Univerzitet u Beogradu
Elektrotehnički fakultet

Domaći zadatak iz
Prepoznavanja oblika

Student:
Ervin Seke
2017/0046

Beograd
školska 2020/21 godina

Zadatak 1.

Tekst zadatka

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

Teorijski osvrt

Za rešavanje ovog zadatka je korišćen Bajesov test minimalne verovatnoće greške. On analizira aposteriornu verovatnoću da je oblik iz neke klase. Formula za aposterionu verovatnoću je:

$$q_i(x) = \frac{P_i f_i(x)}{\sum_{k=1}^{n} P_k f_k(x)},$$

gde je P_i apriorna verovatnoća pojave i-te klase, $f_i(x)$ funkcija gustine verovatnoće(fgv) odbiraka iz i-te klase, n broj klasa. Pošto je imenioc navedenog izraza isti za sve klase i pošto je apriorna verovatnoća pojave klase ista za sve klase dovoljno je posmatrati samo fgv klasa. Ovaj metod kaže da odbirak x pripada onoj klasi za koju je $f_i(x)$ najveće.

Baza podataka

Pri realizaciji ovog klasifikatora je korišćena baza koja sadrži po 120 slika samoglasnika A, E, I, O i U. Po 100 slika svakog samoglasnika je korišćeno da bi se odredila fgv odbiraka odgovarajuće klase, dok je po 20 slika korišćeno za testiranje klasifikatora. Kao fgv za svaki samoglasnik(klasu) je pretpostavljena Normalna Gausova raspodela. Srednja vrednosti i kovarijaciona matrica svake klase je određena na osnovu odgovarajućih 100 odbiraka klase.

Odabrana obeležja

Da bi se izdvojila obeležja bile je potrebno izvršiti predobradu slika. Prvo je slika binarizovana sa pragom 0.92, zatim je isečen samo deo slike koji sadrži slovo, a beli i crni okvir su uklonjeni.

Odabrana su sledeća 3 obeležja:

1. Odnos broja piksela koji imaju vrednost 1 u levoj polovini slike u odnosu na celu sliku. Ovo obeležje je odabrano zato što je primetno da kod slova E ovo obeležje ima vrednost manju od 0.5, kod slova U veću od 0.5, a dok kod slova A, I i O ovo obeležje bi trebalo da ima vrednost oko 0.5.
2. Za svaku kolonu slike je posmatran absolutna razlika broja piksela koji imaju vrednost 1 u gornjoj i donjoj polovini kolone slike. Kao obeležje je korišćena suma svih ovih apsolutnih razlika po svim kolonama. Ovo obeležje je korišćeno pošto je ova suma kod slova E i O blizu 0, dok je kod ostalih slova veća od 0.

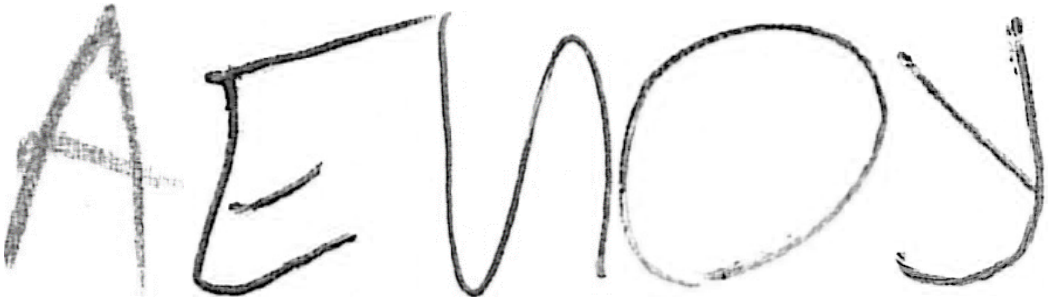
3. Posmatran je broj piksela koji imaju vrednost 1 u okviru pravougaonika koji je centriran u centru cele slike, dok su dužine njegovih stranica jednake 1/4 dužine odgovarajući stranica cele slike. Kao obeležje je korišćen prethodno naveden broj pisela u odnosu na broj piksela koji imaju vrednost 1 u celoj slici.

Rezultati

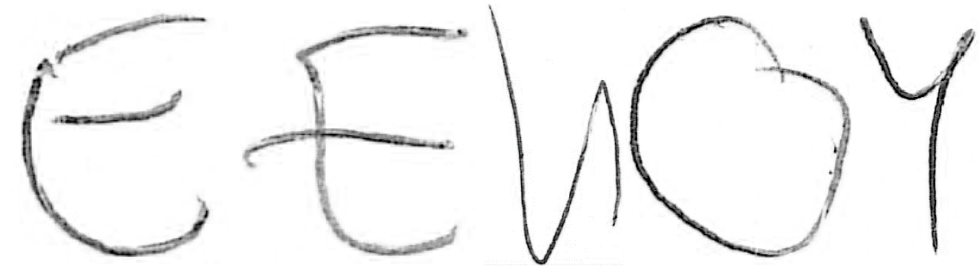
U okviru sledeće tabele se nalazi konfunziona matrica dobijenih rezultata na osnovu test skupa podataka koji sadrži po 20 oblika od svake klase. Za svaki element matrice iznad njega se nalazi slovo klase kojem pripada odgovarajući oblik , dok je sa leve strane slovo klase u koju je klasifkovanon. Greška klasifikacije je 10%.

	A	E	I	O	U
$\hat{A}$	18	0	1	0	0
$\hat{E}$	0	18	0	1	0
$\hat{I}$	0	0	16	0	1
$\hat{O}$	0	1	1	19	0
$\hat{U}$	2	2	2	0	19

Na sledećim slikama se nalaze pravilno klasifikovana slova:



Na sledećim slikama se nalaze nepravilno klasifikovana slova:



Navedeni nepravilno klasifikovani oblici su klasifikovani redom kao: O,U,U,E i I. Na nepravilno klasifikovanim oblicima se vidi da odstupaju od pretpostavki koje su uzete za obeležja, ali i on izgleda slova koji su korišćeni za treniranje. Da bi se dobili bolji rezultati klasifikacije mogla bi da se uzmu neka obeležja koja uzimaju više heurističkih informacija o zapisu samoglasnika, više obeležja u obzir klasifikatora ili veća baza podataka.

Zadatak 2.

Tekst zadatka

Po ugledu na primer 4.3 iz dokumenta Predavanje 2, generisati po N = 500 odbiraka iz dveju dvodimenzionih klasa.

a) Na dijagramu prikazati odbirke.

b) Generisati geometrijsko mesto tačaka sa konstantnom vrednošću funkcija gustina verovatnoće pa ih prikazati na dijagramu u prostoru oblika.

c) Isprojektovati Bajesov klasifikator minimalne greške i na dijagramu, zajedno sa odbircima, skicirati klasifikacionu liniju, pa proceniti verovatnoću greške.

d) Ponoviti prethodnu tačku za neki drugi klasifikator po izboru.

e) Za klase oblika generisanih u prethodnim tačkama, isprojektovati Wald-ov sekvencijalni test pa skicirati zavisnost broja potrebnih odbiraka od usvojene verovatnoće grešaka prvog, odnosno drugog tipa.

Teorijski osvrt

U okviru ovog zadatke su pored Bajesovog testa minimalne verovatnoće greške korišćeni i Neyman-Pearson-ov test hipoteza i Wald-ov sekvencijalni test.

Neyman-Pearson-ov test hipoteza pretpostavlja da je verovatnoća greške drugog tipa ε_2 jednak nekoj konstantnoj vrednošću ε_0, dok verovatnoću greške prvog tipa hoćemo da minimzujemo. Uglavnom se za vrednost ε_2 uzima neka vrednost koja je bliska vrednosti ukupne verovatnoće greške kod Bajesovog testa minimalne cene. Odluka u okviru ovog testa se donosi na osnovu formule:

$$-ln\frac{f_1(x)}{f_2(x)}\begin{cases}< -\ln(\mu)\,, & x \in \omega_1\\ > -\ln(\mu), & x \in \omega_2\end{cases},$$

dok se vrednost parametra μ dobija iterativnim postupkom iz jednačine:

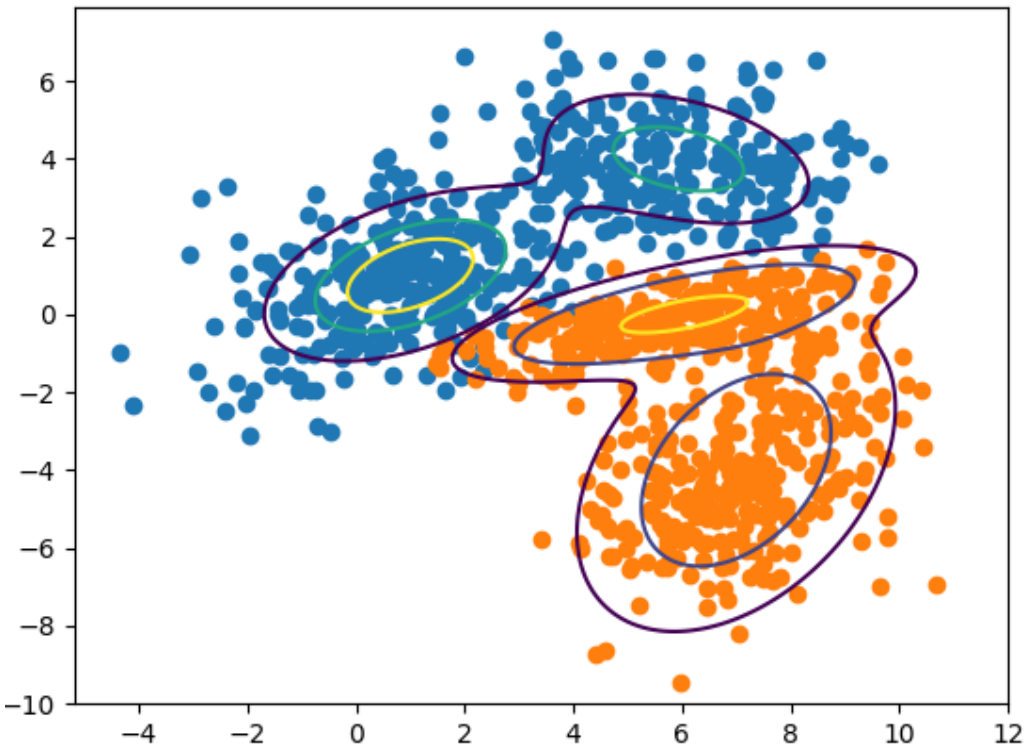
$$\int_{-\infty}^{-\ln(\mu)} f(h/\omega_2)dh = \varepsilon_2.$$

Ideja Wald-ovog sekvencijalnog testa je da se odluka donese na osnovu više oblika koji nam dolaze iz iste klase na osnovu sledeće formule:

$$\sum_{i=1}^{m} -ln\frac{f_1(x)}{f_2(x)}\begin{cases} > -ln\frac{1-\varepsilon_1}{\varepsilon_2}, & x \in \omega_1\\ ostalo,\ traži\ se\ m+1\ odbirak. & \\ < -ln\frac{\varepsilon_1}{1-\varepsilon_2}, & x \in \omega_2\end{cases}$$

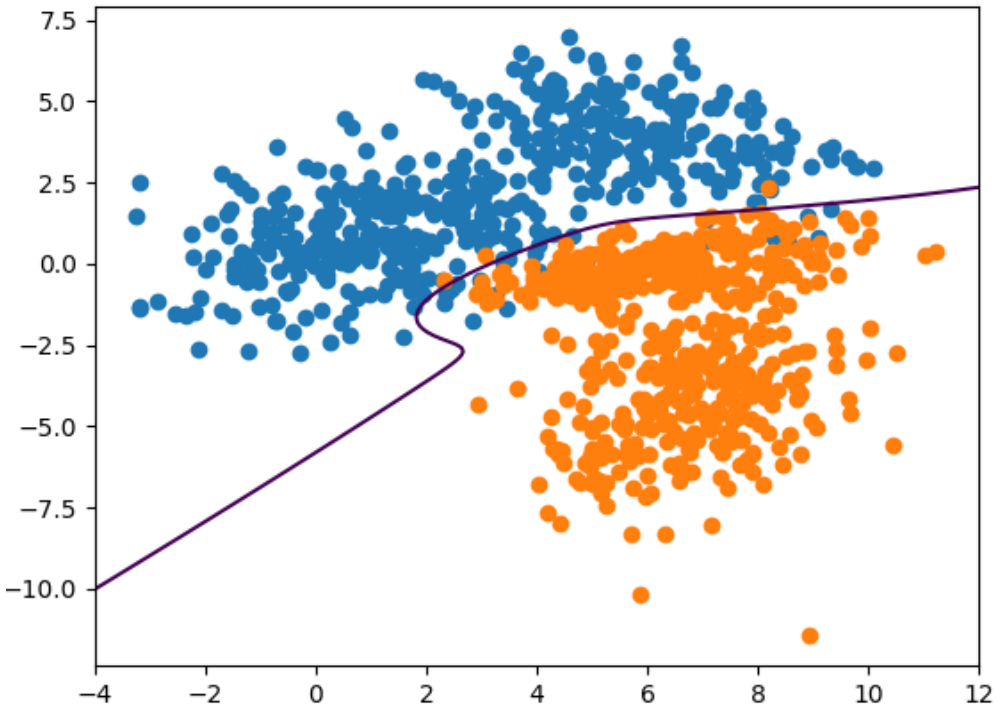
Dobijeni rezultati

Na grafiku 1. se nalaze odbirici 2 klase i geometrijska mesta tačaka sa konstantnom funkcijom gustine verovatnoće odgovarajućih klasa.



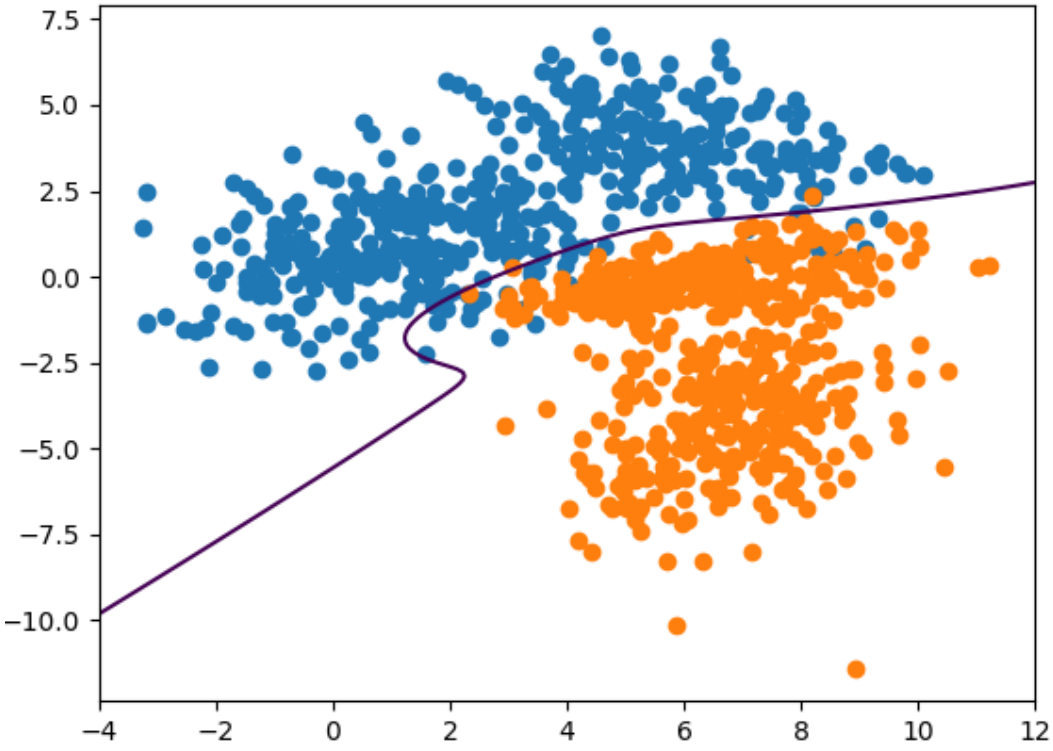
Grafik 1. Odbirci klasa i gmt sa konstantnom fgv

Na grafiku 2. se nalazi klasifikaciona linija dobijena pomoću Bajesovog testa minimalne verovatnoće greške. Dobijene greške su $\varepsilon_1 = 0{,}73\%$, $\varepsilon_2 = 1{,}12\%$ i $\varepsilon = 1{,}85\%$.



Grafik 2. Bajesov test minimalne verovatnoće greške

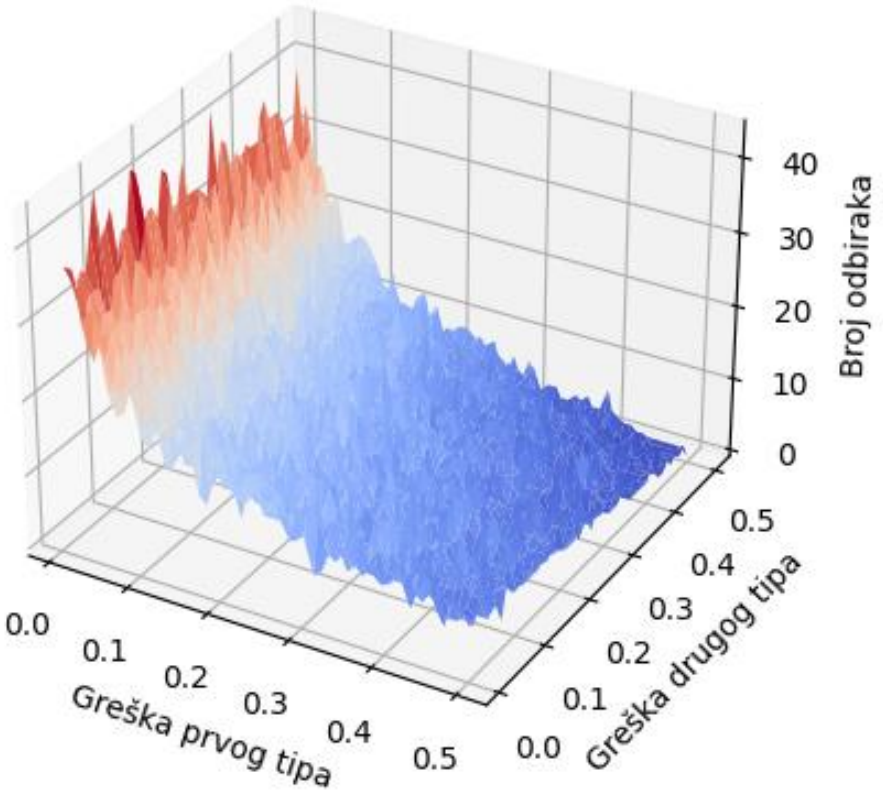
Na grafiku 3. se nalazi klasifikaciona linija dobijena pomoću Neyman-Pearson-ovog testa hipoteza za vrednost $\varepsilon_2 = \varepsilon_0 = 1{,}8\%$, . Dobijene greške su $\varepsilon_1 = 0{,}33\%$, $\varepsilon_2 = 1{,}8\%$ i $\varepsilon = 2{,}13\%$.



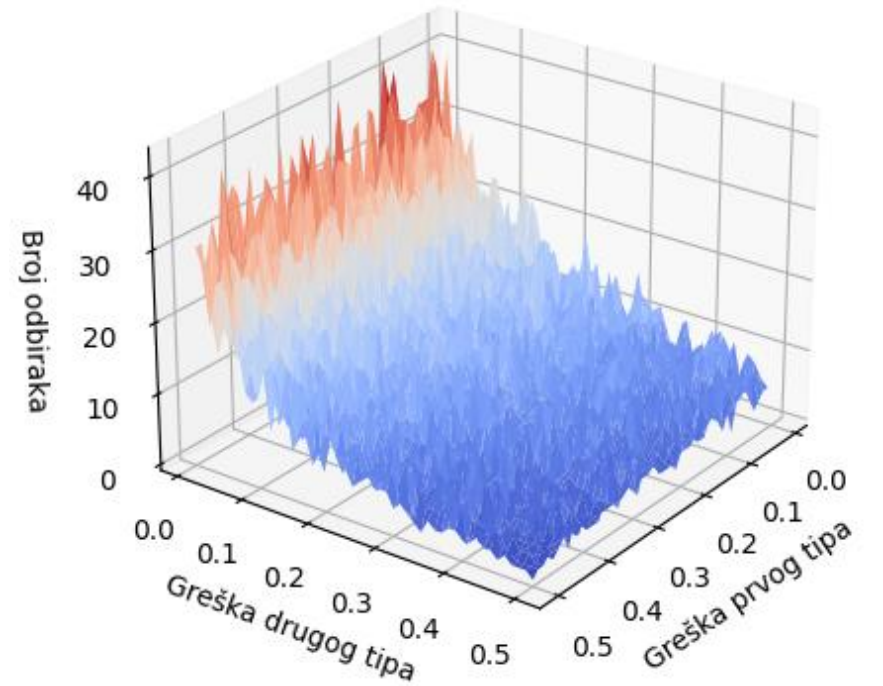
Grafik 3. NP test

Sa prethodna 2 grafika se vidi da će se više elemenata prve klase pravilno klasifikovati kod NP testa, dok će se greška za drugu klasu povećati. NP test ima veću ukupna greška, iz čega zaključujemo da Bajesov test minimalne verovatnoće greške projektuje klasifikator sa minimalnom verovatnoćom greške i da ne može da se projektuje klasifikator zasnovan na fgv koji će dati manju ukupnu verovatnoću greške.

Na grafiku 4. se nalazi zavisnot potrebnog broja odbiraka od grešaka prvog i drugog tipa kada dolaze odbirci iz prve klase, dok se na grafiku 5. vidi ista zavisnost ali kada dolaze odbirci iz druge klase. Sa grafika 4. se vidi da broj potrebnih odbiraka brže opada sa porastom greške prvog tipa nego sa porastom greške drugog tipa, u slučaju kada nam dolaze odbirci iz druge klase je obrnuto.



Grafik 4.



Grafik 5.

Zadatak 3.

Tekst zadatka

Izabrao sam opciju 2

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.

Teorijski osvrt

Linearni klasifikator ima oblik:

$$V^T X + v_0 \begin{cases} > 0, & X \in \omega_2 \\ < 0, & X \in \omega_1 \end{cases},$$

gde je V vektor i zajedno sa v_0 čini koeficijene klasifikatora, dok je X vektor obeležja oblika koji želimo da klasifikujemo.

Kvadratni klasifikator ima oblik:

$$X^T Q X + V^T X + v_0 \begin{cases} > 0, & X \in \omega_2 \\ < 0, & X \in \omega_1 \end{cases},$$

gde je Q matrica parametara klasifikatora. Kvadratni klasifikator dvodimenzionih oblika se može linearizovati na sledeći način:

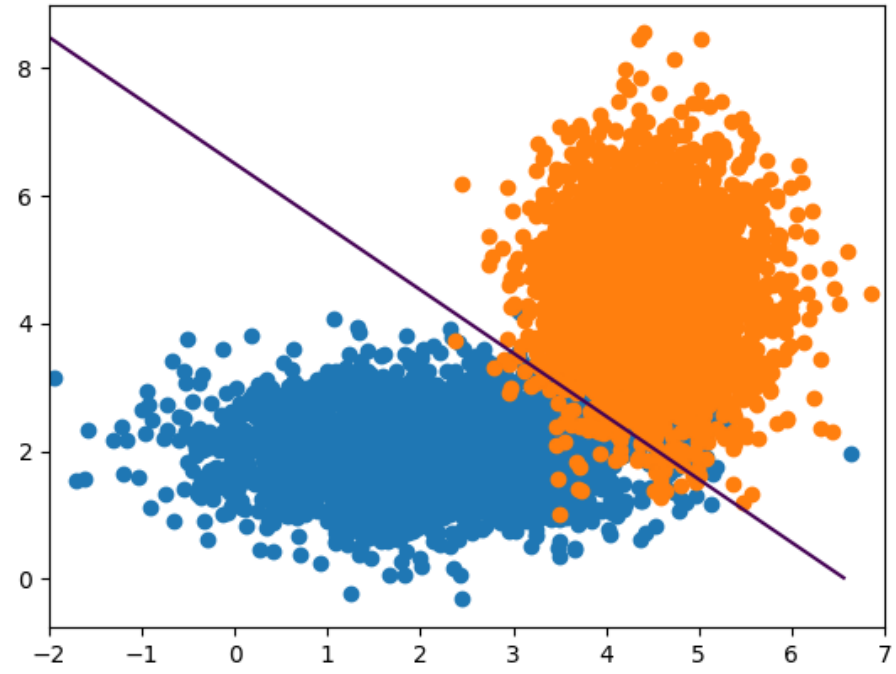
$$\begin{bmatrix} 1 & X_1 & X_2 & {X_1}^2 & 2X_1X_2 & {X_2}^2 \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ q_{11} \\ q_{12} \\ q_{22} \end{bmatrix} \begin{cases} > 0, & X \in \omega_2 \\ < 0, & X \in \omega_1 \end{cases}.$$

Optimalni linearni klasifikator želi da minimizira funkciju verovatnoće greške klasifikatora. Pri izradi zadatka je za projektovanje linearnog klasifikatora korišćena iterativna metoda koja koristi dva različita skupa podataka, jedan za obučavanje, drugi za testiranje.

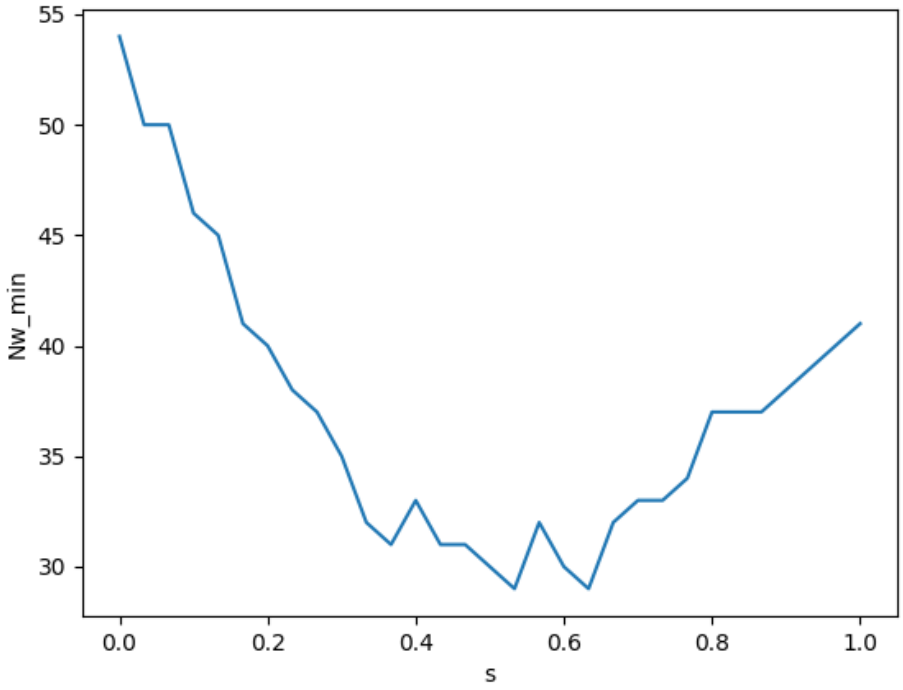
Kod metoda željenog izlaza ideja je da za svaki oblik desna strana prethodnih jednačina ima neku vrednost γ_i. To znači da imamo N(broj oblika) jednačina i M(broj parametara klasifikatora) parametara. Ovo nije uvek moguće rešiti. Jedan od načina za rešavanje ovog problema je presudolinearna regresija. Ovo rešenje sistema jednačina se svodi na metod najmanjih kvadrata.

Rezultati

Na grafiku 6. se nalazi rezultat primene linearnog klasifikatora opisanom iterativnom metodom. U slučaju da su klase skroz separabilne ovaj metod bi pronašao jako puno vrednosti parametara s i v_0 za koje je broj pogrešno klasifikovanih oblika u test skupu jednak 0,tako da je prikazana efikasnost ovog algoritma za klase koje se malo preklapaju. Na grafiku 7. je prikazana zavisnost minimalnog broja pogrešno klasifikovanih oblika u test skupu od vrednosti parametra s.



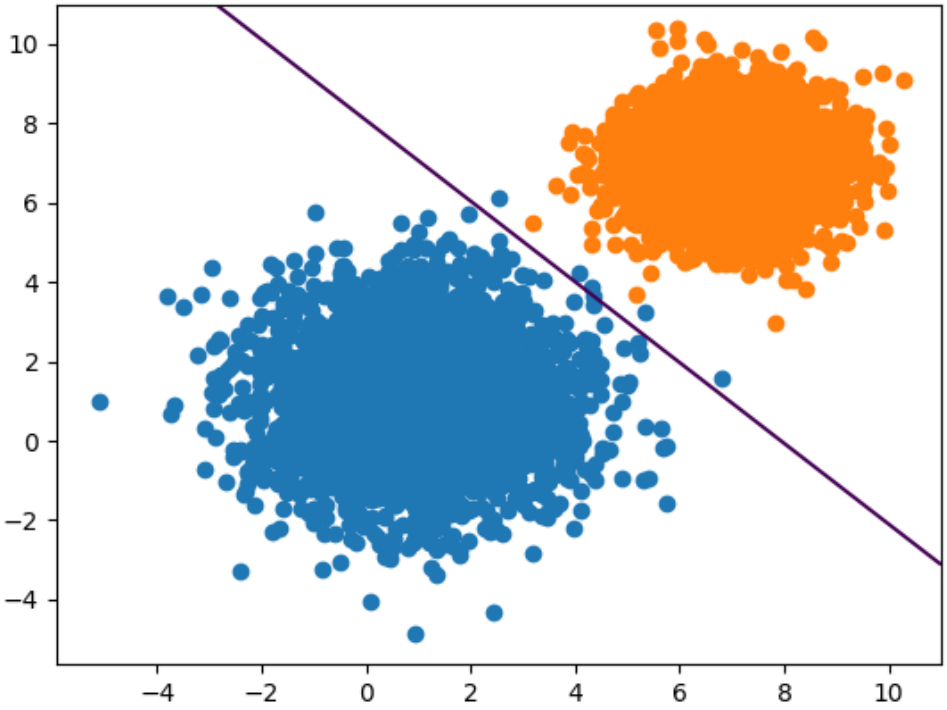
Grafik 6.



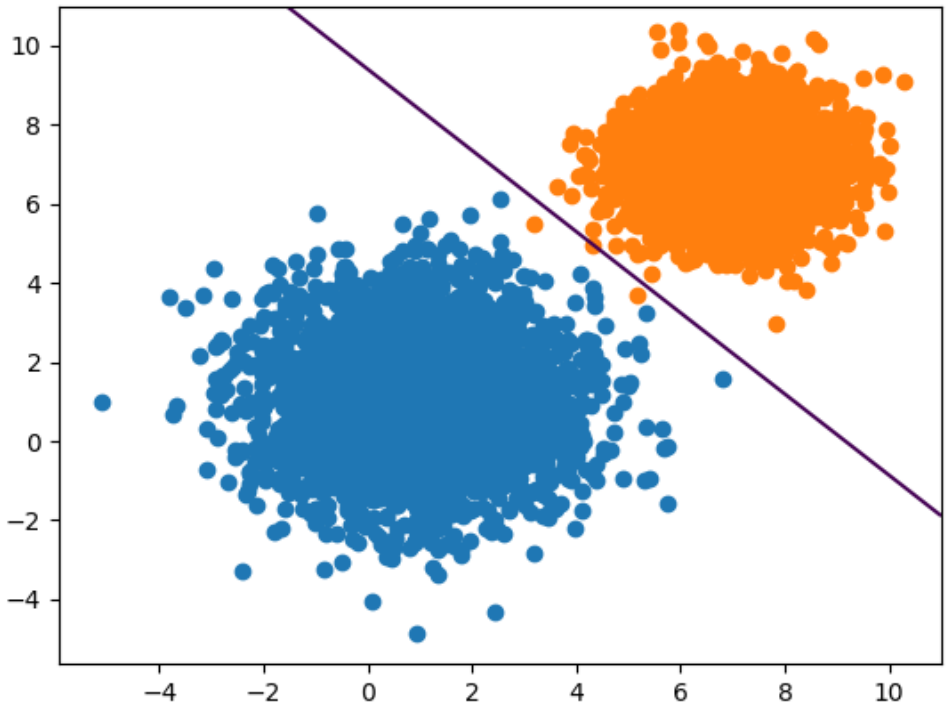
Grafik 7.

Dobili su se rezultati koji se poklapaju sa teorijskim pretpostavkama.

Na grafiku 8. se nalazi rešenje dobijeno primenom metoda željenog izlaza kada su sve vrednosti željenog izlaza jednake 1. Zbog toga što jedna klasa ima veću varijansu pojedini odbirci te klase su klasifikovani da pripadaju drugoj klasi. To se može popraviti tako što se oblicima koji se nalaze blizu diskriminacione linije zadaje veći željeni izlaz. To je prikazano na grafiku 9., gde je oblicima prve klase čije su vredosti obe koordinate veći od 2 date vrednosti željenog izlaza 8.

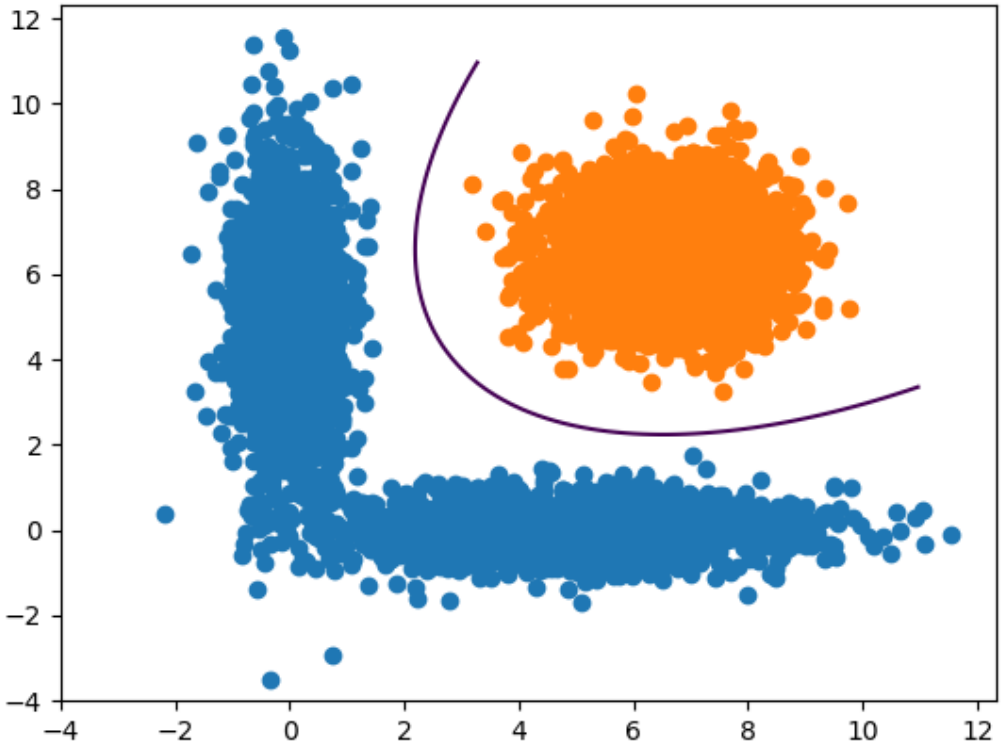


Grafik 8.

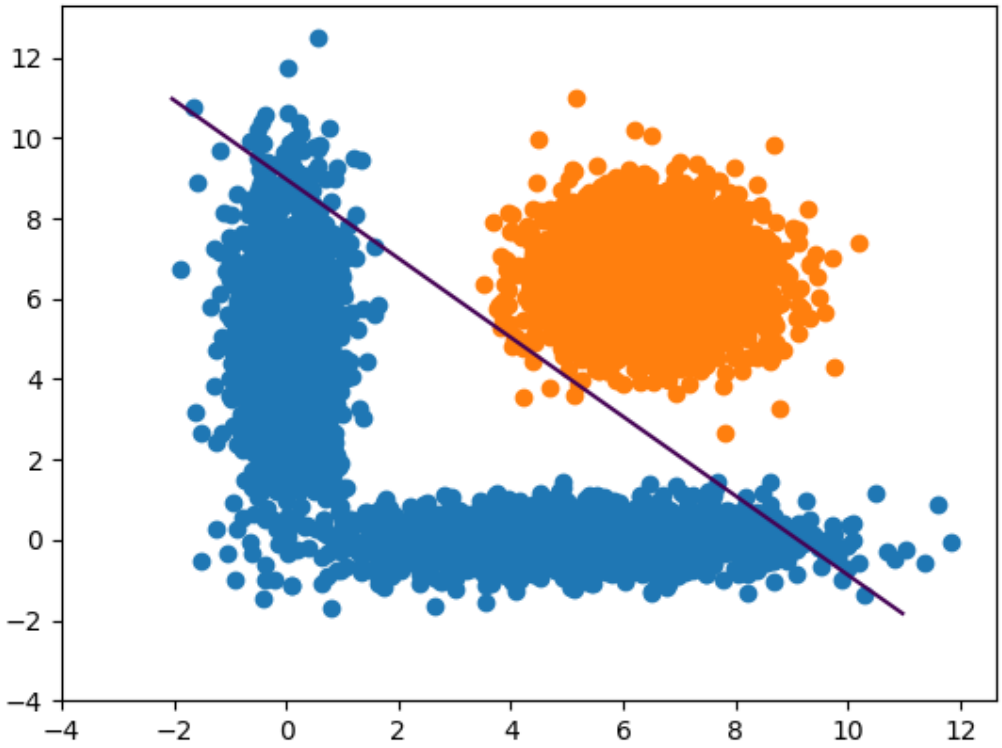


Grafik 9.

Na grafiku 10. je prikazan rezultat dobijen primenom kvadratnog klasifikatora metodom željenog izlaza. Sa grafika se vidi da je ovaj metod uspešan za klase koje su upotpunosti nelinearno separabilne. Na grafiku 11. je prikazana primena linearne klasifikatora primenom metode željenog izlaza na nelinearno separabilne klase.



Grafik 10.



Grafik 11.

Zadatak 4.

Tekst zadatka

1. Generisati po N = 500 dvodimenzionih odbiraka iz četiri klase koje će biti linearno separabilne. Preporučujem da to budu Gausovski raspodeljeni dvodimenzioni oblici. Izabrati jednu od metoda za klasterizaciju (c mean metod, metod kvadratne dekompozicije, 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.
2. 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 prethodne tačke.

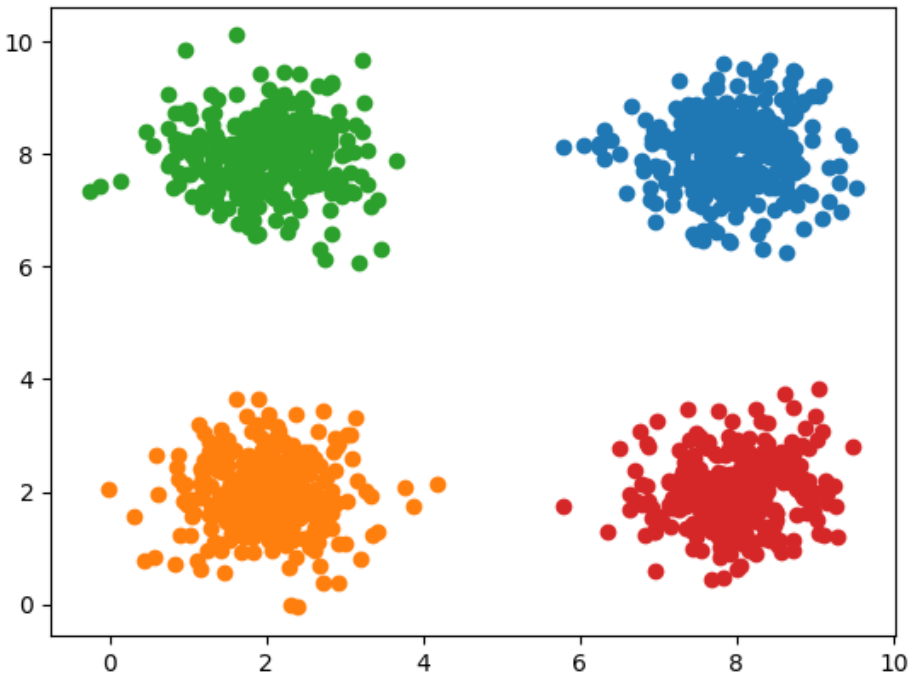
Teorijski osvrt

Osnovna ideja iterativnih metoda klasterizacije je da se iterativno prolazi kroz raspoložive oblike i da se za svaki oblik traži klasa tako da vrednost globalne kriterijumske funkcije bude što manja. Ovaj iterativni postupak se prekida posle iteracije u kojoj ni jedan oblik nije promenio klasu u kojoj se nalazi.

Kod C-mean klasterizacije oblik se uvek premešta u onu klasu čiji centar mu je najbliži. Dok se kod kvadratne dekompozicije uzima u obzir i koliko se nalazi oblika u toj klasi i kolika je varijansa te klase.

Rezultati

Kao algoritam za klasterizaciju linearno separabilnih oblike je korišćen C-mean algoritam. Na grafiku 12. su prikazane raspodele klasa korišćene pri testiranju algoritma.

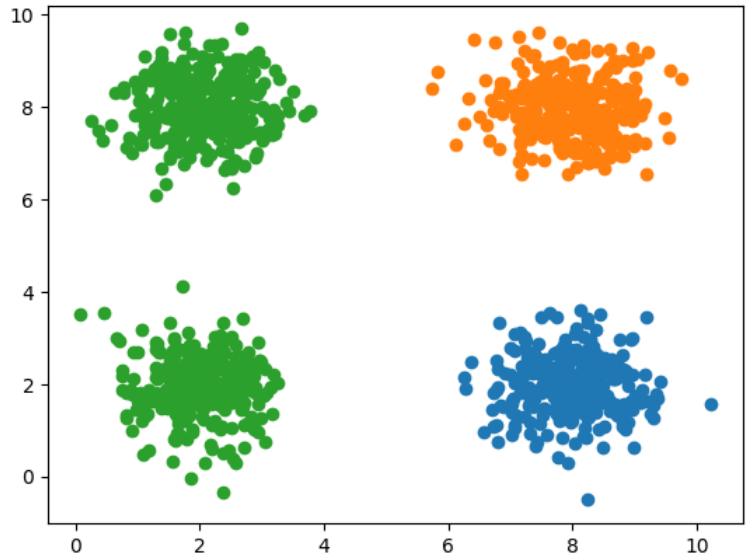


Grafik 12.

U slučaju da se na početku klasterizacije odbirci nasumično rasporede po klasama u određenom broju pokretanja algoritma se dešavalo da se svi oblici klasifikuju u istu klasu. To može da bude posledice simetričnosti među srednjim vrednostima i varijansama klasa. U slučaju kada se klasterizacija izvrši uspešno potrebno je u proseku 4 iteracije.

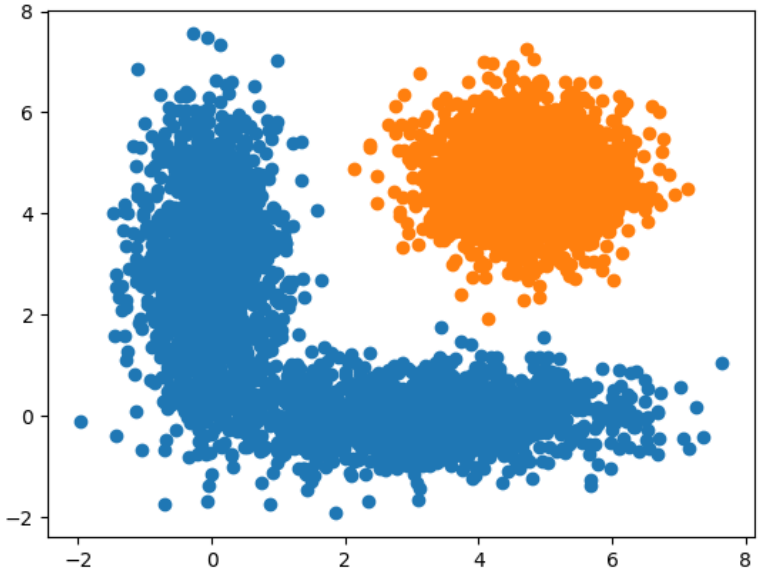
U slučaju kada je na početku klasterizacije 10% oblika ispravno klasterizovano potrebno je 3 iteracije da se uspešno izvrši klasterizacija. U slučaju kada je na početku klasterizacije 50% oblika ispravno klasterizovano potrebno je 2 iteracije da se uspešno izvrši klasterizacija.

Pošto C-mean algoritam zahteva poznavanje broja klasa unapred algoritam je za prethodno korišćene skupove oblike pokretan pod pretpostavkom da ima 3 ili 5 klasa. Pod pretpostavkom da postoji 3 klase se dobija uvek rezultat prikazan na grafiku 13. Pod pretpostavkom da postoji 5 klasa svi oblici se klasterizuju u jednu klasu.



Grafik 13.

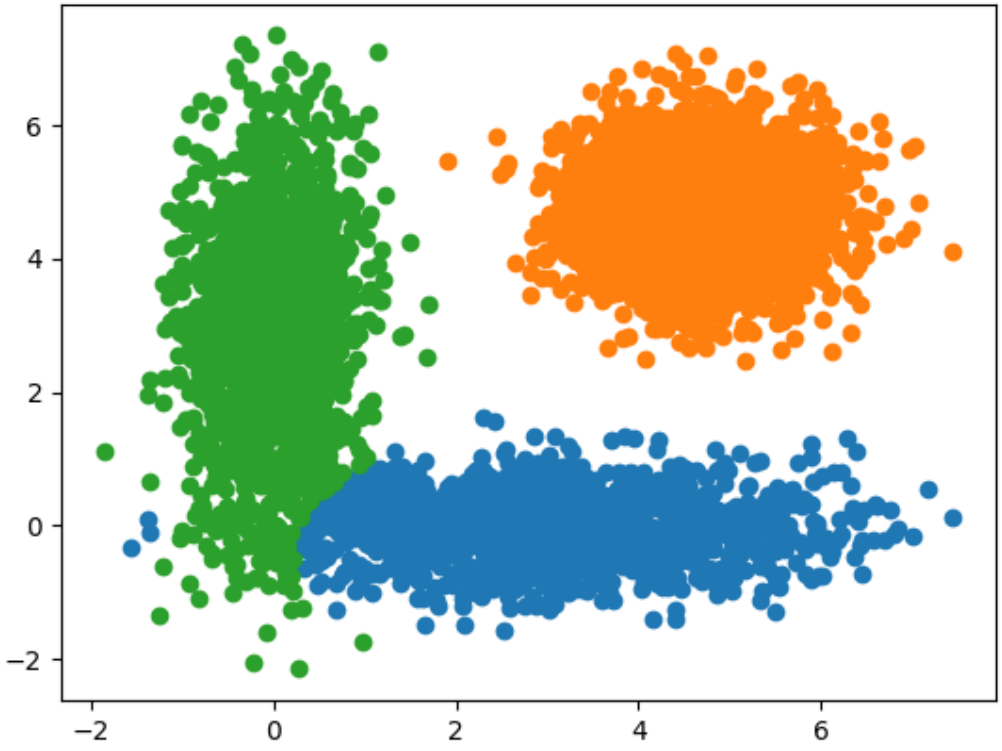
Kao algoritam za klasterizaciju nelinearno separabilnih oblike je korišćena metod kvadratne dekompozicije. Na grafiku 14. su prikazane raspodele klasa korišćene pri testiranju ovog algoritma.



Grafik 14.

Ovaj algoritam je pravilno klasterizovao oblike. U slučaju da su svi oblici nasumično raspodeljeni potrebno mu je u proseku 8 iteracija da izvrši klasterizaciju. U slučaju da je 10% oblika na početku pravilno klasterizovano u proseku je potrebno 4 iteracije, dok kad je na početku 50% oblika pravilno klasterizovano potrebno je 3 iteracije.

Pošto metod kvadratne dekompozicije zahteva poznavanje broja klasa unapred algoritam je za prethodno korišćene oblike pokretan pod pretpostavkom da ima 3. Dobijeni rezultat je prikazan na grafiku 15. Dobijeni rezultat je očekivan zbog načina na koji je formirana klasa sa bimodalnom raspodelom.



Grafik 15.