrži oko 700 slika, pa je trening skup uzeto 600 slika, a ostatak za test. Korišćen je klasifikator više hipoteza, i dobijena je tačnost od 0.9275. Konfuziona matrica i klasifikacija su dati na sledećim graficima.



Slika 9. Konfuziona matrica za korišćenje klasifikatora više hipoteza

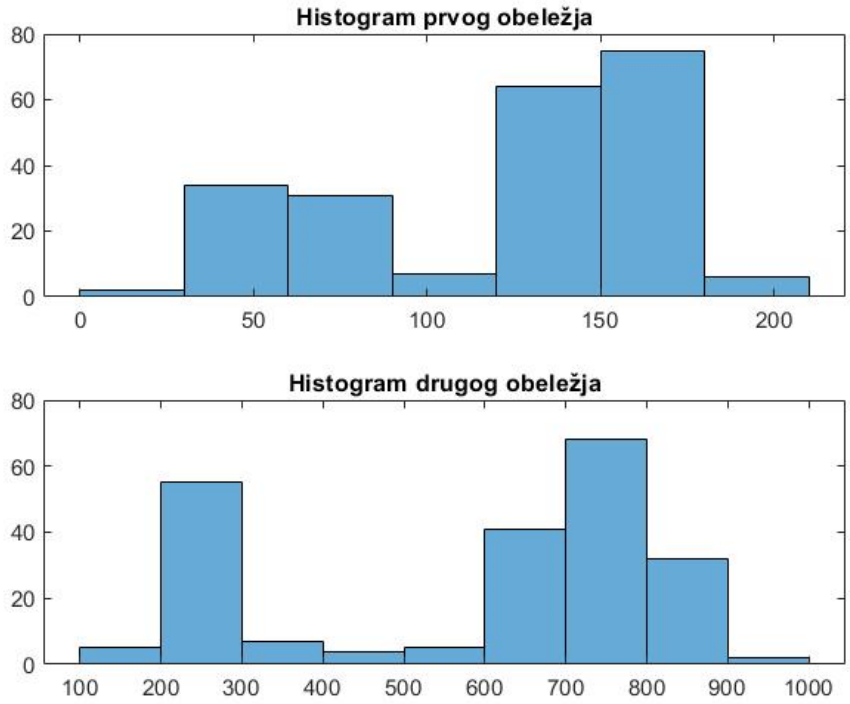


Slika 10. Klasifikacija korišćenjem klasifikatora više hipoteza

Zadatak 1.c)

Odabrati dva znaka i dva obeležja takva da su odabrani znakovi što separabilniji u tom prostoru. Prikazati histogram obeležja za oba slova i prokomentarisati njihov oblik.

Za dalju klasifikaciju odabrani su znakovi makaze i kamen, a zadržana su gore navedena obeležja. Histogrami su prikazani na sledećim slikama, i možemo zaključiti da su obeležja lako separabilna.

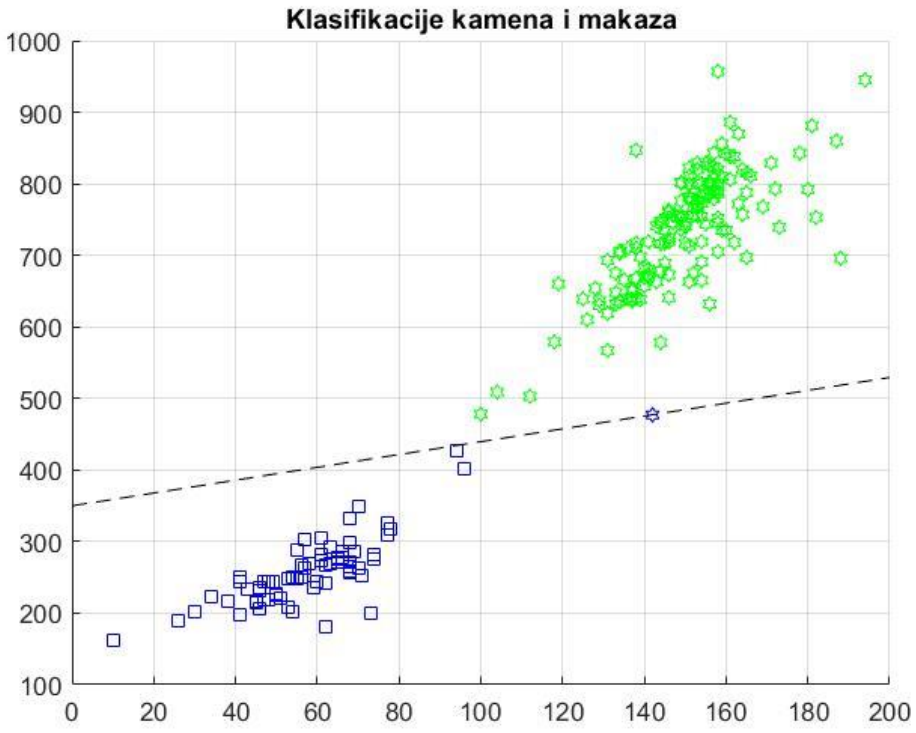


Slika 11. Histogrami odabranih obeležja

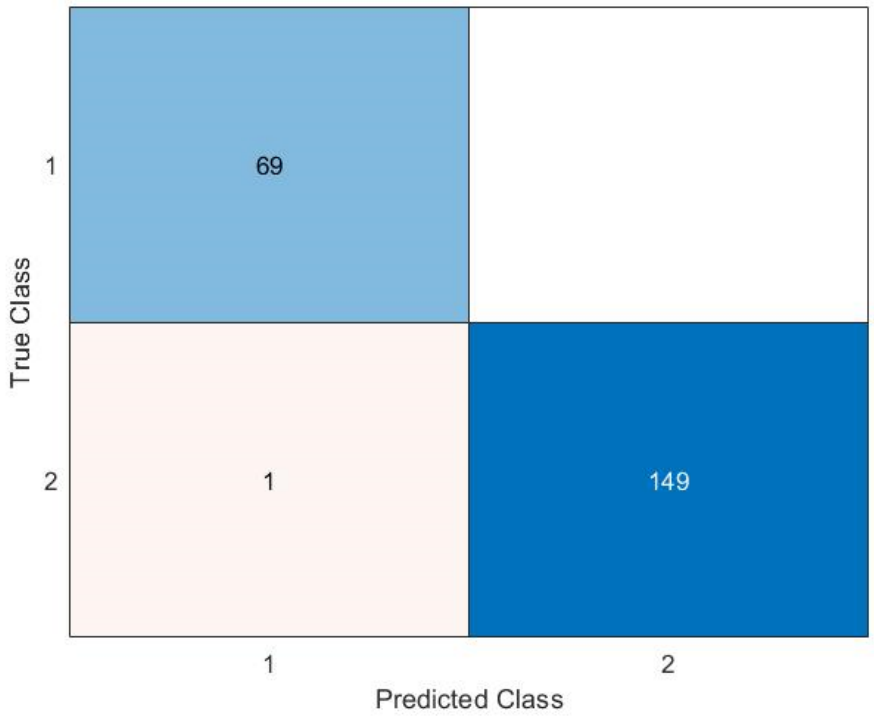
Zadatak 1.d)

Za slova i obeležja pod c) projektovati parametarski klasifikator po izboru i iscrtati klasifikacionu liniju.

Za klasifikaciju odabranih znakova korišćen je linearni klasifikator na bazi željenog izlaza, pri čemu je jedna od klasa (makaze) otežinjena u odnosu na drugu radi veće tačnosti. Dobijena tačnost iznosi 0.9954. Grafik sa klasifikacionom krivom, kao i konfuziona matrica, prikazani su na sledećim slikama.



Slika 12. Klasifikacija znakova kamen i makaze



Slika 13. Konfuziona matrica za klasifikaciju znakova kamen i makaze

Zadatak 2

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

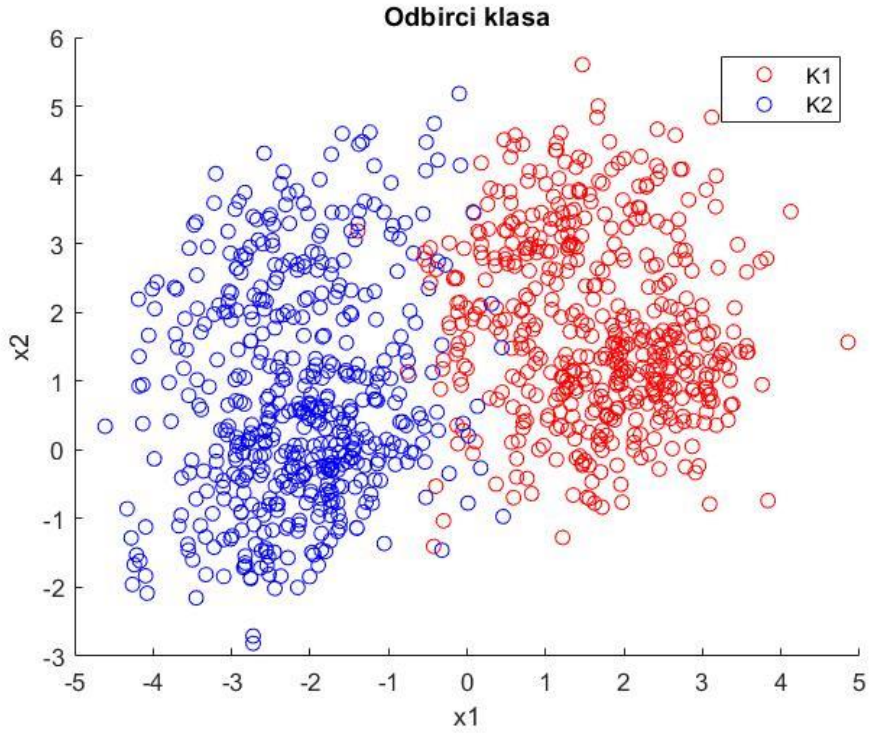
$$\Omega_1 \sim P_{11} \cdot N(M_{11}, \Sigma_{11}) + P_{12} \cdot N(M_{12}, \Sigma_{12})$$

$$\Omega_2 \sim P_{21} \cdot N(M_{21}, \Sigma_{21}) + P_{22} \cdot N(M_{22}, \Sigma_{22})$$

Parametre klasa samostalno izabrati.

Zadatak 2.a)

Na dijagramu prikazati odbirke.

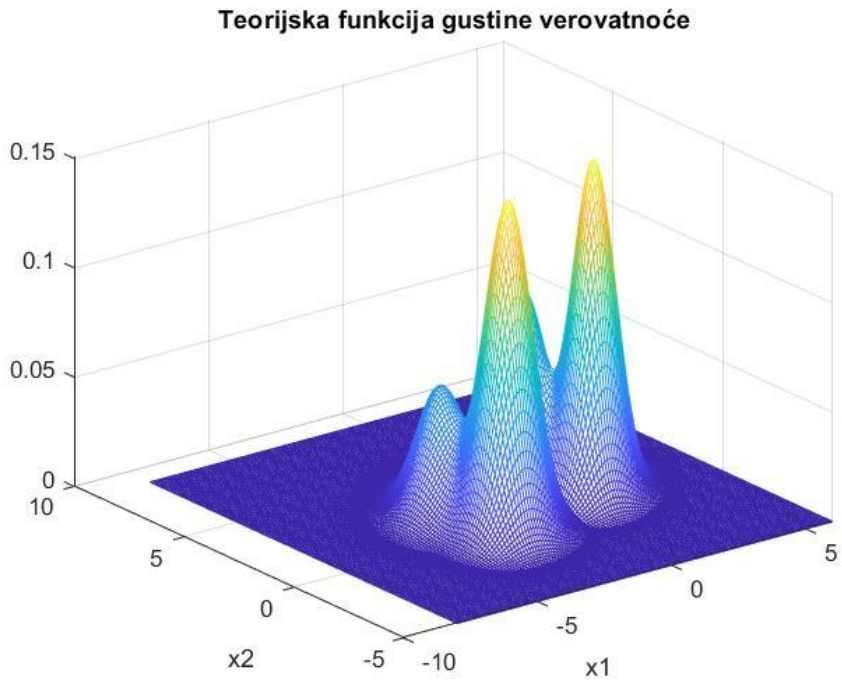


Slika 14. Odbirci dveju dvodimenzionalnih bimodalnih klasa

Zadatak 2.b)

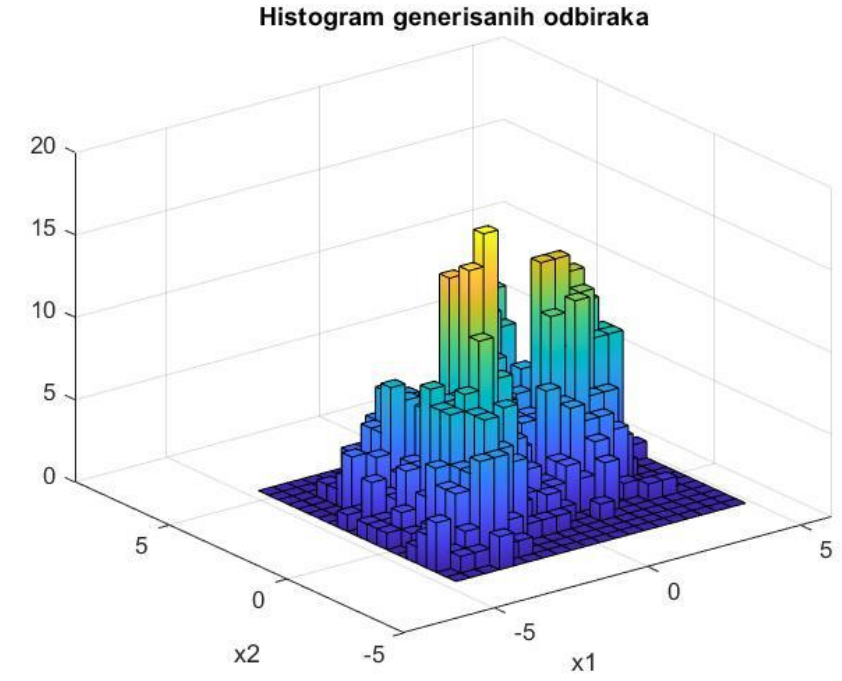
Iscrtati kako teorijski izgledaju funkcije gustine verovatnoće za raspodele klasa i uporediti ih sa histogramom generisanih odbiraka.

Teorijska funkcija gustina verovatnoće se računa za celu mrežu xy i prikazana je na sledećoj slici.



Slika 15. Teorijska funkcija gustine verovatnoće

Histogram generisanih odbiraka je prikazan na narednoj slici. Može se primetiti da su oblici funkcije gustine verovatnoće i histograma veoma slični.



Slika 16. Histogram generisanih odbiraka

Zadatak 2.c)

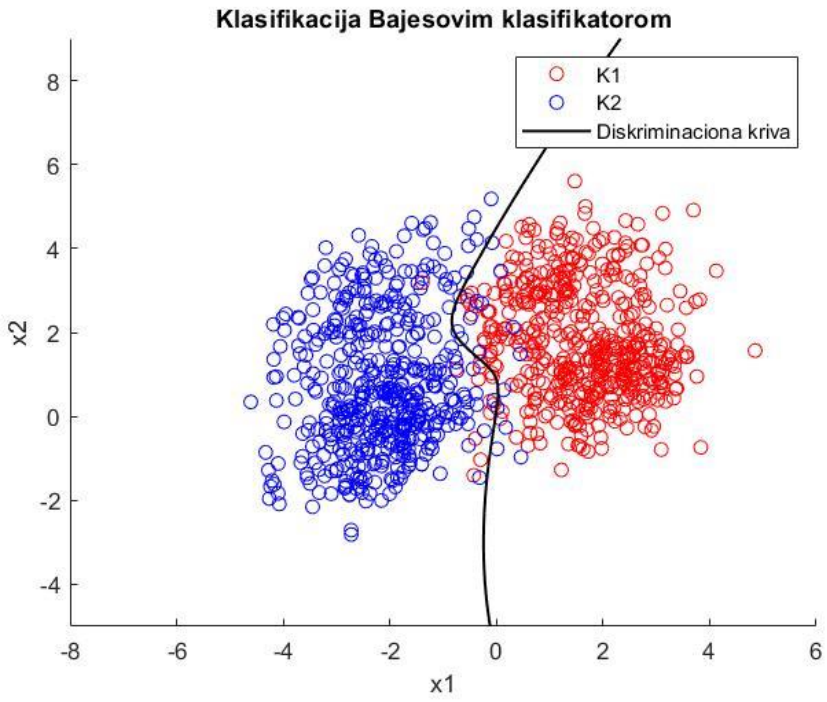
Projektovati Bajesov klasifikator minimalne greške i na dijagramu, zajedno sa odbircima, skicirati klasifikacionu liniju. Uporediti grešku klasifikacije konkretnih odbiraka sa teorijskom greškom klasifikacije prve i druge vrste za datu postavku.

Bajesov klaisfikator minimalne greške ima sledeći oblik

$$h(X) = -log\frac{f_1(X)}{f_2(X)} < 0 \quad \rightarrow X \in \omega_1$$

$$h(X) = -log\frac{f_1(X)}{f_2(X)} > 0 \quad \rightarrow X \in \omega_2$$

Na sledećoj slici je prikazana diskriminaciona kriva za Bajesov klasifikator minimalne verovatnoće greške.



Slika 17. Klasifikacija pomoću Bajesovog klasifikatora

Dobijene greške klasifikacije prvog tipa (eksperimentalna i teorijska) su respektivno 0.024 i 0.025. S druge strane, greške drugog tipa su 0.018 i 0.023.

Zadatak 2.d)

Projektovati klasifikator minimalne cene tako da se više penalizuje pogrešna klasifikacija odbiraka iz prve klase.

Klasifikator minimalne cene daje mogućnost da se pogrešna odluka za jednu klasu vrednuje više u odnosu na drugu klasu, što je veoma korisno u mnogim medicinskim i ekonomskim problemima. Cena odluke se definiše na sledeći način:

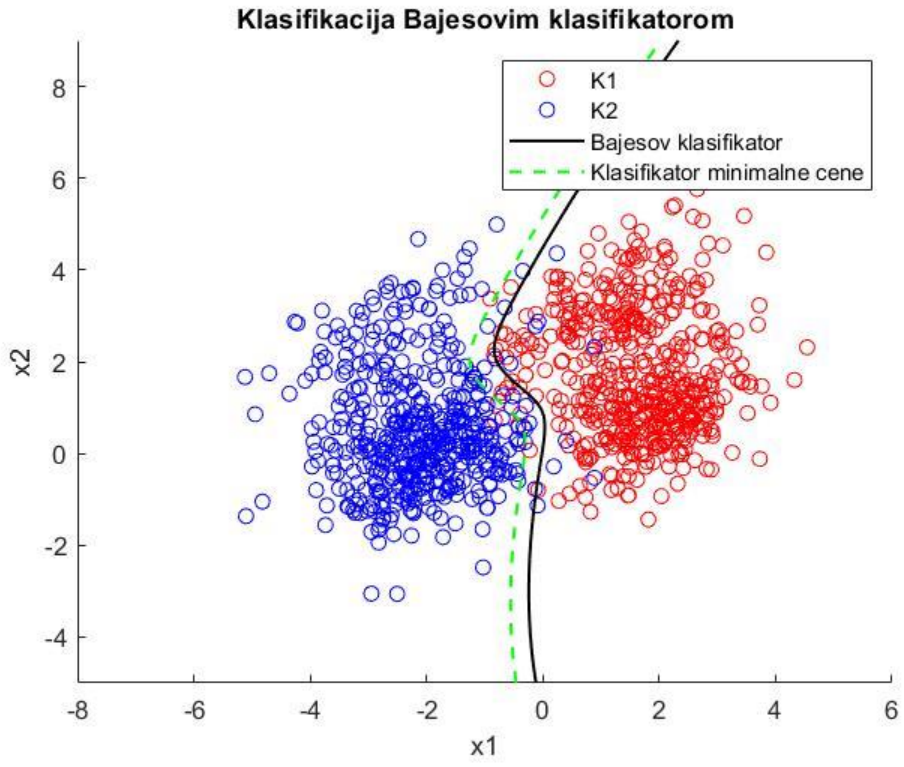
$$c_{ij} = cena\ odluke\ X\ \in\ \omega_i\ kada\ zapravo\ X\ \in\ \omega_j$$

Opšti oblik klasifikatora ima sledeći oblik:

$$h(X) = -log\frac{f_1(X)}{f_2(X)} < -log\frac{(c_{12}-c_{22})\cdot P_2}{(c_{21}-c_{11})\cdot P_1} \quad \rightarrow\ X \in \omega_1$$

$$h(X) = -log\frac{f_1(X)}{f_2(X)} > -log\frac{(c_{12}-c_{22})\cdot P_2}{(c_{21}-c_{11})\cdot P_1} \quad \rightarrow\ X \in \omega_2$$

Rezultati klasifikacije dobijeni klasifikatorom minimalne cene su prikazani na sledećem grafiku. Kako je potrebno više penalizovati grešku klasifikacije odbiraka iz prve klase, diskriminaciona kriva je u odnosu na Bajesov klasifikator pomerena ka odbircima druge klase.



Slika 18. Klasifikacija pomoću Bajesovog i klasifikatora minimalne cene

Zadatak 2.e)

Ponoviti prethodnu tačku za Neuman-Pearson-ov klasifikator. Obrazložiti izbor $\varepsilon 2 = \varepsilon 0$.

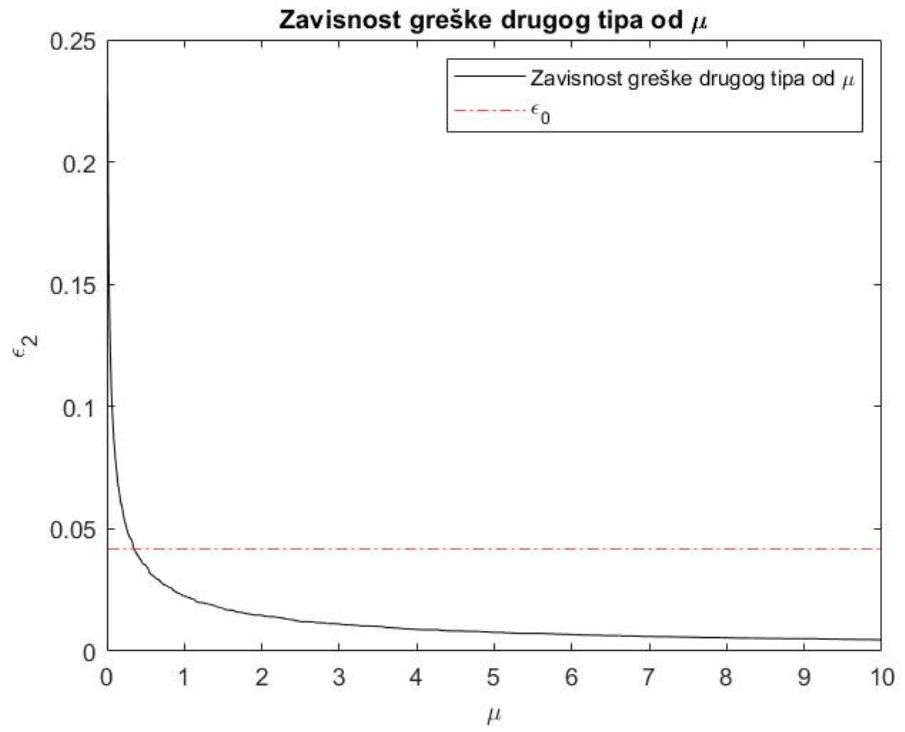
Nojman-Pirson-ov klasifikator podrazumeva fiksiranje jednog tipa greške, dok se drugi tip minimizuje. Njegova prednost u odnosu na Bajesov test je što ne vrši minimizaciju zbira grešaka prvog i drugog tipa, već se radi minimizacija važnije od njih, a druga (često se u literaturi naziva FA – False Alarm) se postavlja na neku prihvatljivu vrednost.

Pošto Bajesov test minimalne verovatnoće daje minimalno ε onda se, ukoliko je recimo grešku prvog tipa potrebno svesti na najmanji mogući nivo, greška drugog tipa postavlja baš na vrednost ε jer je tada veoma izvesno da će ε_1 biti veoma mala pošto njih dve kada se saberu u zbiru trebaju da daju vrednost blisku ε.

Međutim, greška ε_2 se računa kao

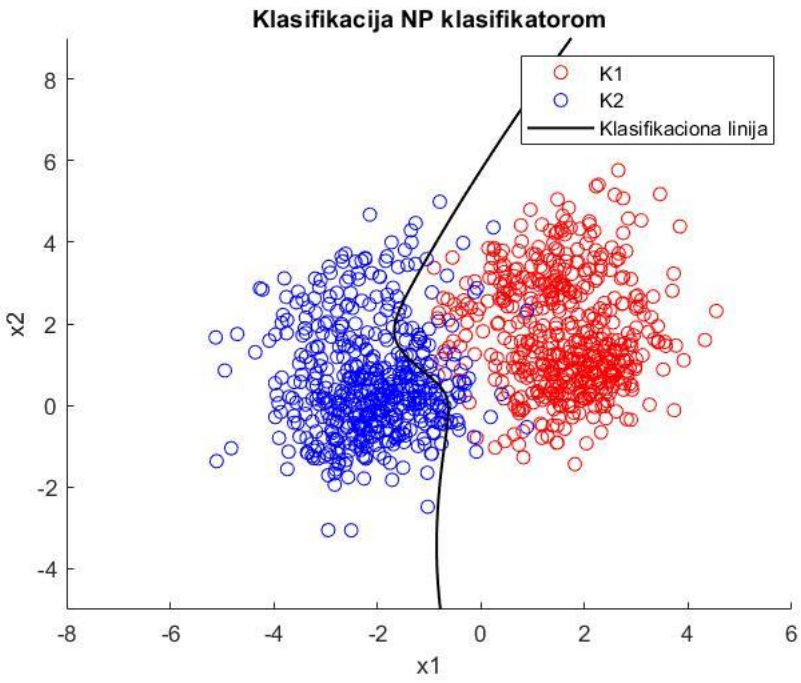
$$\varepsilon_2 = \int_{-\infty}^{-\ln(\mu)} f_h(h/\omega_2)dh$$

pa je potrebno odrediti parametar μ koji se koristi za vrednost praga odluke. On se određuje numerički, gde se različite greške drugog tipa crta vrednost parametra μ i na kraju se samo odabere ona vrednost koja daje željeno ε_2.



Slika 19. Zavisnost greške drugog tipa od parametra mi

S obzirom da je ε iz prve tačke 0.042, μ koje odgovara toj vrednosti iznosi 0.0421. Na osnovu ovog μ je konstruisana diskriminaciona kriva koja je prikazana na sledećem grafiku.



Slika 20. Klasifikacija pomoću Nojman-Pirsonovog klasifikatora

Zadatak 2.f)

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.

Wald-ov sekvencijalni test pripada grupi sekvencijalnih testova za testiranje hipoteza koji prikuplja podatke koji pristižu, a odluku donosi na sledeći način

$$s_m \leq a \quad \rightarrow X \in \omega_1$$

$$a < s_m < b \quad \rightarrow uzeti\ (m+1) - vi\ vektor$$

$$s_m \geq b \quad \rightarrow X \in \omega_2$$

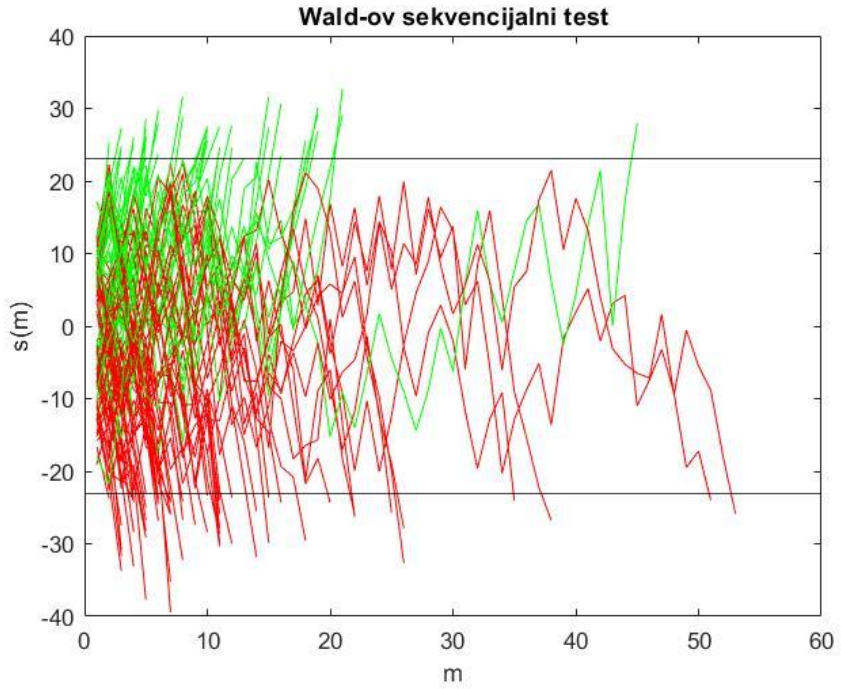
Dakle, ovaj test donosi odluku tek kada se postignu neki pragovi, a oni sami zavise od definisanih željenih verovatnoća grešaka prvog i drugog tipa

$$a = -ln\frac{1-\varepsilon_1}{\varepsilon_2}$$

$$b = -ln\frac{\varepsilon_1}{1-\varepsilon_2}$$

što znači da se mogu odrediti na osvu zadatih parametara.

Rezultati Wald-ovog testa za 100 klasifikacija prikazani su na sledećoj slici.



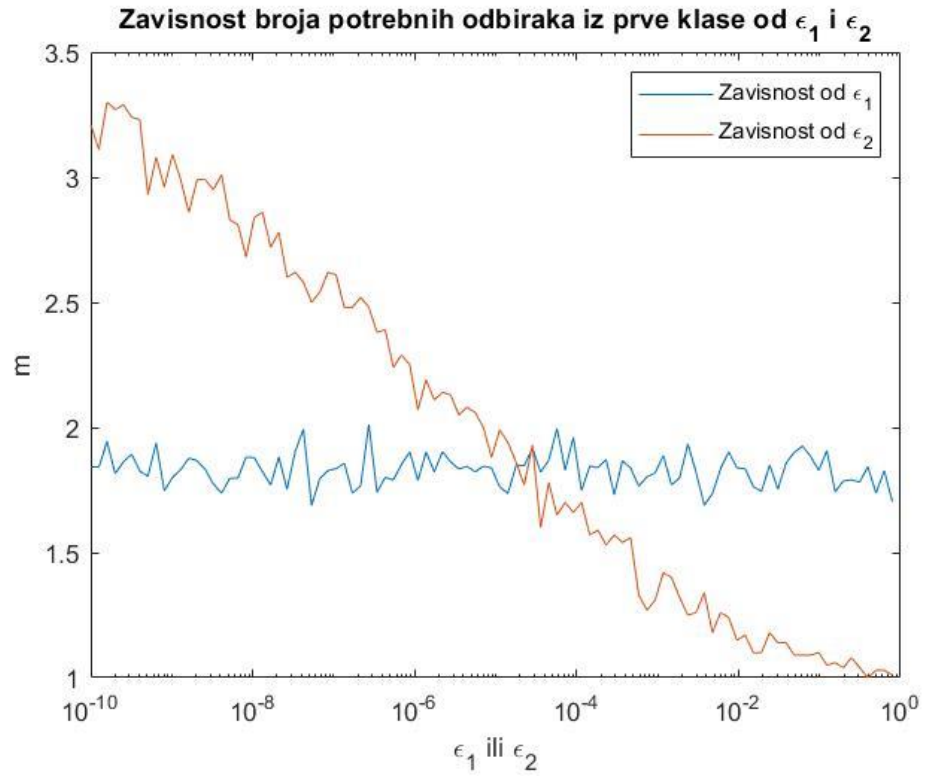
Slika 21. Grafički prikaz Waldovog sekvencijalnog testa

Srednji broj odbiraka prve i druge klase zavisi od željenih grešaka prvog i drugog tipa kao i od η_i koje predstavlja očekivanje diskriminacione funkcije pod uslovom da odbirci dolaze iz i-te klase

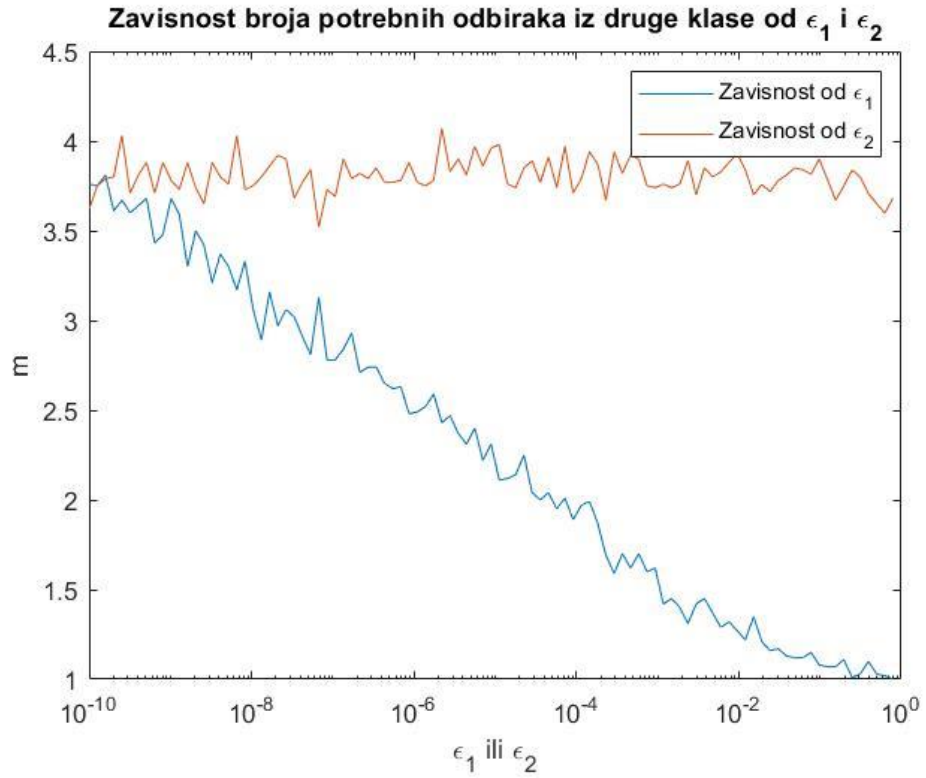
$$m_1 = \frac{a(1-\varepsilon_1) + b\varepsilon_1}{\eta_1}$$

$$m_2 = \frac{b(1-\varepsilon_2) + a\varepsilon_2}{\eta_2}$$

Za greške prvog i drugog tipa je uzeta vrednost 10^{-10}, a grafici za srednji broj potrebnih odbiraka iz prve i druge klase zavisni od tako zadatih grešaka su prikazani na sledećim graficima.



Slika 22. Zavisnost broja potrebnih odbiraka iz prve klase od grešaka prvog i drugog tipa

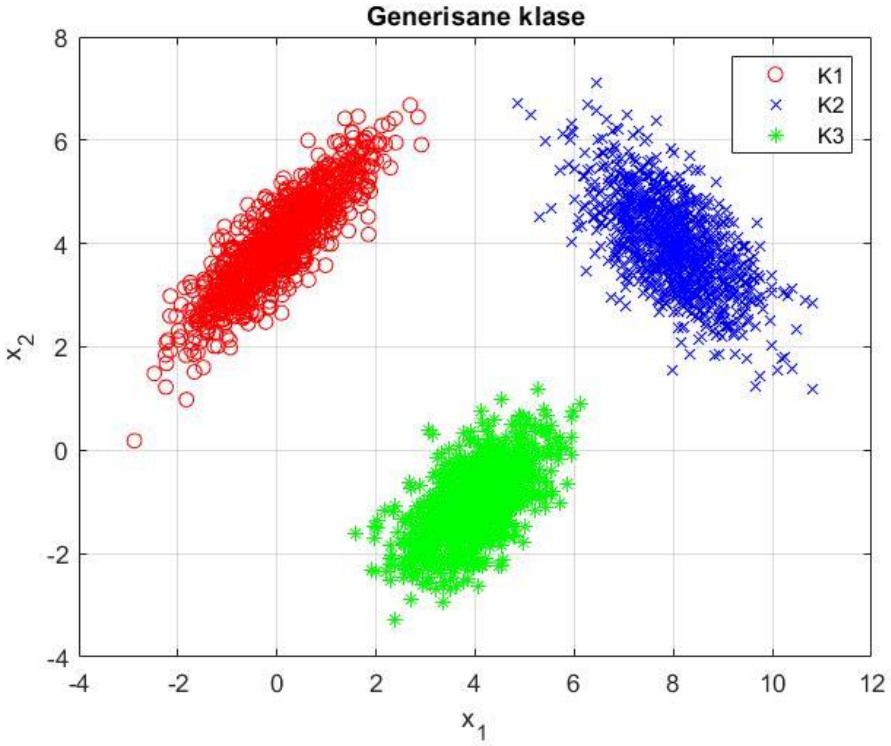


Slika 23. Zavisnost broja potrebnih odbiraka iz druge klase od grešaka prvog i drugog tipa

Zadatak 3 – Opcija 1

Zadatak 3.1.

Generisati tri klase dvodimenzionalnih oblika. Izabrati funkciju gustine verovatnoće oblika tako da klase budu linearno separabilne.



Slika 24. Tri generisane dvodimenzionalne klase

Zadatak 3.1.a)

Za tako generisane oblike izvšiti projektovanje linearnog klasifikatora jednom od tri iterativne procedure. Rezultate prikazati u obliku matrice konfuzije. Detaljno opisati postupak klasifikacije.

Linearni klasifikator je sledećeg oblika:

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

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

Dok se parametri klasifikatora izračunavaju na osnovu sledećih izraza:

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

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

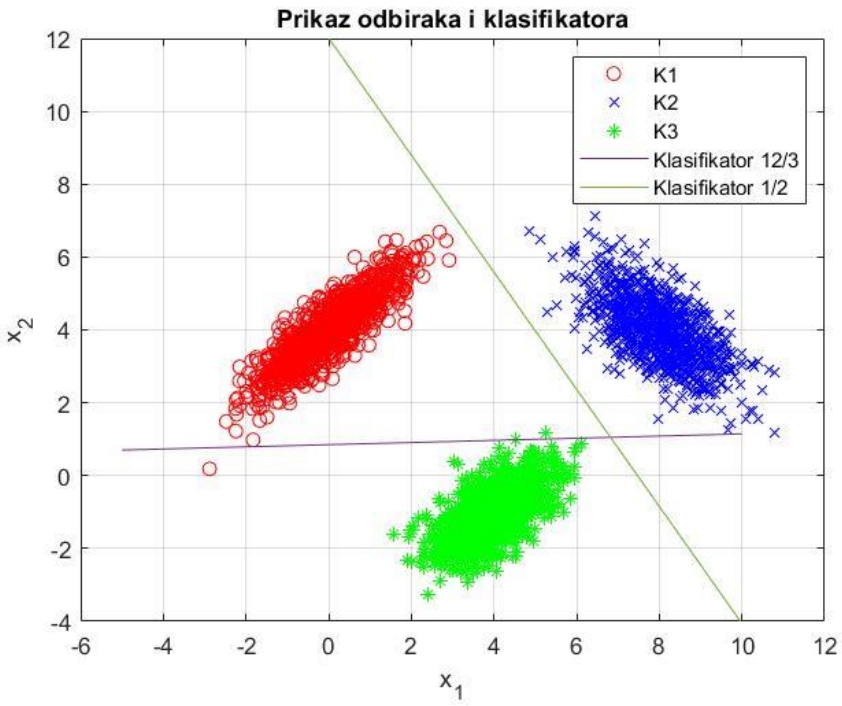
$$\sigma_i = V^T \Sigma_i V, \ i = 1,2 \quad (3)$$

$$\eta_i = V^T M_i + v_0, \quad i = 1,2 \tag{4}$$

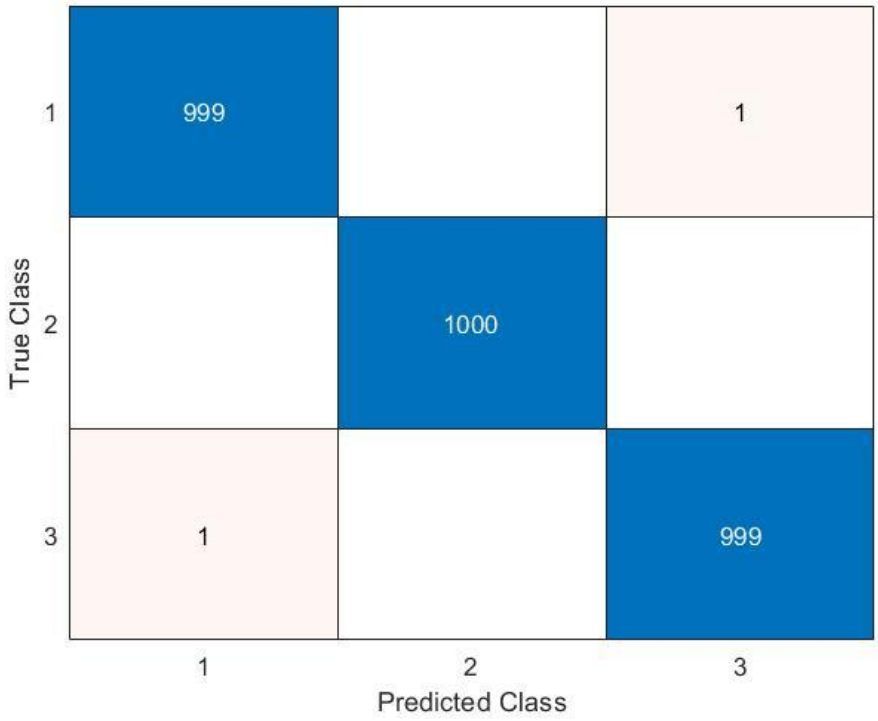
Pošto je dati sistem jednačina implicitan, za dobijanje parametara klasifikatora se koristi neka od numerčkih metoda. Odabrana je druga iterativna metoda koja se često može sresti i pod nazivom metod resupstitucije, i koja se sastoji iz sledećih koraka:

1) Na osnovu raspoloživog broja oblika iz obe klase se procene $\widetilde{M_i}$ i $\widetilde{\Sigma_i}$.
2) Odredi se vektor V na osnovu prethodno određenih procena srednjih vrednosti i kovarijacionih matrica, i za zadato s koje se menja u opsegu između 0 i 1 sa mailim korakom.
3) Na osnovu dobijenog V se izračunaju
$$y_j^{(i)} = V^T X_j^{(i)}; \quad i = 1,2; \; j = 1, \dots, N_i$$
gde je sa N_i označen broj odbiraka iz svake klase ponaosob.
4) Oni $y_j^{(1)}$ i $y_j^{(2)}$ koji ne zadovoljavaju $y_j^{(1)} < -v_0$ i $y_j^{(2)} > -v_0$ se broje kao greške pri čemu se v_0 menja u opsegu od $-\max\left(\max\left(y_j^{(1)}\right), \max\left(y_j^{(2)}\right)\right)$ do $-\min\left(\min\left(y_j^{(1)}\right), \min\left(y_j^{(2)}\right)\right)$ i pri tome se pamti ona vrednost v_0 koja rezultuje najmanjim brojem grešaka.
5) Menja se vrednost parametra s od 0 do 1 sa korakom Δs i skicira se broj grešaka od s. Bira se ono s za koje je broj grešaka najmanji, i na osnovu njega se projektuju traženi parametri linearnog klasifikatora.

S obzirom da ovaj zadatak zahteva klasifikaciju na 3 klase, potrebno je projektovani deo po deo linearni klasifikator. Inicijalno je formiran linearni klasifikator koji razdvaja 1. i 2. klasu od 3., a zatim i klasifikator koji razdvaja 1. i 2. klasu. Na sledećim slikama prikazani su odbirci klasa i odgovarajuće diskriminacione krive, kao i konfuziona matrica.



Slika 25. Klasifikacija pomoću linearnog klasifikatora



Slika 26. Konfuziona matrica za projektovani linearni klasifikator

Zadatak 3.1.b)

Ponoviti prethodni postupak korišćenjem metode željenog izlaza. Analizirati uticaj elemenata u matrici željenih izlaza na konačnu formu linearnog klasifikatora.

Metod željenog izlaza podrazumeva da klasifikator ima sledeći oblik

$$h(X) = -V^T \cdot X - v_0 > 0 \;\; \rightarrow X \in \omega_1$$

$$h(X) = V^T \cdot X + v_0 > 0 \;\; \rightarrow X \in \omega_2$$

i da se uvodi novi vektor Z kojim se predstavljaju oblici

$$Z = [-1 \; -X_1 \; -X_2 \ldots -X_n]^T; \;\; X \in \omega_1$$

$$Z = [1 \;\; X_1 \;\; X_2 \;\; \ldots \;\; X_n]^T; \;\; X \in \omega_2$$

Tada diskriminacion funkcija ima sledeći oblik

$$h(Z) = W^T Z > 0$$

Ukoliko se svakom semplu Z sada pridruži žejeni izlaz, potrebno je odrediti vektor W koji yz rezultat daje najviše tačnih željenih izlaza. On se dobija minimizacijom sledeće kriterijumske funkcije

$$\overline{\varepsilon^2} = \frac{1}{N}\sum_{j=1}^{N}(W^T Z - \gamma(Z_j))^2$$

Sređivanjem jednačina i traženjem izvoda po greške W dobija se sledeći rezultat

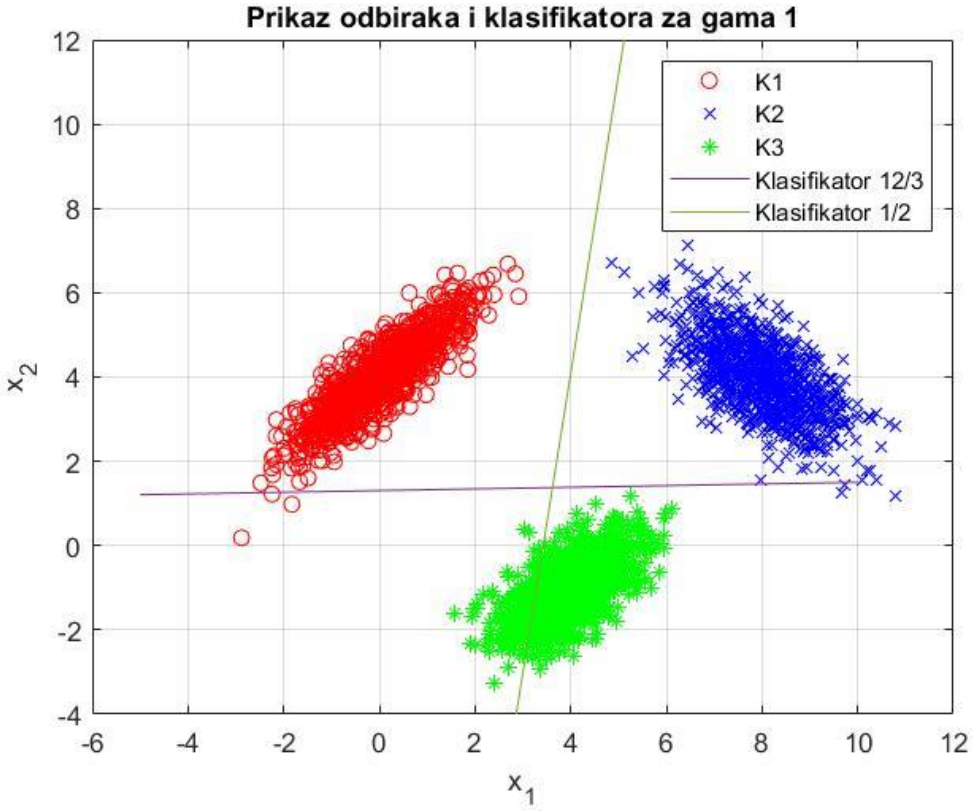
$$W = (UU^T)^{-1}U\Gamma$$

gde je matrica U matrica uzoraka a matrica Γ matrica željenih izlaza

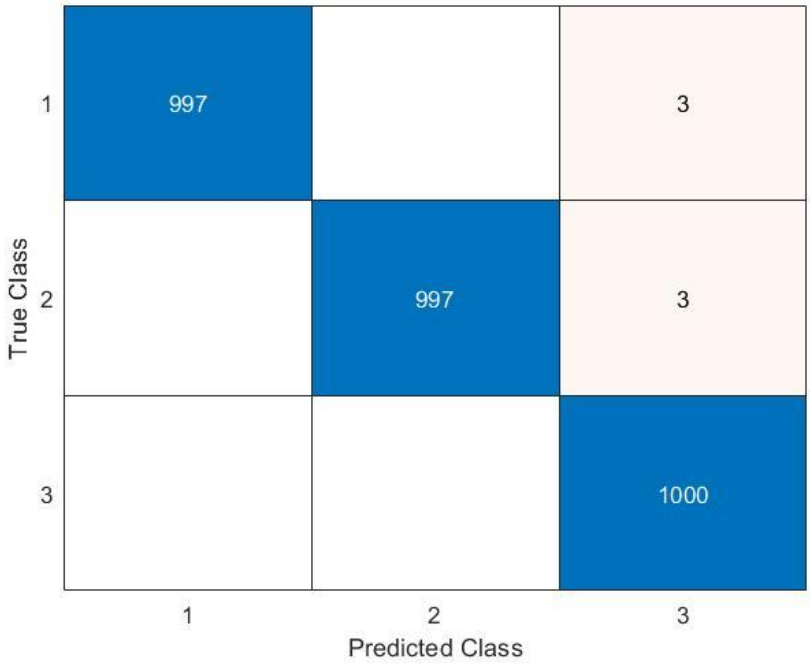
$$U = [Z_1 \;\; Z_2 \;\; \ldots \;\; Z_N \;]$$

$$\Gamma = [\,\gamma(Z_1) \;\; \gamma(Z_2) \ldots \;\; \gamma(Z_N)]^T$$

Klasifikacija 3 klase je kao i u prošlom zahtevu rađena projektovanjem deo po deo linearnog klasifikatora. Na sledećem grafiku prikazan je dobijeni klasifikator na bazi željenog izlaza, gde je uzeto da su matrice Γ za oba linearna klasifikatora ispunjene jedinicama.

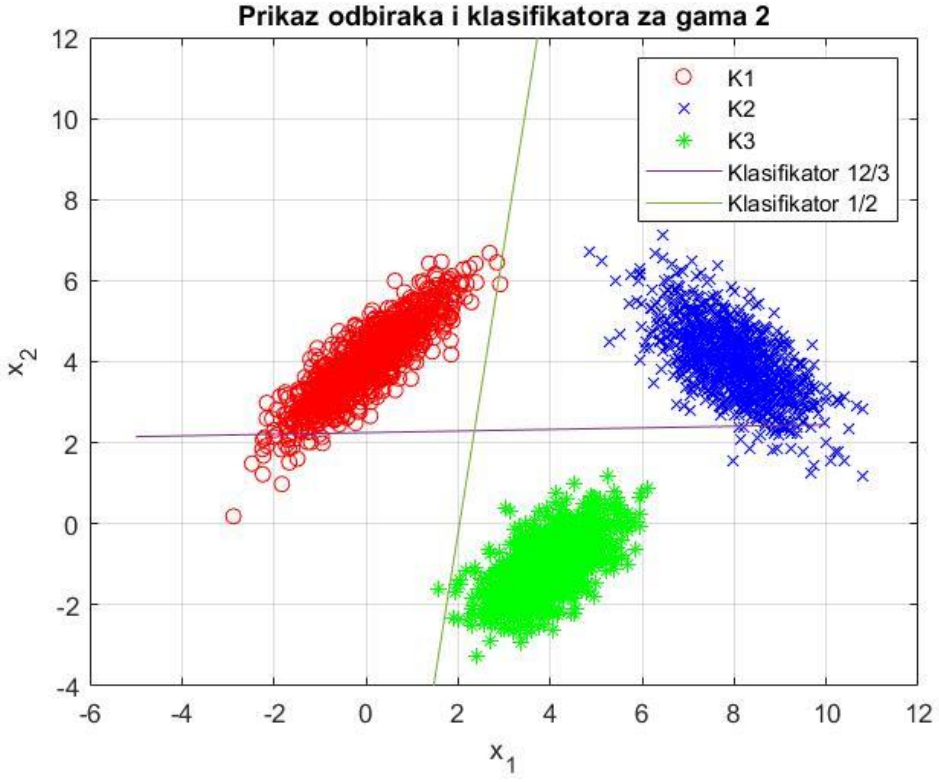


Slika 27. Klasifikacija pomoću metode željenog izlaza

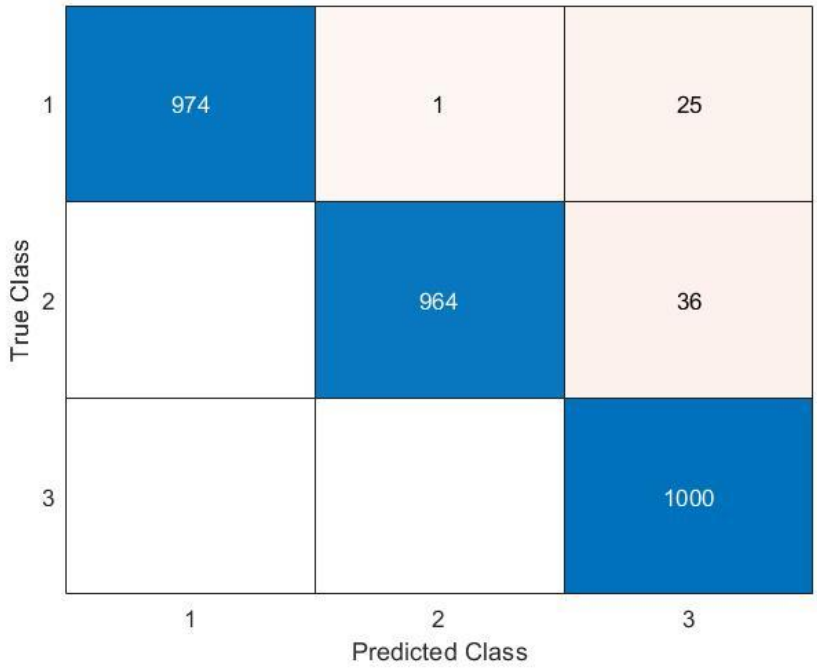


Slika 28. Konfuziona matrica za klasifikaciju pomoću metode željenog izlaza

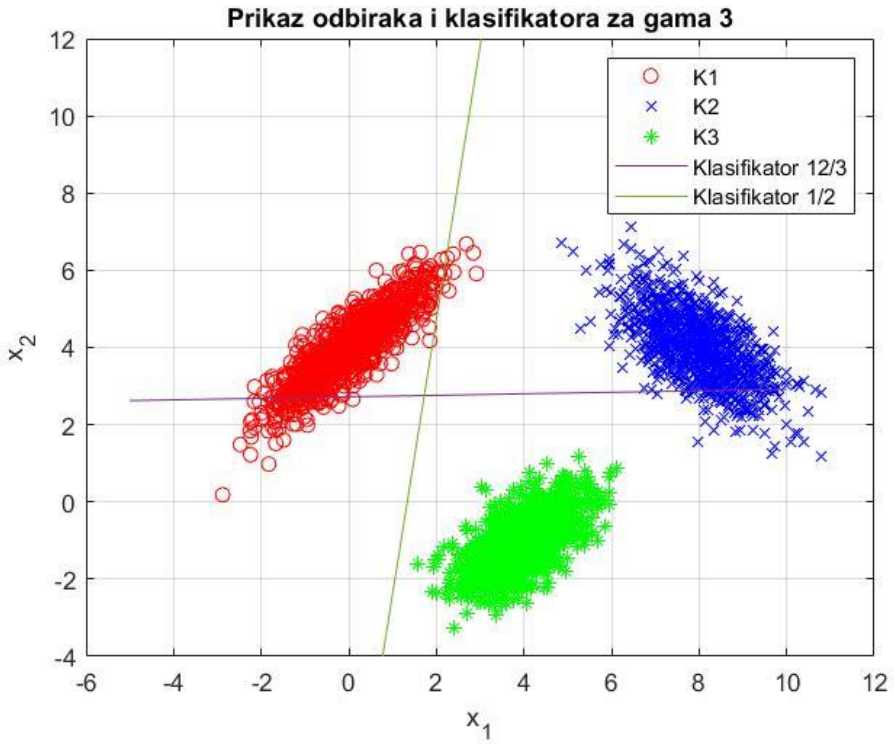
Pošto je u zadatku traženo i da se ispita kako utiču vrednosti elemenata matrice željenog izlaza Γ na dobijeni klasifikator, isti postupak ćemo ponoviti i sa nešto drugačijim matricama Γ. Prvi klasifikator daje veći težinu klasi 3 u odnosu na klase 1 i 2, a zatim drugi klasifikator daje veću težinu klasi 2 u odnosu na klasu 1. Ovakvi klasifikatori dobijeni metodom željenog izlaza prikazani su na sledećim slikama, gde je na prve dve slike prikazan grafik i konfuziona matrica za težinski faktor 2, a na druge dve za težinski faktor 3.



Slika 29. Klasifikacija pomoću metode željenog izlaza sa težinskim faktorom



Slika 30. Konfuziona matrica za klasifikaciju pomoću metode željenog izlaza sa težinskim faktorom

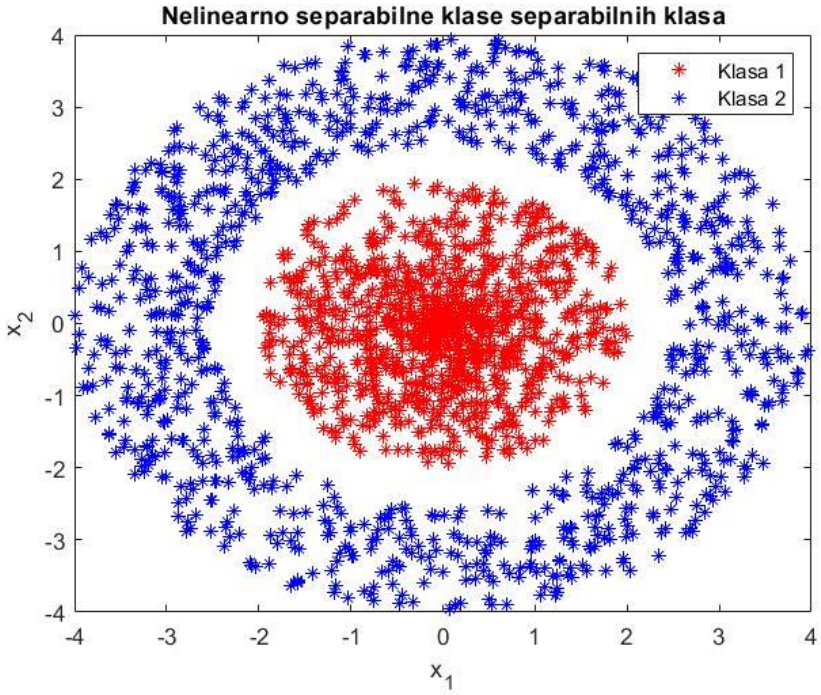


Slika 29. Klasifikacija pomoću metode željenog izlaza sa težinskim faktorom

Slika 32. Konfuziona matrica za klasifikaciju pomoću metode željenog izlaza sa težinskim faktorom

Zadatak 3.2.

Generisati dve klase dvodimenzionalnih oblika koje jesu seperabilne, ali ne linearno, pa isprojektovati kvadratni klasifikator metodom po želji.



Slika 33. Nelinearno separabilne klase

Za zadate klase se mora projektovati kvadratni klasifikator čiji je opšti oblik dat formulom

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

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

Sada je dodatno potrebno odrediti i matricu Q, a jedan od načina kako se to može rešiti jeste prividnom linearizacijom kvadratnog klasifikatora.

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

$$= \sum_{i=1}^{\frac{n(n+1)}{2}} \alpha_i y_i + \sum_{i=1}^{n} v_i x_i + v_0$$

$$= \begin{bmatrix} \alpha_1 & \alpha_2 \dots & \alpha_{\frac{n(n+1)}{2}} & v_1 & v_2 \dots & v_n & v_0 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \dots \\ y_{\frac{n(n+1)}{2}} \\ x_1 \\ \dots \\ x_n \end{bmatrix} =$$

$$= W^T Z + v_0$$

Iz priloženih jednačina se vidi da je kvadratni klasifikator sveden na linearni, međutim sada je dimenzija vektora Z mnogo veća i iznosi $n + \frac{n(n+1)}{2} = \frac{n(n+3)}{2}$ pa je potreban veliki broj uzoraka za računanje statistika višeg reda.

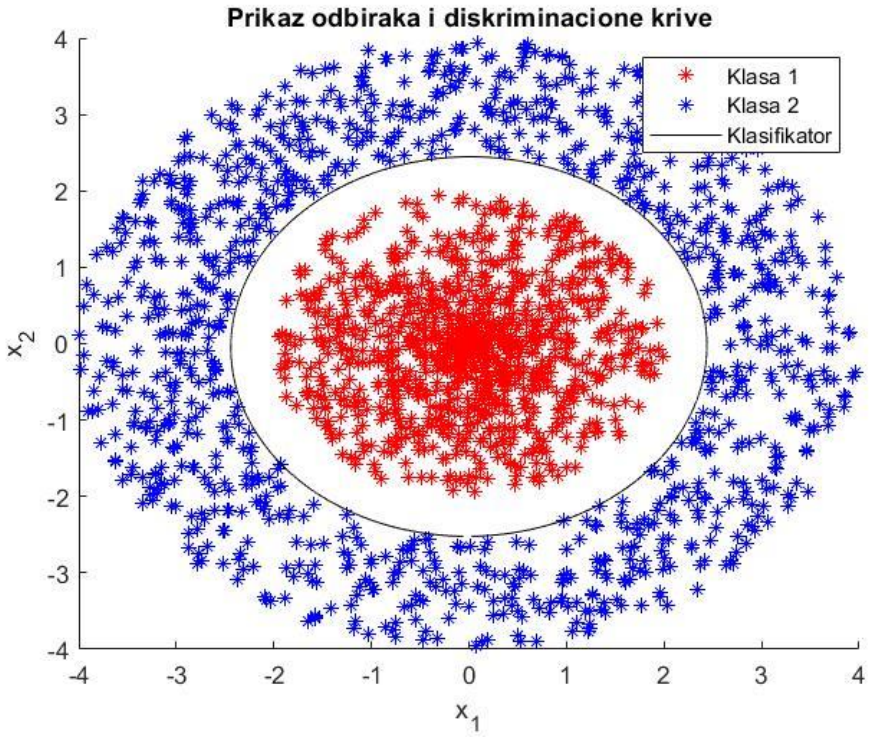
Pošto u zadatom problemu postoje dve klase nije veliki problem odrediti tražene vektore, i to će se raditi metodom željenog izlaza kao za linearan klasifikator, a traženi vektori će imati sledeći oblik

$$U = \begin{bmatrix} -1 & 1 \\ -X & Y \\ -X_1^2 & Y_1^2 \\ -X_2^2 & Y_2^2 \\ -2X_1X_2 & 2Y_1Y_2 \end{bmatrix}$$

$$\Gamma = [1 \;\; 1 \dots \;\; 1]^T$$

gde je uzeto da su svi željeni izlazi jednaki 1.

Diskriminaciona kriva kvadratnog klasifikatora koja razdvaja nelienarno separabilne klase generisane za ovaj zadatak prikazana je na sledećem grafiku.



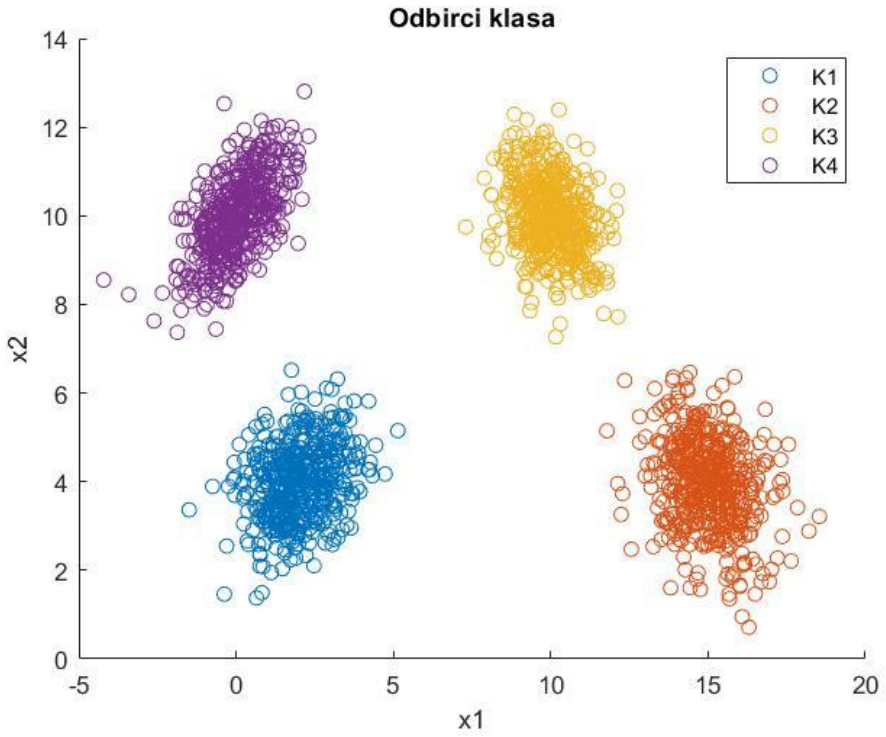
Slika 34. Klasifikacija pomoću kvadratnog klasifikatora

Zadatak 4

Zadatak 4.1.

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.

Na sledećem grafiku je prikazano 500 odbiraka iz 4 linearno separabilne klase.



Slika 35. Četiri generisane linearno separabilne klase

Potrebno je izvršiti klasterizaciju u 4 klastera nekom od metoda klasterizacije – za ovaj deo zadatka je odabrana C-mean klasterizacija, koja se bazira na najbližim srednjim vrednostima. Postoji veliki broj metoda klasterizacije koji se bazira na srednjoj vrednosti jer je minimzacija kovarijacionih matrica zahtevna.

Algoritam podrazumeva postavljanje kriterijuma na osnovu kog se vrši klasterizacija, a najčešće se koristi

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

gde je je S_w matrica unutarklasnog rasejanja, a S_m miksovana matrica rasejanja. Ukoliko se usvoji da je zajednička srednja vrednost odbiraka M_0 jednaka 0, a matrica unutarklasnog rasejanja jednaka jediničnoj matrici, pri reklasifikaciji X_j-tog sempla iz klase k_i u j-tu klasu tokom l-te iteracije priraštaj kriterijuma postaje

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

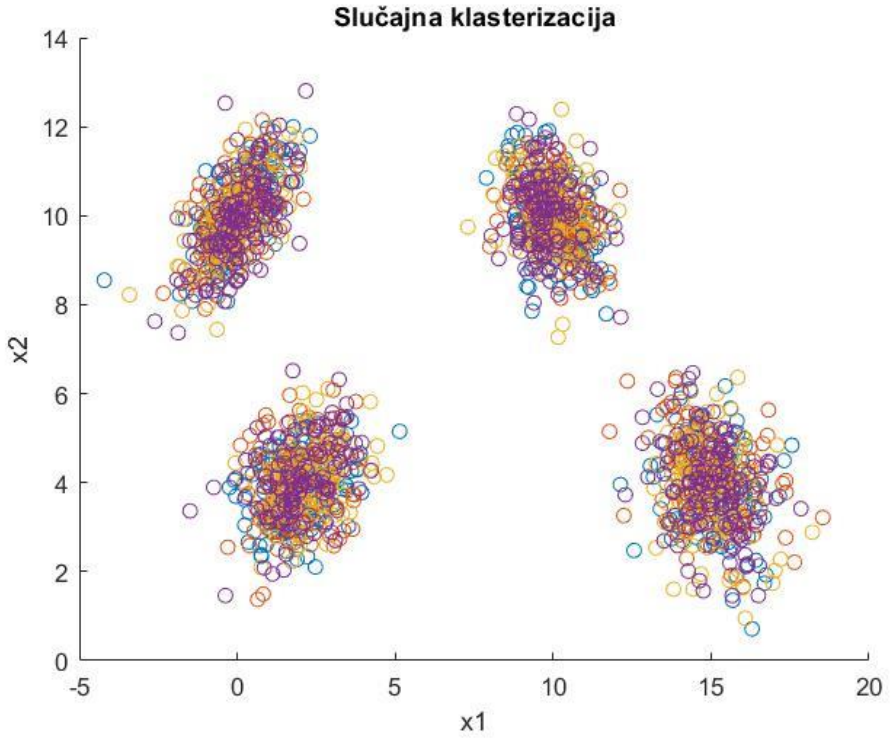
Pošto je drugi član nezavistan od j, reklasifikacija zavisi samo od prvog člana, i što je on manji to je priraštaj negativniji i X_i se pridružuje onoj klasi za koju je on najmanji:

$$\|X_i - M_t(l)\| = \min_j \|X_i - M_j(l)\| \quad \rightarrow X \in \omega_t$$

Iz zadatih jednačina može se dobiti postupak c-mean klasterizacije:

1) Izabere se inicijalna klasterizacija $\Omega(0)$ i izračuna $M_1(0), M_2(0), \dots\ M_L(0)$
2) Na osnovu izračunatih $M_1(l), M_2(l), \dots\ M_L(l)$ u l-toj iteraciji reklasifikacija za svaki sempl $X_i,\ i = 1,2, \dots N$ se vrši prema najbližem vektoru srednjih vrednosti $M_j(l)$
3) Ako je bar neki sempl X_i reklasifikovan u l-toj iteraciji, ulazi se u novu $l+1$-vu iteraciju i vraća se na korak (2). Ukoliko u tekućoj iteraciji nijedan sempl nije reklasifikovan, algoritam reklasifikacije se završava.

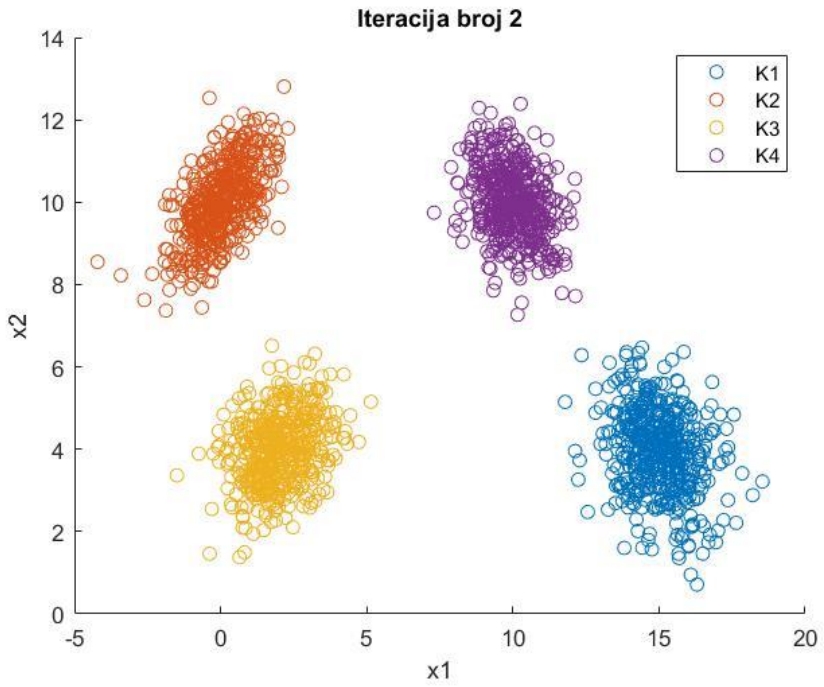
Na narednoj slici prikazana je inicijalna slučajna klasterizacija, sa jednakom verovatnoćom za svaku od klasa.



Slika 36. Slučajna klasterizacija klasa

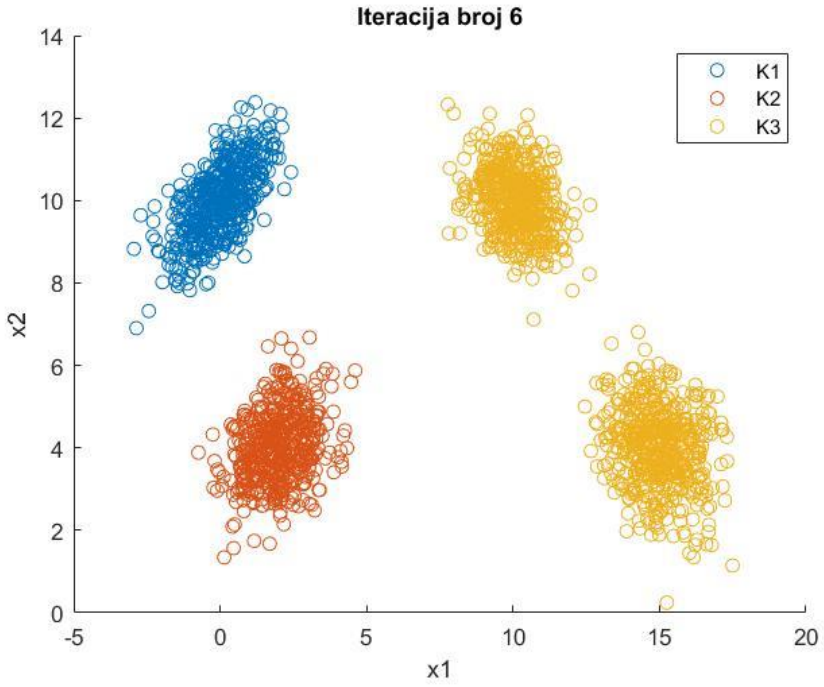
Nakon što je algoritam pokrenut par puta, zaključujemo da bez obzira na početnu klasterizaciju, on uvek uspešno separatiše klase, bar linearno separabilne. Iz ovoga zaključujemo da je C-mean metoda neosetljiva na početnu klasterizaciju, i ona eventualno može uticati samo na broj iteracija koji je potreban da bi se došlo do uspešne klasifikacije.

Na sledećoj slici je prikazan konačan rezultat klasterizacije, odakle vidimo da je raspodela oblika identična početnoj, te da je algoritam konvergirao ka tačnom rešenju.



Slika 37. Klasterizacija pomoću C-mean algoritma

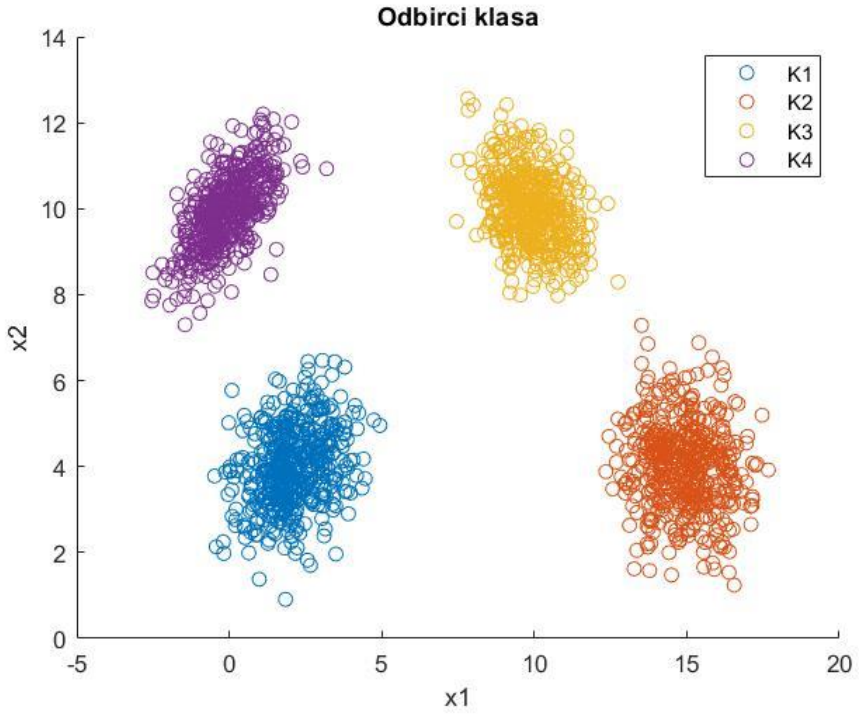
Međutim, mana ovog algoritma je što neće dati tačan rezultat ako nemamo apriorno znanje o broju klasa, ili ako je ono pogrešno. Na narednom grafiku prikazan je jedan primer gde je umesto 4 zadate klase, bilo zadato 3 klase.



Slika 38. Klasterizacija pomoću C-mean algoritma za pogrešan broj klasa

Zadatak 4.2.

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.



Slika 39. Četiri generisane linearno separabilne klase

U ovoj tački zadatka će biti izvršena klasterizacija metodom maksimalne verodostojnosti. Ovaj metod podrazumeva da je poznat broj klasa L i da su funkcije gustine verovatnoća oblika iz pojedinih klasa Gausovske i da je stoga združena funkcija gustine verovatnoće svih oblika zapravo mešavina Gausovskih raspodela:

$$f(X) = \sum_{i=1}^{L} P_i f_i(X), \qquad f_i(X) \sim N(M_i, \Sigma_i), \qquad i = 1,2, \dots L$$

Pod ovom pretopostavkom se primena maksimalne verodostojnosti svodi na maksimizaciju funkcije

$$\prod_{j=1}^{N} f(X_j)$$

po parametrima $P_i, M_i, \Sigma_i,\ i = 1,2, \dots L$ uz uslov $\sum_{j=1}^{L} P_i = 1$. Kriterijum J koji treba optimizovati je

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

Pronalaženjem izvoda datog kriterijuma redom po P_i, M_i, Σ_i se dobijaju sledeće jednačine, koje se koriste u algoritmu klasterizacije metodom maksimalne verodostojnosti:

(1) $$q_{ij} = \frac{P_j f_j(X_i)}{f(X_i)}$$

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

(3) $$M_j = \frac{1}{NP_j}\sum_{i=1}^{N} q_{ij} X_i$$

(4) $$\Sigma_j = \sum_{j=1}^{N} q_{ij}(X_i - M_j)(X_i - M_j)^T$$

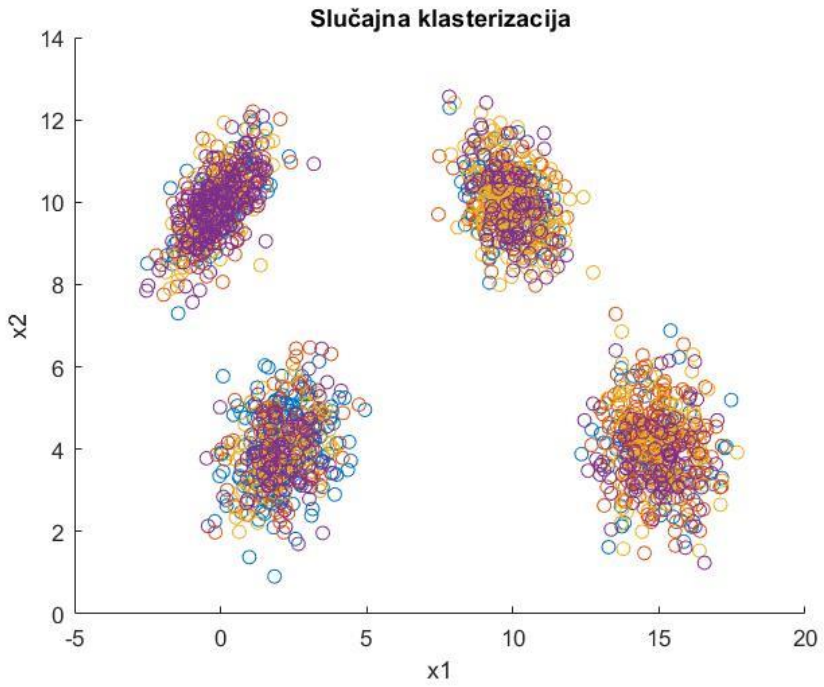
Sistem jednačina (1)-(4) je implicitan, i nema rešenje u zatvorenoj formi pa je potrebno rešavati ga numerički. Zato se i algoritam za klasterizaciju metodom maksimalne verodostojnosti svodi na numeričko rešavanje problema, kroz sledeće korake:

(1) Odredi se proizvoljna inicijalna početna klasterizacija *Ω(0).*
(2) Brojač iteracija l se inicijalizuje na 0.
(3) Na osnovu početne klasterizacije se estimiraju $P_j(l), M_j(l)\ i\ \Sigma_j(l)$ iz formula (2)-(4).
(4) Iz fromule (1) se odredi q_{ij}.
(5) Brojač iteracija l se uveća za 1, a zatim se na osnovu izračunatog q_{ij} odrede $P_j(l), M_j(l)\ i\ \Sigma_j(l)$ iz formula (2)-(4).
(6) Ponovo se iz formule (2) odredi q_{ij}.
(7) Pamti se prethodno izračunato $q_{ij}(l-1)$ i trenutno $q_{ij}(l)$, i ukoliko važi

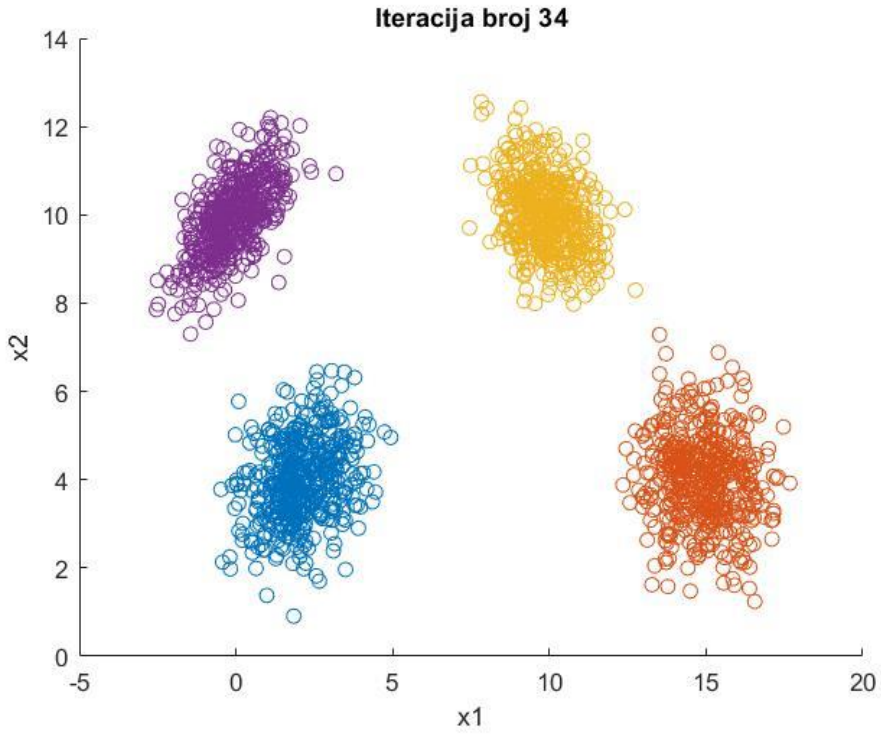
$$\max_{i,j}\left(q_{ij}(l) - q_{ij}(l-1)\right) < err$$

onda se algoritam završava i određeni su svi klasteri. Ukoliko ovaj uslov nije ispunjen, vraća se na korak (4). Parametar err predstavlja preciznost koja mora biti zadovoljena, i najčešće se nalazi u opsegu $(10^{-2}, 10^{-3})$.

Na narednim slikama prikazana je inicijalna slučajna klasterizacija, a zatim i 4 potpuno odvojena klastera kao rezultat.

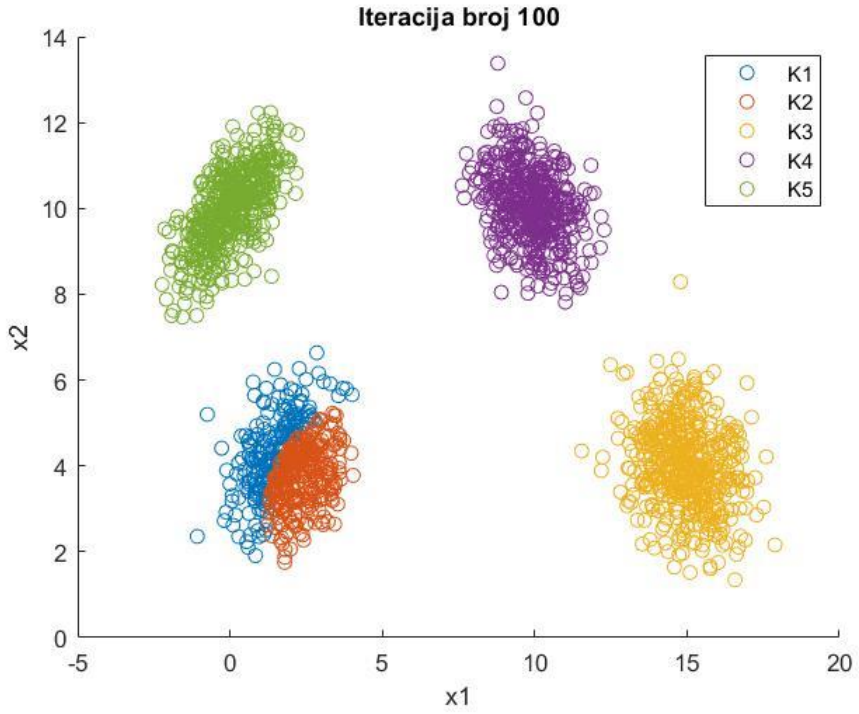


Slika 40. Slučajna klasterizacija



Slika 41. Klasterizacija metodom maksimalne verodostojnosti

Kao i algoritam C-means, i metod maksimalne verodostojnosti je osetljiv na apriorno poznavanje broja klasa, i neće na ispravan način predvideti klastere ukoliko je broj klasa pogrešan. Na sledećoj slici je prikazano klasterovanje 4 klasa, sa unetim pogrešnim brojem klasa 5.

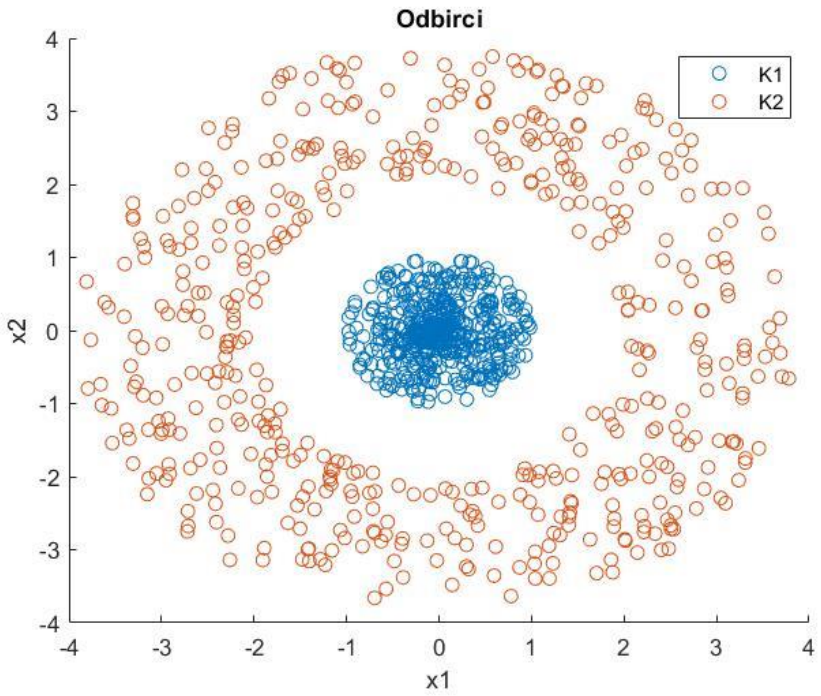


Slika 42. Klasterizacija metodom maksimalne verodostojnosti za pogrešan broj klasa

Zadatak 4.3.

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.

Generisani odbirci za dve linearno nesaparabilne klase prikazani su na sledećem grafiku.



Slika 43. Dve generisane linearno neseparabilne klase

Metod kvadratne klasterizacije je još jedan od metoda parametarkse klasifikacije odbiraka, koji liči na c-mean metod klasterizacije ali za razliku od njega kod metoda kvadratne klasterizacije se pri određivanju granice između klastera koriste i kovarijacione matrica odbiraka iz klasa, zbog čega se ovaj algoritam može primeniti i na nelinearno separabilne klase kao u ovom primeru. Algoritam se sastoji iz sledećih koraka:

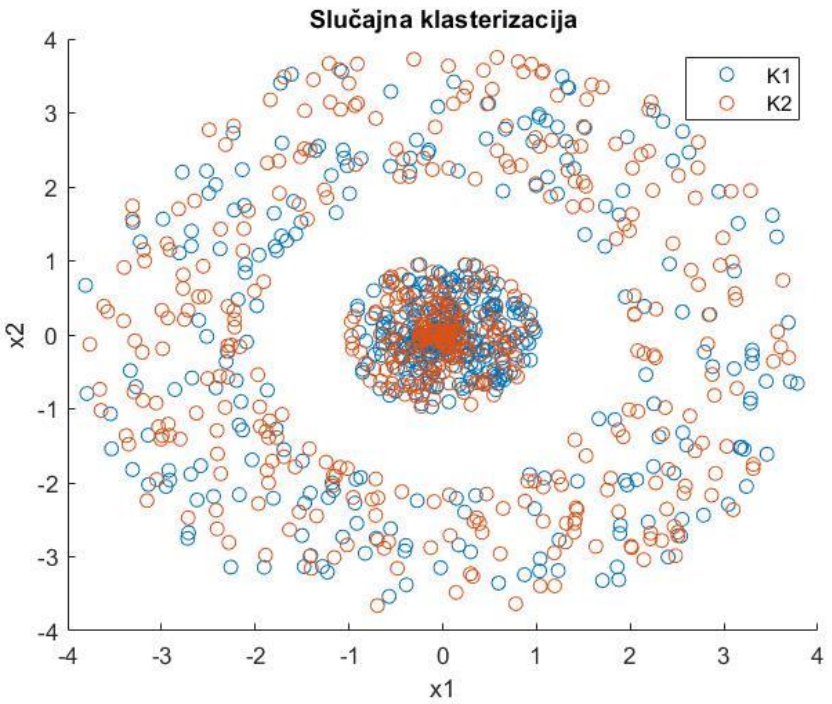
(1) Odredi se proizvoljna inicijalna početna klasterizacija $\Omega(0)$.
(2) Brojač iteracija l se inicijalizuje na 0.
(3) Na osnovu početne klasterizacije se estimiraju $P_i(0), M_i(0)\ i\ \Sigma_i(0)$ za $i = 1,2, \dots L$.
(4) U l-toj iteraciji se svako odbirak X_j reklasifikuje prema sledećem izrazu

$$\frac{1}{2}\left(X_j - M_t(l)\right)^T {\Sigma_t}^{-1}\left(X_j - M_t(l)\right) + \frac{1}{2}\ln|\Sigma_t(l)| - \frac{1}{2}\ln P_t(l), \quad t = 1,2, \dots L$$

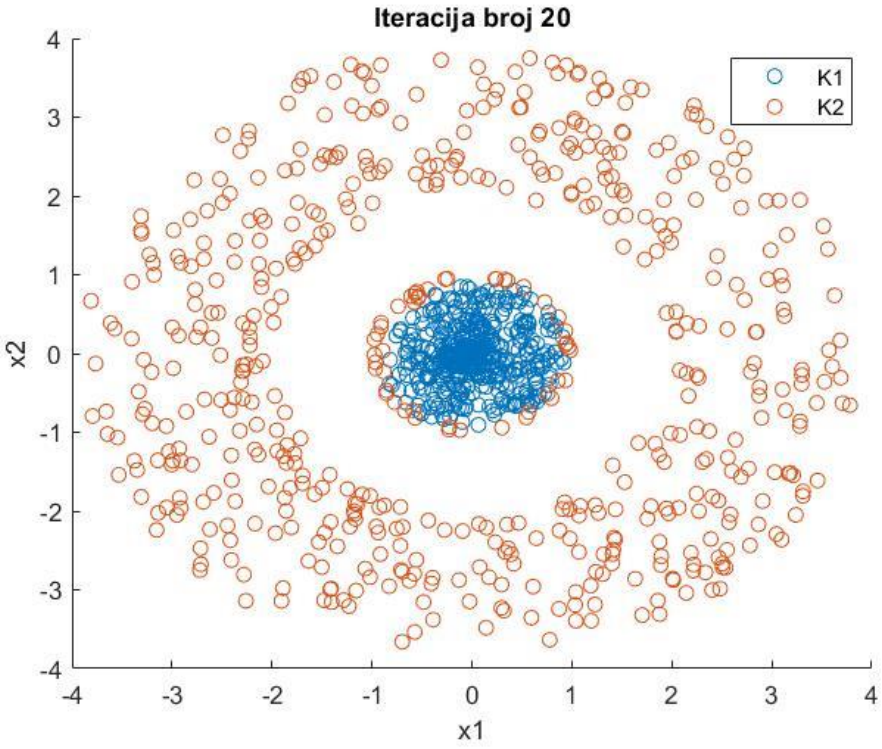
(5) Ako je bar neki sempl X_i reklasifikovan u l-toj iteraciji, ulazi se u novu $l+1$-vu iteraciju i vraća se na korak (2). Ukoliko u tekućoj iteraciji nijedan sempl nije reklasifikovan, algoritam reklasifikacije se završava.

Primenom ovog algoritma se dobijaju hiperpovrši drugog reda između klastera, a za njega se takođe zahteva apriorono poznavanje broja klastera. Takođe, algoritam je vrlo osetljiv na počenu klasterizaciju, pa je u nju potrebno ugraditi apriorno znanje da bi algoritam konvergirao.

Na narednom grafiku je ponovo prikazana slučajna klasterizacija, a zatim i finalni rezultat.

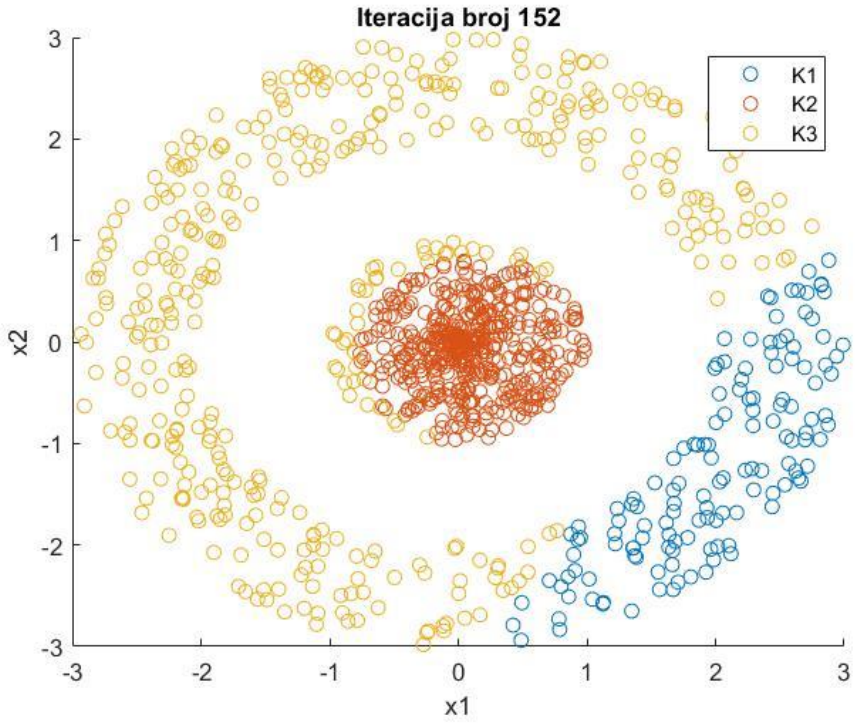


Slika 44. Slučajna klasterizacija



Slika 45. Klasterizacija metodom kvadratne klasterizacije

Ukoliko je pretpostavljeni broj klasa netačan, na primer 3, 2 klase će biti podeljene na 3, tako da zaključujemo da i ovaj algoritam zahteva apriorno poznavanje broja klastera Ovakav primer je prikazan na sledećoj slici.



Slika 46. Klasterizacija metodom kvadratne klasterizacije za pogrešan broj klasa