Linearna diskriminantna analiza

Zrinščak, Anamarija

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:906516

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Anamarija Zrinščak

LINEARNA DISKRIMINANTNA ANALIZA

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, veljača, 2022.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Od srca zahvaljujem svojoj mentorici, prof. dr. sc. Anamariji Jazbec na uloženom vremenu, podršci i korisnim savjetima tijekom izrade ovog diplomskog rada.
Neizmjerno hvala mojim roditeljima, Stjepanu i Verici, braći, Barbari, Katarini i Marku, te suprugu Davoru, na bezuvjetnoj podršci tijekom svih ovih godina studiranja.
Hvala Vladi na vjeri u mene, i hvala mojim prijateljicama, Ivani i Mateji na podršci i razumijevanju.

Sadržaj

Sadržaj **iv**

Uvod **2**

1 Diskriminantna analiza **3**
1.1 Opća namjena i opis metode 3
1.2 Pitanja koja se pojavljuju tijekom istraživanja 4
1.3 Ograničenja linearne diskriminantne analize 6
1.4 Bazične jednadžbe za diskriminantnu analizu 9
1.5 Vrste diskriminantnih analiza 14
1.6 Kriteriji za ukupnu statističku značajnost 15
1.7 Prikaz diskriminacijskih funkcija 16
1.8 Vrednovanje prediktorskih varijabli 19
1.9 Veličina efekta 20
1.10 Provjera uspješnosti klasifikacije 21

2 Primjena linearne diskriminantne analize na primjeru daljinskih istraživanja usjeva **24**
2.1 Deskriptivna statistika 24
2.2 Provjera pretpostavki linearne diskriminantne analize 27
2.3 Udaljenost između grupa 30
2.4 Statistička značajnost 31
2.5 Diskriminacijske funkcije 33
2.6 Klasifikacija testnih podataka 38
2.7 Poboljšanje klasifikacije 40
2.8 Kanonička diskriminantna analiza 43
2.9 Zaključak 45

3 Primjena linearne diskriminantne analize u istraživanju svojstava zrna triju sorti pšenice **46**

3.1 Deskriptivna statistika 46
3.2 Provjera pretpostavki linearne diskriminantne analize 48
3.3 Udaljenost između grupa 51
3.4 Statistička značajnost 52
3.5 Rezultati klasifikacije 59
3.6 Klasifikacija testnih podataka 62
3.7 Kanonička diskriminantna analiza 63
3.8 Zaključak 67

A Korišteni SAS kod **68**

Bibliografija **77**

Uvod

Statistika je, prema definiciji, grana primijenjene matematike koja se bavi prikupljanjem, uređivanjem, analizom i tumačenjem podataka. Kao takva, pronašla je svoju primjenu u raznim područjima znanstvenih i stručnih djelatnosti; u ekonomiji, medicini, biologiji, psihologiji, demografiji, itd. U mnogim područjima primjene, od interesa je sposobnost razlikovanja dviju ili više promatranih skupina prema njihovim određenim svojstvima. Na primjer, u liječenju oboljelih osoba, isti lijek jednoj skupini pacijenata pomaže, a drugoj odmaže. Cilj je otkriti po kojim osobinama se te dvije skupine pacijenata razlikuju, kako bismo nove pacijente mogli svrstati u ispravnu skupinu i potom im dati lijek ili ne. Metode kojima pokušavamo svrstati objekte u ispravne skupine nazivaju se klasifikacijske metode, a jedna od njih je i diskriminantna analiza.

Skupine u diskriminantnoj analizi nazivamo grupe, a promatrana svojstva ili osobine nazivamo prediktorima. Glavna svrha diskriminantne analize je pokušati predvidjeti pripadnost grupi na temelju skupa prediktora, tj. pronaći funkcije prediktorskih varijabli koje maksimalno separiraju promatrane grupe objekata, a zatim pomoću njih klasificirati nove objekte u ispravne grupe. Diskriminantna analiza je tehnika koja se koristi kada je zavisna varijabla kategorijska (grupe), a nezavisne varijable su kontinuirane (prediktori). Dva su osnovna cilja: pronalaženje prediktivne jednadžbe za klasificiranje novih slučajeva u grupe, ili tumačenje prediktivne jednadžbe kako bi se bolje razumjeli odnosi koji mogu postojati među varijablama.

Izvornu dihotomnu diskriminantnu analizu razvio je Ronald Fisher 1936. godine. Iako je početno proučavanje diskriminantne analize uključivalo primjenu u biološkim i medicinskim znanostima, usredotočenim na predviđanje pripadnosti grupi, razvila se potreba za tumačenjem učinka otkrivenog analizom i primjenom analize i u područjima poslovanja, obrazovanja, inženjerstva i psihologije. Danas je primjena diskriminantne analize česta u analizi uspjeha novog proizvoda, procjeni kreditnog rizika i bankrota, i sl.

Osnovna podjela diskriminantne analize je na linearnu i kvadratnu diskriminantnu analizu. U ovom radu, koncentrirat ćemo se na linearnu diskriminantnu analizu, te pokazati njezinu primjenu kroz dva primjera. U prvom poglavlju upoznat ćemo se sa samom metodom, osnovnim pojmovima i ciljevima, te uvjetima i problemima analize. Opisat ćemo izvođenje i tumačenje diskriminacijskih funkcija i klasifikacijskih jednadžbi. U drugom

poglavlju, ilustrirat ćemo opisanu analizu na konkretnom primjeru podataka, te pokazati kako pristupiti podacima koji ne daju zadovoljavajuće rezultate. U trećem poglavlju, provest ćemo linearnu diskriminantnu analizu na novom skupu podataka, te dati malo detaljniju analizu varijabli prediktora kada je njihov doprinos klasifikaciji i separaciji značajan. Pri analizi podataka, koristit ćemo programski jezik SAS.

Poglavlje 1

Diskriminantna analiza

1.1 Opća namjena i opis metode

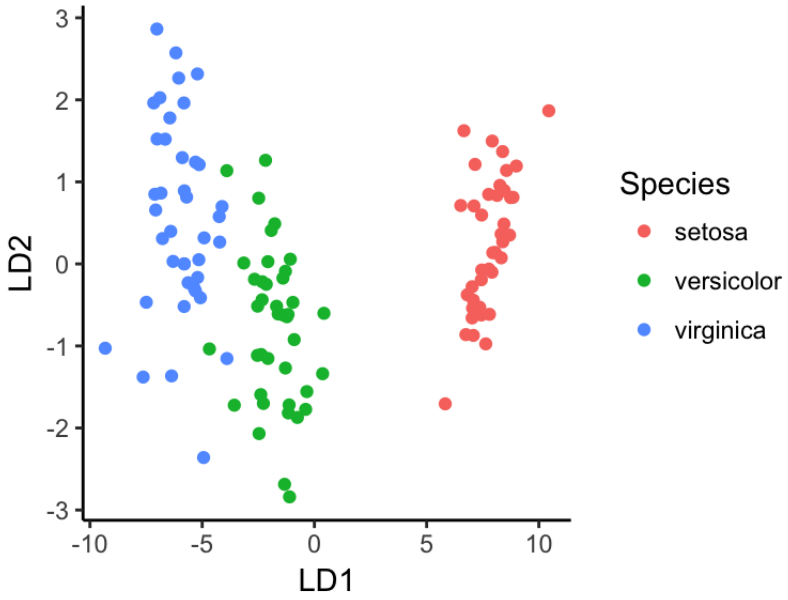
Linearna diskriminantna analiza (LDA) nastoji procijeniti linearnu kombinaciju varijabli koja najbolje diskriminira pripadnost individualnih elemenata određenoj grupi. Diskriminacija se postiže izračunom linearnog odnosa varijabli za svaki individualni element na način da se maksimizira razlika među grupama (relativni odnos varijance među grupama i unutar grupa) pri čemu je moguće, ovisno o istraživačkom okviru, procijeniti više od jedne diskriminacijske funkcije. Statistički, LDA testira nultu hipotezu jednakosti grupnih prosjeka (centroida) za skup nezavisnih varijabli gdje je mjera statističke značajnosti izračunata na osnovi generalizirane mjere udaljenosti između grupnih centroida.

Diskriminantna analiza slična je višestrukoj regresijskoj analizi, no glavna razlika jest što LDA dozvoljava kategorijsku zavisnu varijablu. Metodologija LDA slična je regresijskoj analizi, tj. svaku nezavisnu varijablu grafički uspoređujemo s grupnim varijablama, zatim biramo varijable kako bismo odredili koje su nezavisne varijable korisne te provodimo analizu reziduala kako bismo odredili točnost diskriminacijskih jednadžbi.

Matematički, diskriminantna analiza usko je povezana s jednosmjernom multivarijatnom analizom varijance (MANOVA). Semantički, međutim, dolazi do zabune između MANOVE i LDA jer su u MANOVI nezavisne varijable (kategorijske) grupe, a zavisne (kontinuirane) prediktori, dok su u LDA prediktori nezavisne (kontinuirane), a grupe zavisne varijable (kategorijske). U ovom radu, da bi izbjegli zabunu, pozivat ćemo se uvijek na nezavisne varijable kao prediktore i na zavisne varijable kao grupe ili varijable grupiranja.

Druga razlika uključuje tumačenje razlika između prediktora. U MANOVI se često nastoji odlučiti koje zavisne varijable su povezane s grupnim razlikama, ali rijetko se nastoji protumačiti obrazac razlika između zavisnih varijabli kao cjeline. U LDA se često nastoji protumačiti obrazac razlika između prediktora kao cjeline s ciljem razumijevanja dimenzija po kojima se grupe razlikuju.

S ovim pokušajem tumačenja razlika, analiza postaje složenija, jer s više od dvije grupe može postojati više načina kombiniranja prediktora za razlikovanje grupa. Zapravo, može postojati onoliko dimenzija koje diskriminiraju grupe, koliko postoji stupnjeva slobode za grupe ili onoliko koliko je prediktora (ovisno o tome što je manje).[7]



Slika 1.1: Grafički prikaz linearnih diskriminacijskih funkcija za razdvajanje triju grupa cvijeta irisa iz poznatog Fisherovog primjera
izvor: http://www.sthda.com/english/sthda-upload/figures/machine-learning-essentials/032-discriminant-analysis-ggplot-lda-1.png

1.2 Pitanja koja se pojavljuju tijekom istraživanja

Primarni cilj istraživanja koja koriste LDA jest pronaći dimenziju ili dimenzije po kojima se grupe razlikuju, te pronaći klasifikacijske funkcije za predviđanje pripadnosti grupi. Stupanj u kojem su ti ciljevi ispunjeni ovisi o izboru prediktora.

Glavno pitanje koje se postavlja u LDA jest pitanje značajnosti predikcije, tj. može li se pripadnost grupi pouzdano predvidjeti pomoću skupa prediktora. U skladu s time, zanima nas i koliko je dimenzija potrebno da bi se grupe pouzdano razlikovale, tj. koje diskriminacijske funkcije su statistički značajne, a koje nisu.

1.2.1 Broj značajnih diskriminacijskih funkcija

Veoma bitno pitanje u diskriminantnoj analizi jest koliko dimenzija je potrebno da bi se grupe pouzdano razlikovale? Koje diskriminacijske funkcije su statistički značajne, a koje nisu?

U LDA, prva diskriminacijska funkcija pruža najbolje razdvajanje grupa. Zatim, druga diskriminacijska funkcija, ortogonalna na prvu, najbolje razdvaja grupe na temelju asocijacija koje se ne koriste u prvoj diskriminacijskoj funkciji. Ovaj postupak pronalaska uzastopnih ortogonalnih diskriminacijskih funkcija nastavlja se sve dok se ne procijene sve moguće dimenzije. Kao što smo već naveli, broj mogućih dimenzija ili je jedan manji od broja grupa ili je jednak broju varijabli prediktora, ovisno o tome što je od toga manje. Obično samo prva ili prve dvije diskriminacijske funkcije pouzdano razlikuju grupe, dok preostale funkcije ne pružaju dodatne informacije o pripadnosti grupi i bolje ih je zanemariti.

Od važnosti je znati protumačiti dimenzije po kojima se grupe razdvajaju, gdje se nalaze grupe u odnosu na diskriminacijske funkcije, kako prediktori koreliraju s tim funkcijama, te koje se linearne jednadžbe mogu koristiti za klasifikaciju novih slučajeva u grupe.

1.2.2 Adekvatnost klasifikacije

S obzirom na klasifikacijske funkcije, zanima nas koliki udio slučajeva je pravilno klasificiran, te kako se pogrešno klasificiraju slučajevi. Klasifikacijske funkcije koriste se za predviđanje pripadnosti grupi za nove slučajeve i za provjeru adekvatnosti klasifikacije za slučajeve u istom uzorku krosvalidacijom. Ako istraživač zna da se neke grupe pojavljuju s većom vjerojatnošću ili ako su neke vrste pogrešne klasifikacije posebno nepoželjne, postupak klasifikacije može se modificirati.

1.2.3 Veličina efekta

Još jedno od pitanja koja se pojavljuju jest koji je stupanj odnosa između pripadnosti grupi i skupa prediktora.

Ako prva diskriminacijska funkcija odvaja jednu grupu od drugih dviju grupa, koliko se varijanca za grupe preklapa s varijancom u kombiniranim rezultatima prediktora? Ovo je u osnovi pitanje postotka varijance koji se uzima u obzir i na to pitanje odgovara se pomoću kanoničke korelacije[1].

[1]Cilj kanoničke korelacije je analizirati odnose između dva skupa varijabli, gdje jedan skup uzimamo kao nezavisnu varijablu, a drugi kao zavisnu. Kanonička korelacija pruža statističku analizu za istraživanje u kojem se svaki subjekt mjeri na dva skupa varijabli, a istraživač želi znati jesu li i kako ta dva skupa međusobno povezana.

Kanonička korelacija mjeri opseg povezanosti između diskriminantnih rezultata i grupa. To je mjera povezanosti između jedne diskriminancijske funkcije i skupa tzv. dummy varijabli koje definiraju pripadnost grupi.

1.2.4 Važnost varijabli prediktora i značajnost predikcije s kovarijatama

Pitanja o važnosti prediktora analogna su pitanjima o značaju zavisnih varijabli u MANOVI, nezavisnih varijabli u višestrukoj regresiji te zavisnih i nezavisnih varijabli u kanoničkoj korelaciji. LDA pokušava tumačiti korelaciju između prediktora i diskriminacijskih funkcija, te procijeniti prediktore prema tome koliko dobro razdvajaju svaku grupu od svih ostalih.

U LDA, kao i u MANOVI, sposobnost nekih prediktora da unaprijede razdvajanje grupa može se procijeniti nakon prilagodbe za prethodne varijable. Ako se vrijednosti jednog prediktora smatraju kovarijatom i prve „ulaze“ u LDA, zanima nas doprinose li tada predviđanju pripadnosti grupi i vrijednosti preostalih prediktora dodavanjem u jednadžbu. Drugim riječima, u smislu sekvencijalne LDA, postavlja se pitanje pružaju li preostali prediktori značajno bolju klasifikaciju između grupa od one klasifikacije koju daje samo prvi prediktor.

Odgovore na ova pitanja dati ćemo kasnije u radu.

1.3 Ograničenja linearne diskriminantne analize

Klasifikacija postavlja manje statističkih zahtjeva nego zaključivanje. Ako je klasifikacija primarni cilj analize, tada je većina sljedećih zahtjeva (osim netipičnih vrijednosti i homogenosti matrica varijance – kovarijance) stavljena po strani. Ako se, na primjer, postigne 95%-tna točnost u klasifikaciji, teško da će nas brinuti oblik distribucija. Ako je stopa klasifikacije nezadovoljavajuća, to može biti zbog kršenja pretpostavki ili ograničenja, što može također i iskriviti testove statističke značajnosti.

1.3.1 Nejednake veličine uzoraka, podaci koji nedostaju i snaga

Kako je LDA tipično jednosmjerna analiza, nejednake veličine uzoraka u grupama ne predstavljaju posebne probleme.[2] U klasifikaciji je, međutim, potrebna odluka o tome želimo li da veličina uzorka utječe na apriorne vjerojatnosti dodjele grupama. Odnosno, želimo li da

[2]Problem se zapravo pojavljuje ako je poželjna rotacija zbog neortogonalnosti diskriminacijskih funkcija s nejednakim veličinama , ali rotacija osi nije uobičajena u diskriminantnoj analizi.

vjerojatnost s kojom je slučaj dodijeljen grupi odražava činjenicu da je sama grupa više (ili manje) vjerojatna u uzorku? Veličina uzorka najmanje grupe trebala bi biti veća od broja varijabli prediktora. Iako sekvencijalna i stepwise LDA izbjegavaju probleme multikolinearnosti i singularnosti testom tolerancije u svakom koraku, tzv. *overfitting* se javlja kod svih oblika LDA ako veličina najmanje grupe nije značajno veća od broja prediktora.[3]

Postoji nekoliko rješenja za problem podataka koji nedostaju, od kojih nijedno nije idealno te je važno donijeti ispravnu odluku o tome koju metodu upotrijebiti. Neke od metoda su: brisanje slučajeva ili varijabli koje nedostaju, procjena podataka koji nedostaju, upotreba korelacijske matrice podataka koji nedostaju, tretiranje podataka koji nedostaju kao podataka pomoću tzv. *dummy* varijabli ili ponavljanje analize sa i bez podataka koji nedostaju.

Snaga analize je smanjena, osim ako u svakoj grupi nema više slučajeva nego što je prediktora, zbog smanjenih stupnjeva slobode za pogreške. Jedan vjerojatni ishod smanjene snage nije značajan multivarijatni F omjer, već jedan ili više značajnih univarijatnih F omjera. Veličine uzoraka u svakoj grupi moraju u svakom slučaju biti dovoljne da osiguraju odgovarajuću snagu. Dostupni su mnogi programski paketi za izračunavanje potrebnih veličina uzorka ovisno o željenoj snazi i očekivanim srednjim vrijednostima i standardnim devijacijama u ANOVI. Internetska pretraga za "statističkom snagom" otkriva nekoliko njih, od kojih su neki besplatni. Jedan brz i „prljav“ način da ih primijenimo jest da odaberemo prediktor s najmanjom očekivanom razlikom koju želimo pokazati statističkom značajnošću – minimalno značajan prediktor.

Snaga za multivarijatni test je najveća kada je združena korelacija unutar grupe između dva prediktora visoka i negativna. Multivarijatni test ima mnogo manju snagu kada je korelacija pozitivna, nula ili umjereno negativna. Međutim, zanimljiva stvar se događa kada je jedan od dvaju prediktora pod utjecajem tretmana, a drugi ne. Što je veća apsolutna vrijednost korelacije između dva prediktora, to je veća snaga multivarijatnog testa.[8]

1.3.2 Multivarijatna normalnost

Kada se koristi statističko zaključivanje u LDA, pretpostavka multivarijatne normalnosti je da su rezultati prediktora neovisno i nasumično odabrani iz populacije, te da je distribucija uzorka bilo koje linearne kombinacije prediktora normalna.

Međutim, LDA je otporna na propuste normalnosti ako je kršenje uzrokovano zakrivljenošću (engl.*skewness*), a ne netipičnim vrijednostima (enlg.*outlier*). Poznato je da bi veličina uzorka koja bi proizvela 20 stupnjeva slobode za pogrešku u slučaju univarijatne ANOVE, trebala osigurati robusnost u odnosu na multivarijatnu normalnost sve dok su veličine uzoraka jednake i dok se koriste dvostrani testovi.[4]

[3]Također, vrlo nejednake veličine uzoraka bolje se rješavaju logističkom regresijom nego diskriminantnom analizom.

Budući da su testovi za LDA obično dvostrani, ovaj zahtjev ne predstavlja poteškoću. Međutim, veličine uzoraka često nisu jednake za primjenu LDA jer se prirodno grupe rijetko pojavljuju ili uzorkuju s jednakim brojem slučajeva u grupama. Kako se razlike u veličini uzorka među grupama povećavaju, potrebne su sve veće ukupne veličine uzoraka kako bi se osigurala robusnost. Kao konzervativna preporuka, robusnost se očekuje s 20 slučajeva u najmanjoj grupi, ako postoji samo nekoliko prediktora (recimo, pet ili manje).

Ako su uzorci mali i nejednake veličine, procjena normalnosti je stvar prosuđivanja. Ako se ne očekuje normalna uzoračka distribucija prediktora u uzorkovanoj populaciji, mogla bi se isplatiti transformacija jednog ili više prediktora.

1.3.3 Homogenost matrica varijance i kovarijance

Kada su veličine uzoraka jednake ili velike, LDA je robusna do kršenja pretpostavke o jednakosti matrica varijance i kovarijance (disperzije) unutar grupe. Međutim, kada su veličine uzoraka nejednake ili male, rezultati testa značajnosti mogu navesti na krivi trag ako postoji heterogenost matrica varijance i kovarijance.

Iako je zaključivanje obično robusno s obzirom na heterogenost matrica varijance i kovarijance kod uzoraka pristojne veličine, klasifikacija nije. Može doći do prekvalifikacije slučajeva u grupe s većom disperzijom. Ako je klasifikacija važan cilj analize, dobro je provjeriti homogenost matrica varijance i kovarijance.

Andersonov test, dostupan u SAS DISCRIM proceduri naredbom *POOL=TEST*, procjenjuje homogenost matrica varijance i kovarijance, ali je osjetljiv na normalnost, odnosno nenormalnost podataka.

Ako se ustanovi heterogenost, mogu se transformirati prediktori ili koristiti zasebne matrice kovarijance tijekom klasifikacije, kvadratna diskriminantna analiza ili neparametarska klasifikacija. Ako se odlučimo za transformaciju prediktora, važno je provjeriti jesu li transformirani podaci normalno distribuirani ili barem blizu normalne distribucije; ako ne, isprobavamo transformacije dok ne dođemo do one koja najbliža normalnoj distribuciji. Klasifikacija koja koristi zasebne matrice često dovodi do tzv.*overfitting*-a pa bi se trebala koristiti samo ako je uzorak dovoljno velik da dopušta krosvalidaciju. Kvadratna diskriminantna analiza izbjegava preklasifikaciju u grupe s većom disperzijom, ali se loše ponaša s malim uzorcima.

Dakle, transformaciju varijabli koristimo ako postoji značajno odstupanje od homogenosti ili ako su uzorci mali i nejednaki. Ako je naglasak na klasifikaciji i varijance nisu jednake, koristimo zasebne matrice kovarijance i/ili kvadratnu diskriminantnu analizu ako su uzorci veliki, a varijable normalne te neparametarske metode klasifikacije ako su varijable nenormalne i/ili su uzorci mali.

1.3.4 Linearnost

LDA model pretpostavlja linearne odnose između svih parova prediktora unutar svake grupe. Pretpostavka je manje stroga od drugih, međutim, kod njezinog kršenja dolazi do smanjenja snage, a ne do povećanja pogreške tipa I.

Ako je uočena nelinearnost, ona se može ispraviti transformacijom nekih kovarijata. Ili, zbog poteškoća u tumačenju transformiranih varijabli, može se i eliminirati kovarijata koja uzrokuje nelinearnost.

1.3.5 Odsustvo multikolinearnosti i singularnosti

Multikolinearnost ili singularnost mogu se pojaviti s izrazito redundantnim prediktorima, što inverziju matrice čini nepouzdanom. Srećom, većina računalnih programa štiti od multikolinearnosti i singularnosti računanjem SMC-a (engl. *squared multiple correlations*) za varijable. SMC je kvadrat višestruke korelacije svake varijable sa svim ostalim varijablama. Ako je SMC visok, varijabla je jako povezana s ostalima u skupu i imamo multikolinearnost. Ako je SMC jednak 1, varijabla je savršeno povezana s ostalima u skupu i imamo singularnost. Mnogi programi pretvaraju SMC vrijednosti za svaku varijablu u toleranciju (1 - SMC) i bave se tolerancijom umjesto SMC. Prediktori s nedovoljnom tolerancijom se isključuju.

1.4 Bazične jednadžbe za diskriminantnu analizu

Bazične jednadžbe prikazat ćemo za dva glavna dijela LDA: diskriminacijske funkcije i klasifikacijske jednadžbe.

1.4.1 Izvod i testiranje diskriminacijskih funkcija

Varijanca u skupu prediktora podijeljena je u dva izvora: varijancu koja se pripisuje razlikama između grupa i varijancu koja se pripisuje razlikama unutar grupa, što daje sljedeću jednadžbu.

$$\sum_i \sum_j (Y_{ij} - \overline{Y})^2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_j + \overline{Y}_j - \overline{Y})^2 =$$

$$= n \sum_j (\overline{Y}_j - \overline{Y})^2 + \sum_i \sum_j (Y_{ij} - \overline{Y}_j)^2 \quad , \quad i = 1, 2, ...n \ , j = 1, 2, ..., broj \ \ grupa. \qquad (1.1)$$

Ukupni zbroj kvadrata razlika rezultata Y (zavisna varijabla) i srednje vrijednosti svih podataka $\overline{Y}$ podijeljen je na zbroj kvadrata razlika srednjih vrijednosti grupa Y_j i srednje vrijednosti svih podataka (tj.varijabilnost između grupa), i zbroj kvadrata razlika pojedinačnih rezultata Y_{ij} i njihovih odgovarajućih grupnih srednjih vrijednosti $\overline{Y}_j$.

Ili, jednadžbu (1.1) možemo napisati u obliku

$$SS_{ukupno} = SS_{između} + SS_{unutar}. \tag{1.2}$$

Ukupna matrica unakrsnih produkata, SS_{ukupno}, podijeljena je u matricu unakrsnih produkata povezanu s razlikama između grupa, $SS_{između}$, i matricu unakrsnih produkata razlika unutar grupa SS_{unutar}.

Nakon što dobijemo matrice $SS_{između}$ i SS_{unutar}, potrebno je izračunati sljedeće determinante[4]

$$|SS_{unutar}|$$

i

$$|SS_{između} + SS_{unutar}|,$$

te zatim i Wilksovu lambdu[5]

$$\Lambda = \frac{|SS_{unutar}|}{|S_{između} + S_{unutar}|}.$$

Da bismo našli aproksimativni F omjer (omjer objašnjene i neobjašnjene varijabilnosti) koristimo formulu

$$Aproksimativni \;\; F(df_1, df_2) = \left(\frac{1-y}{y}\right)\left(\frac{df_2}{df_1}\right),$$

gdje je

$$y = \Lambda^{1/s},$$

$$s = min(p, df_{između}),$$

[4]Determinanta se može promatrati kao mjera generalizirane varijance matrice.

[5]Objašnjenje u poglavlju 1.6. U ovim jednadžbama koristimo *između* i *unutar*, redom, umjesto $effect$ i *error*.

$$df_1 = p \cdot df_{između},$$

$$df_2 = s \cdot \left[df_{unutar} - \frac{p - df_{između} + 1}{2} \right] - \left[\frac{p \cdot df_{između} - 2}{2} \right]$$

i

$$df = broj \;\; stupnjeva \;\; slobode.$$

Koristimo podatke:

- p - broj varijabli prediktora
- $df_{između}$ - broj grupa minus 1 $(k - 1)$
- df_{unutar} - broj grupa pomnožen s $(n - 1)$, gdje je n broj slučajeva po grupi.Budući da n često nije jednak za sve grupe, alternativa za df_{unutar} je $N - k$, gdje je N ukupan broj slučajeva u svim grupama.

Ovo je test ukupnog odnosa između grupa i prediktora. Sljedeći je korak ispitati diskriminacijske funkcije koje čine cjelokupni odnos.

Maksimalni broj diskriminacijskih funkcija je, kao što smo već rekli, ili broj prediktora ili broj stupnjeva slobode za grupe, ovisno o tome što je manje. Na primjer, s tri grupe i četiri prediktora, potencijalno postoje dvije diskriminacijske funkcije koje doprinose ukupnom odnosu (jer $df = 3 - 1 = 2 < 4$). Ako je ukupni odnos statistički značajan, vrlo je vjerojatno da će barem prva diskriminacijska funkcija biti značajna, a mogu biti i obje.

Diskriminacijske funkcije su poput regresijskih jednadžbi; rezultat diskriminacijske funkcije za pojedini slučaj predviđa se iz zbroja niza prediktora, svaki ponderiran koeficijentom. Postoji jedan skup koeficijenata diskriminacijske funkcije za prvu diskriminacijsku funkciju, drugi skup koeficijenata za drugu diskriminacijsku funkciju itd. Slučajevi dobivaju zasebne rezultate diskriminacijske funkcije za svaku diskriminacijsku funkciju kada se u jednadžbe umetnu njihovi rezultati prediktora.

Za dobivanje (standardiziranih) rezultata diskriminacijske funkcije za i-tu funkciju koristi se jednadžba (1.3).

$$D_{zi} = d_{i1}z_1 + d_{i2}z2 + \cdots + d_{ip}z_p \tag{1.3}$$

Standardizirani rezultat za i-tu diskriminacijsku funkciju (D_i) nalazi se množenjem standardizirane vrijednosti svakog prediktora (z) s njegovim standardiziranim koeficijentom diskriminacijske funkcije (d_i), a zatim se dodaju produkti za sve prediktore.

U LDA, d_i i su odabrani tako da povećaju razlike između grupa u odnosu na razlike unutar grupa.

Baš kao i u višestrukoj regresiji, jednadžba (1.3) može se napisati ili za originalne rezultate ili za standardizirane rezultate. Procjena diskriminacijske funkcije za zasebni slučaj se tada može dobiti množenjem originalnog rezultata svakog prediktora s pripadajućim nestandardnim koeficijentom diskriminacijske funkcije, dodavanjem produkta na sve prediktore i dodavanjem konstante za prilagođavanje srednjih vrijednosti. Rezultat dobiven na ovaj način jednak je kao i D_i dobiven u jednadžbi (1.3). Srednja vrijednost svake diskriminacijske funkcije u svim slučajevima je nula, jer je sredina svakog prediktora, kada je standardiziran, nula. Standardna devijacija svakog D_i jednaka je 1.

Baš kao što se D_i može izračunati za svaki slučaj, može se izračunati i njegova srednja vrijednost za svaku grupu. Članovi svake grupe, promatrani zajedno, imaju prosječan rezultat diskriminacijske funkcije koji je udaljenost grupe, u jedinicama standardne devijacije, od nulte sredine diskriminacijske funkcije. Grupne D_i sredine obično se zovu centroidi u reduciranom prostoru, pri čemu je prostor smanjen s onog sa p prediktora na jednu dimenziju, ili diskriminacijsku funkciju.

Za svaku diskriminacijsku funkciju pronađena je kanonička korelacija. Kanoničke korelacije pronalaze se rješavanjem svojstvenih vrijednosti i svojstvenih vektora korelacijske matrice. Svojstvena vrijednost oblik je kvadratne kanoničke korelacije koja, kao što je uobičajeno za kvadratne koeficijente korelacije, predstavlja preklapanje varijance među varijablama, u ovom slučaju između prediktora i grupa. Uzastopne diskriminacijske funkcije procjenjuju se zbog značajnosti.

Ako postoje samo dvije grupe, rezultati diskriminacijske funkcije mogu se koristiti za klasifikaciju slučajeva u grupe. Slučaj se klasificira u jednu grupu ako je rezultat D_i iznad nule, a u drugu grupu ako je rezultat D_i ispod nule. S velikim brojem grupa, moguće je klasificiranje prema diskriminacijskim funkcijama, ali je jednostavnije koristiti postupak klasifikacije opisan u sljedećem poglavlju.

1.4.2 Klasifikacija

Kako bi se slučajevi klasificirali u grupe, za svaku grupu razvijena je klasifikacijska jednadžba. Podaci za svaki slučaj ubacuju se u svaku klasifikacijsku jednadžbu kako bi se dobio rezultat klasifikacije slučaja za svaku grupu. Slučaj se dodjeljuje grupi za koju ima najveći rezultat klasifikacije.

U svom najjednostavnijem obliku, osnovna klasifikacijska jednadžba za j-tu grupu ($j = 1, 2, ..., k$) je

$$C_j = c_{j0} + c_{j1}X_1 + c_{j2}X_2 + \cdots + c_{jp}X_p \tag{1.4}$$

Rezultat klasifikacijske funkcije za grupu j (C_j) dobiva se množenjem originalnog rezultata svakog prediktora (X) s pripadajućim koeficijentom klasifikacijske funkcije (c_j), zbrajanjem svih prediktora i dodavanjem konstante c_{j0}.

Klasifikacijski koeficijenti, c_j nalaze se pomoću aritmetičkih sredina p prediktora i združene matrice varijance i kovarijance unutar skupine, $\mathbf{W}$. Matrica kovarijance unutar grupe dobivena je dijeljenjem svakog elementa u matrici unakrsnih produkata, SS_{unutar}, s brojem stupnjeva slobode unutar grupe $N - k$. U matričnom obliku,

$$\mathbf{C}_j = \mathbf{W}^{-1}\mathbf{M}_j \, . \tag{1.5}$$

Stupčana matrica klasifikacijskih koeficijenata za skupinu j ($\mathbf{C}_j = c_{j1}, c_{j2}, ..., c_{jp}$) nalazi se množenjem inverza matrice varijance i kovarijance unutar grupe, $\mathbf{W}^{-1}$, stupčanom matricom sredina za grupu j na p varijabli ($\mathbf{M}_j = X_{j1}, X_{j2}, ..., X_{jp}$).

Konstanta za skupinu j, c_{j0}, nalazi se pomoću jednadžbe 1.6.

$$c_{j0} = \left(-\frac{1}{2}\right)\mathbf{C}'_j\mathbf{M}_j \, . \tag{1.6}$$

Konstanta za klasifikacijsku funkciju grupe j (c_{j0}) dobiva se množenjem $\left(-\frac{1}{2}\right)$ transponiranom stupčanom matricom klasifikacijskih koeficijenata za grupu j ($\mathbf{C}'_j$) i stupčanom matricom srednjih vrijednosti za grupu j ($\mathbf{M}_j$).

Ova jednostavna klasifikacijska shema najprikladnija je kada se u populaciji očekuju jednake veličine grupa. Ako se očekuju nejednake veličine grupa, postupak klasifikacije može se izmijeniti postavljanjem apriornih vjerojatnosti na veličinu grupe. Klasifikacijska jednadžba za skupinu j (C_j) tada postaje

$$C_j = c_{j0} + \sum_{i=1}^{p} c_{ji}X_i + ln(n_j/N) \, , \tag{1.7}$$

gdje je n_j veličina grupe j, a N ukupna veličina uzorka.

Valja ponovno naglasiti da su klasifikacijski postupci vrlo osjetljivi na heterogenost matrica varijance i kovarijance. Vjerojatnije je da će slučajevi biti klasificirani u grupu s najvećom disperzijom, odnosno u grupu za koju je determinanta matrice kovarijance unutar grupe najveća.

1.5 Vrste diskriminantnih analiza

Postoje tri vrste linearnih diskriminantnih analiza – direktne (standardne), sekvencijalne i statističke (engl. *stepwise*).

U nastavku poglavlja, navest ćemo glavne karakteristike za svaku od spomenutih vrsta analize.

1.5.1 Direktna diskriminantna analiza

U direktnoj (standardnoj) LDA, svi prediktori ulaze u jednadžbe odjednom i svakom prediktoru se dodjeljuje samo jedinstvena povezanost koju ima s grupama. Zajednička varijanca među prediktorima doprinosi ukupnom odnosu, ali ne niti jednom prediktoru.

U ovom cjelokupnom testu odnosa između grupa i prediktora, sve diskriminacijske funkcije su kombinirane i sve zavisne varijable razmatraju se istovremeno. Opisani postupak izvoda bazičnih jednadžbi odgovara upravo direktnoj diskriminantnoj analizi.

1.5.2 Sekvencijalna diskriminantna analiza

Sekvencijalna (ili, drugim nazivom, hijerarhijska) LDA koristi se za procjenu doprinosa prediktora predviđanju pripadnosti grupi kako ulaze u jednadžbe, redoslijedom koji odredi istraživač. Istraživač procjenjuje poboljšanje klasifikacije kada se skupu prethodnih prediktora doda novi prediktor. Glavno pitanje na koje želimo otkriti odgovor jest poboljšava li se pouzdano klasifikacija slučajeva u grupe dodavanjem novog prediktora ili novih prediktora?

Ako se prediktori koji uđu ranije od ostalih promatraju kao kovarijate, a dodani prediktor kao zavisne varijable, LDA se može koristiti za analizu kovarijance.

Sekvencijalna LDA je također korisna kada postoji neka osnova za uspostavljanje prioritetnog reda među prediktorima. Ako je, na primjer, neke prediktore lako ili „jeftino“ dobiti i ako im se daje raniji ulaz, koristan i isplativ skup prediktora može se pronaći kroz sekvencijalni postupak.

SAS DISCRIM procedure u SAS-u ne pružaju prikladne metode za unos prediktora prema redoslijedu prioriteta. Umjesto toga, slijed se postavlja pokretanjem zasebne diskriminantne analize za svaki korak, prvi s varijablom najvišeg prioriteta, drugi s dvije varijable najvišeg prioriteta koje ulaze istovremeno, i tako dalje. U svakom koraku može se dodati jedna ili više varijabli . Međutim, test značajnosti poboljšanja predviđanja je zamoran u odsudstvu vrlo velikih uzoraka. Ako imamo samo dvije grupe i veličine uzorka su približno jednake, bolja opcija bila bi izvođenje sekvencijalne diskriminantne analize putem interaktivne SAS REGRESS procedure, gdje je zavisna varijabla dihotomna varijabla koja predstavlja pripadnost grupi, s grupama kodiranim 0 i 1. Ako je klasifikacija poželjna,

preliminarna višestruka regresijska analiza s potpuno fleksibilnim ulazom prediktora mogla bi biti praćena diskriminantnom analizom kako bi se osigurala klasifikacija.

1.5.3 Stepwise (statistička) diskriminantna analiza

Kada istraživač nema razloga nekim prediktorima dodijeliti veći prioritet od drugih, mogu se koristiti statistički krireriji za određivanje redoslijeda ulaska prediktora. Odnosno, ako istraživač želi reducirati skup prediktora, ali nema preferencija među njima, stepwise LDA se može koristiti da bi dobili željeni reducirani skup. Ulaz prediktora određen je statističkim kriterijima koje je odredio korisnik.

Stepwise LDA ima iste kontroverzne aspekte kao i stepwise procedure općenito. Redoslijed ulaska može ovisiti o trivijalnim razlikama u odnosima među prediktorima u uzorku koji ne odražavaju razlike u populaciji. Međutim, ova pristranost se smanjuje ako se koristi krosvalidacija. Preporučuje se vjerojatnost za unošenje liberalnijeg kriterija od 0.05. Bolji izbor bio bi u rasponu od 0.15 do 0.20 kako bi se osigurao ulaz važnih varijabli. [2]

U SAS-u se stepwise diskriminantna analiza pruža putem zasebne procedure — STEPDISC. Dostupne su tri metode unosa, kao i dodatni statistički kriteriji za dvije od njih.

1.6 Kriteriji za ukupnu statističku značajnost

Neki od kriterija za procjenu ukupne statističke značajnosti u LDA su: Wilksova lambda (engl. *Wilks' lambda*), Royev najveći karakteristični korijen (engl. *Roy's greatest characteristic root*), Hotellingov trag (enlg. *Hotelling trace*) i Pillaijev kriterij (engl. *Pillai's trace*). U SAS DISCRIM proceduri dostupna su sva četiri navedena kriterija.

Dva dodatna statistička kriterija, Mahalanobisov D^2 i Raov V, posebno su relevantna za stepwise LDA. Mahalanobisov D^2 temelji se na udaljenosti između parova grupnih centroida koja se onda generalizira na udaljenosti preko više parova grupa. Raov V je još jedna generalizirana mjera udaljenosti koja svoju najveću vrijednost postiže kada postoji najveća sveukupna separacija među grupama.

Ova dva kriterija dostupna su i za usmjeravanje napredovanja stepwise diskriminantne analize i za procjenu pouzdanosti skupa prediktora za predviđanje pripadnosti grupi. Slično Wilksovoj lambdi, Mahalanobisov D^2 i Raov V temelje se na svim diskriminantnim funkcijama, a ne na jednoj. Važno je imati na umu da su lambda, D^2 i V deskriptivne statistike; one same po sebi nisu inferencijalne statistike, iako se na njih primjenjuje inferencijalna statistika.

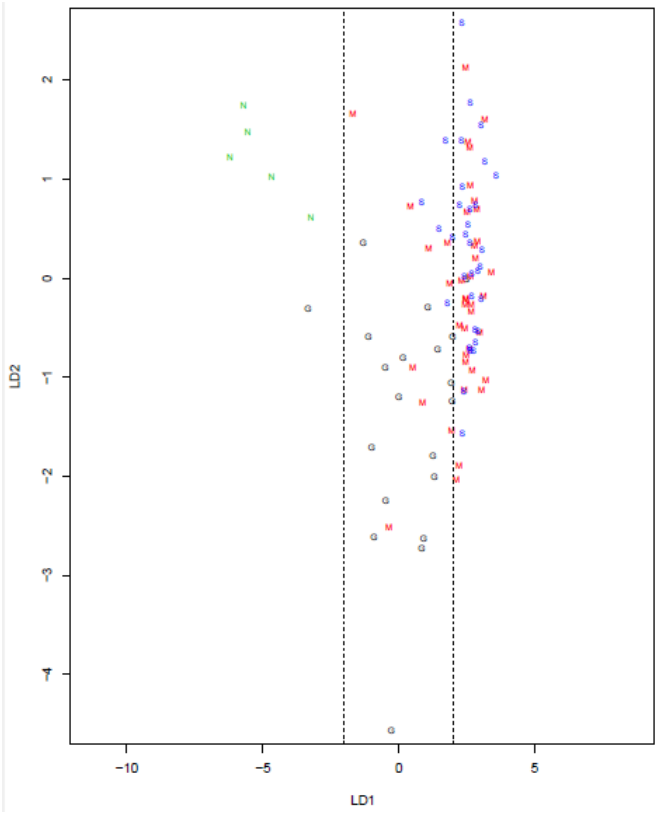
Wilksova lambda, definirana u jednadžbi (1.8) je statistika omjera vjerojatnosti podataka pod pretpostavkom jednakih vektora srednjih vrijednosti populacije za sve grupe, u odnosu na vjerojatnost pod pretpostavkom da su vektori srednjih vrijednosti populacije identični vektorima srednjih vrijednosti uzorka za različite grupe. Wilksova lambda je združeni omjer varijance pogreške i zbroja varijance efekta i varijance pogreške.

$$\Lambda = \frac{|S_{error}|}{|S_{effect} + S_{error}|} \tag{1.8}$$

Hotellingov trag je združeni omjer varijance efekta i varijance pogreške, a Pillaijev kriterij su jednostavno združene varijance efekta. Wilksova lambda, Hotellingov trag i Royev gcr kriterij često su moćniji od Pillaijevog kriterija kada postoji više od jedne dimenzije, ali prva dimenzija osigurava većinu separacije među grupama; oni imaju manju moć kada je separacija grupa raspoređena po dimenzijama. Pillaijev kriterij robusniji je od ostala tri. Kako se veličina uzorka smanjuje, pojavljuju se nejednaki *n*-ovi i krši se pretpostavka homogenosti matrica varijance-kovarijance te tu dolazi do izražaja prednost Pillaijevog kriterija u smislu robusnosti. Kada dizajn istraživanja nije idealan, tada je kriterij izbora upravo Pillaijev kriterij.[5]

1.7 Prikaz diskriminacijskih funkcija

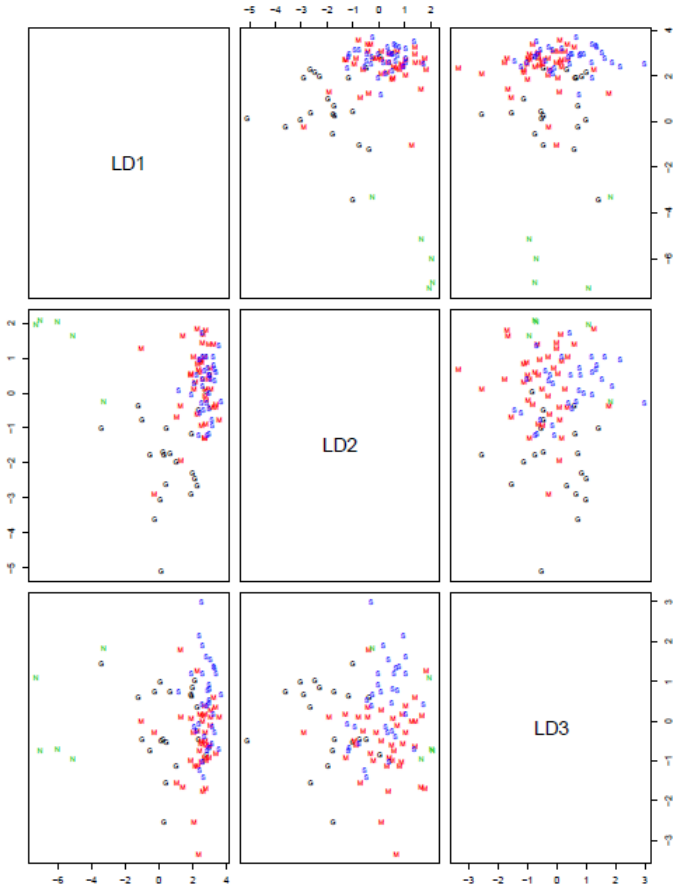
Grupe su raspoređene duž različitih diskriminacijskih funkcija ovisno o njihovim centroidima. Diskriminacijske funkcije tvore osi, a grupni centroidi ucrtani su duž tih osi. Ako postoji velika razlika između centroida jedne grupe i centroida druge grupe duž osi diskriminacijske funkcije, tada diskriminacijska funkcija razdvaja te dvije grupe. Ako ne postoji velika udaljenost između grupnih centroida, diskriminacijska funkcija ne razdvaja grupe.



Slika 1.2: Grafički prikaz dviju diskriminacijskih funkcija za razdvajanje četiriju grupa
izvor: https://online.stat.psu.edu/stat555/node/101/

Na slici 1.2 vidimo da prva diskriminacijska funkcija, *LD1*, dobro razdvaja grupe ”N” i ”G”, dok je orisnost druge diskriminacijske funkcije, *LD2*, manje jasna.

Ako postoje četiri ili više grupa i prema tome, više od dvije statistički značajne diskriminacijske funkcije, tada se koriste dijagrami osi u paru. Jedna diskriminacijska funkcija je os X, a druga je os Y. Svaka grupa ima centroid za svaku diskriminacijsku funkciju; upareni centroidi su ucrtani u odnosu na njihove vrijednosti na osi X i Y. Budući da su centroidi prikazani samo u paru, tri značajne diskriminacijske funkcije zahtijevaju tri dijagrama (funkcija 1 naspram funkcije 2; funkcija 1 naspram funkcije 3; i funkcija 2 naspram funkcije 3) i tako dalje.



Slika 1.3: Grafički prikaz parova linearnih diskriminacijskih funkcija za primjer s 4 grupe i 3 diskriminacijske funkcije
izvor: https://online.stat.psu.edu/stat555/node/101/

Na slici 1.3 vidimo prikaz dijagrama triju linearnih diskriminacijskih funkcija u paru, za 4 grupe. U ovom primjeru, duž osi nisu ucrtani samo centroidi, već svi podaci za svaku grupu.

1.7.1 Matrica strukture

Grafički prikazi centroida govore nam kako su grupe odvojene diskriminacijskom funkcijom, ali ne otkrivaju značenje diskriminacijske funkcije. Postoje razne matrice koje otkrivaju prirodu kombinacije prediktora koji razdvajaju grupe. Matrice standardiziranih diskri-

minacijskih (kanoničkih) funkcija su u osnovi koeficijenti regresije, koje bismo primijenili na rezultat svakog slučaja kako bismo pronašli standardizirani diskriminantni rezultat za taj slučaj (jednadžba 1.3).

Matrica strukture (poznata i kao matrica opterećenja, engl. *loadings matrix*) sadrži korelacije između prediktora i diskriminacijskih funkcija. Značenje funkcije istraživač zaključuje iz ovog obrasca korelacija (opterećenja). Korelacije između prediktora i funkcija nazivaju se opterećenja u analizi diskriminacijskih funkcija. Ako prediktori X_1, X_2 i X_3 jako koreliraju s funkcijom, ali prediktori X_4 i X_5 ne, istraživač pokušava razumjeti što X_1, X_2 i X_3 imaju međusobno zajedničko, a što se razlikuje od X_4 i X_5; značenje funkcije određeno je ovim shvaćanjem.

Matematički gledano, matrica opterećenja je združena matrica korelacije unutar grupe pomnožena s matricom standardiziranih koeficijenata diskriminacijske funkcije.

$$\mathbf{A} = \mathbf{R}_{unutar} \cdot \mathbf{D} \tag{1.9}$$

Matrica strukture korelacija između prediktora i diskriminacijskih funkcija, **A**, dobiva se množenjem matrice korelacija unutar grupe među prediktorima, $\mathbf{R}_{unutar}$, s matricom standardiziranih koeficijenata diskriminacijske funkcije, **D** (standardizirani korištenjem združenih standardnih devijacija unutar grupe).

Nedostaje konsenzus o tome koliko visoke moraju biti korelacije u matrici strukture da bi ih mogli interpretirati. Prema dogovoru, korelacije veće od 0.33 (10% varijance) mogu se smatrati prihvatljivima, dok one niže ne.[1] Međutim, veličina opterećenja ovisi i o vrijednosti korelacije u populaciji i o homogenosti rezultata u uzorku preuzetom iz nje. Ako je uzorak neobično homogen u odnosu na prediktor, opterećenja za prediktor su niža i možda bi bilo mudro sniziti kriterij za određivanje hoće li se prediktor tumačiti kao dio diskriminacijske funkcije.

Međutim, uvijek je potreban oprez u tumačenju opterećenja, jer su to pune, a ne djelomične (enlg. *partial*) ili poludjelomične (engl. *semipartial*) korelacije. Opterećenje bi moglo biti znatno niže kada bi se djelomično uklonile korelacije s drugim prediktorima.

U nekim slučajevima, rotacija matrice strukture može olakšati interpretaciju. No rotacija matrica diskriminantne strukture smatra se problematičnom i stoga se ne preporučuje.

1.8 Vrednovanje prediktorskih varijabli

Drugi dostupan alat za procjenu doprinosa prediktora dostupan je u SAS-u i u njemu su srednje vrijednosti za prediktore za svaku grupu suprotstavljene združenim srednjim vrijednostima za druge grupe. Na primjer, ako postoje tri grupe, srednje vrijednosti na pre-

diktorima za Grupu 1 suprotstavljaju se združenim srednjim vrijednostima iz Grupa 2 i 3. Zatim se srednje vrijednosti za Grupu 2 suprotstavljaju združenim srednjim vrijednostima iz Grupa 1 i 3, te se konačno, srednje vrijednosti za Grupu 3 suprotstavljaju združenim srednjim vrijednostima iz Grupa 1 i 2. Ovaj postupak se koristi za određivanje koji su prediktori važni za izoliranje jedne grupe od ostalih.

U slučaju da imamo 3 grupe i 4 prediktora potrebno je dvanaest GLM (enlg. *general linear models*) procedura, četiri za svaku od triju usporedba. Unutar svake usporedbe, koja izolira srednje vrijednosti iz svake grupe i suprotstavlja ih srednjim vrijednostima za druge grupe, postoje odvojene procedure za svaki od četiriju prediktora, u kojima se svaki prediktor prilagođava za preostale prediktore. U tim procedurama, prediktor od interesa je označen kao zavisna varijabla, a preostali prediktori su označeni kao kovarijate. Kao rezultat dobije se niz testova značajnosti svakog prediktora nakon prilagodbe za sve ostale prediktore u razdvajanju svake grupe od preostalih grupa.

Najbolje je uzeti u obzir samo prediktore sa "značajnim" F omjerima nakon podešavanja pogreške za broj prediktora u skupu. Čak i uz ovu prilagodbu, postoji opasnost od inflacije stope pogreške tipa I jer se izvodi više neortogonalnih usporedbi. Ako postoji velik broj grupa, može se razmotriti daljnja prilagodba kao što je množenje kritičnog F sa (k–1), gdje je k broj grupa.

Postupci detaljno opisani u ovom poglavlju najkorisniji su kada je broj grupa mali i kada su podjele među grupama prilično ujednačene na dijagramu diskriminacijskih funkcija za prve dvije funkcije. Druge vrste dijagrama diskriminacijskih funkcija mogu se predložiti ako postoji velik broj grupa, od kojih su neke blisko grupirane (npr. Grupe 1 i 2 mogu se združiti i suprotstaviti združenim Grupama 3, 4 i 5).

Ako postoji logična osnova za dodjeljivanje prioriteta prediktorima, može se koristiti sekvencijalni, a ne standardni pristup usporedbama. Umjesto procjene svakog prediktora nakon prilagodbe za sve ostale prediktore, procjenjuju ga samo prediktori višeg prioriteta nakon prilagodbe.

1.9 Veličina efekta

U diskriminantnoj analizi, tri tipa veličine efekta su od interesa: jedan tip koji opisuje varijancu povezanu s cjelokupnom analizom i dva tipa koja opisuju varijance povezane s pojedinačnim prediktorima. η^2 (ili parcijalni η^2), koji se može pronaći iz Wilksove lambde ili iz povezanih F i stupnjeva slobode, daje veličinu efekta za cijelu analizu.

Izračunavanje parcijalnog η^2 pomoću Λ

$$parcijalni\ \eta^2 = 1\text{–}\Lambda^{1/3}.$$

Zasebne veličine efekta za svaku diskriminacijsku funkciju dostupne su kao kvadrirane kanoničke korelacije (engl. *squared canonical correlation*).

Matrica strukture daje opterećenja u obliku korelacije svakog prediktora sa svakom diskriminacijskom funkcijom. Ove korelacije, kada ih kvadriramo, zapravo su veličine efekta, što ukazuje na udio varijance podijeljene između svakog prediktora i svake funkcije. Matrica strukture u SAS DISCRIM proceduri označena je kao *Pooled Within Canonical Structure*.

Drugi oblik veličine efekta je η^2 koji se može pronaći kada se izvode usporedbe između svake grupe i preostalih grupa, pri čemu je svaki prediktor prilagođen za sve ostale prediktore.

1.10 Provjera uspješnosti klasifikacije

Osnovna tehnika za klasificiranje slučajeva u grupe navedena je u poglavlju 1.4. Rezultati klasifikacije prikazani su SAS tablicama kao što su *Classification results* ili *Number of Observations and Percents Classified into GROUP*, gdje se stvarna pripadnost grupi uspoređuje s predviđenom pripadnosti grupi. Iz ovih se tablica iščitava postotak ispravno klasificiranih slučajeva te broj i priroda grešaka u klasifikaciji.

Kada postoji jednak broj slučajeva u svakoj grupi, lako je odrediti postotak slučajeva koji bi se sami slučajno trebali ispravno klasificirati, za usporedbu s postotkom koji je ispravno klasificiran postupkom klasifikacije. Ako postoje dvije grupe jednake veličine, 50% slučajeva trebalo bi biti ispravno klasificirano na slučajan način. Međutim, kada postoji nejednak broj slučajeva u grupama, izračunavanje postotka slučajeva koji bi slučajno trebali biti ispravno klasificirani je malo kompliciranije.

Lakši način[6] za pronalaženje postotka točnosti je da prvo izračunamo broj slučajeva u svakoj grupi, koji bi slučajno trebali biti ispravno klasificirani, a zatim zbrajamo po grupama kako bismo pronašli ukupni očekivani postotak ispravno klasificiranih podataka. Razmotrimo primjer u kojem postoji 60 slučajeva: 10 u Grupi 1, 20 u Grupi 2 i 30 u Grupi 3. Ako su prethodne vjerojatnosti određene kao 0.17, 0.33 i 0.50, redom, programi će dodijeliti 10, 20, i 30 slučajeva grupama. Ako je 10 slučajeva nasumično dodijeljeno Grupi 1, 0.17 od njih (ili 1.7 slučajeva) bi slučajno trebalo biti ispravno klasificirano. Ako je 20 slučajeva nasumično dodijeljeno Grupi 2, 0.33 (ili 6.6 slučajeva) od njih bi trebalo biti slučajno ispravno klasificirano, a ako je 30 slučajeva dodijeljeno Grupi 3, 0.50 od njih (ili 15 slučajeva) bi trebalo biti slučajno ispravno klasificirano. Zbrajanjem 1.7, 6.6 i 15 dobiju se 23.3 slučajno ispravno klasificirana slučaja, tj. 39% od ukupnog broja slučajeva. Posto-

[6]Teži način za pronalaženje je proširiti multinomnu distribuciju, postupak koji je tehnički ispravniji, ali daje identične rezultate kao i jednostavnija metoda koja je ovdje prikazana.

tak točnosti pomoću klasifikacijskih jednadžbi mora biti znatno veći od postotka očekivane slučajne točnosti da bi jednadžbe bile korisne.

1.10.1 Krosvalidacija i novi slučajevi

Klasifikacija se temelji na klasifikacijskim koeficijentima izvedenim iz uzoraka i oni obično rade dobro za uzorak iz kojeg su izvedeni (engl. *resubstitution*). Budući da su koeficijenti samo procjene populacijskih klasifikacijskih koeficijenata, često je najpoželjnije znati koliko se koeficijenti generaliziraju na novi uzorak slučajeva. Testiranje korisnosti koeficijenata na novom uzorku naziva se krosvalidacija (engl. *cross-validation*). Jedan oblik krosvalidacije uključuje podjelu jednog velikog uzorka nasumično na dva dijela, te zatim izvođenje klasifikacijskih funkcija na jednom dijelu i njihovo testiranje na drugom. Drugi oblik krosvalidacije uključuje izvođenje klasifikacijskih funkcija iz uzorka izmjerenog u jednom trenutku i njihovo testiranje na uzorku izmjerenom kasnije.

Za veliki uzorak nasumično podijeljen na dijelove, jednostavno za neke slučajeve izostavimo informacije o stvarnoj pripadnosti grupi (sakrijemo ih u programu). U SAS DISCRIM proceduri, zadržani slučajevi stavljaju se u zasebnu podatkovnu datoteku, a zatim se za njih ispituje točnost s kojom funkcije klasifikacije predviđaju pripadnost grupi za nove podatke (u SAS-u se ovaj postupak naziva kalibracija (engl. *calibration*).

Mogući problemi pojavljuju se ako je originalni uzorak malen.

1.10.2 Klasifikacija Jackknifed

Pristranost ulazi u klasifikaciju ako su koeficijenti koji se koriste za dodjeljivanje slučaja grupi izvedeni, djelomično, iz slučaja. U Jackknifed klasifikaciji, podaci iz slučaja se izostavljaju kada se izračunavaju koeficijenti koji se koriste za dodjeljivanje grupi. Svaki slučaj ima skup koeficijenata koji se razvijaju iz svih ostalih slučajeva. Jackknifed klasifikacija daje realniju procjenu sposobnosti prediktora da razdvoje grupe.

1.10.3 Ocjenjivanje poboljšanja u klasifikaciji

U sekvencijalnoj LDA, korisno je utvrditi poboljšava li se klasifikacija dodavanjem novog skupa prediktora u analizu. McNemarov χ^2 s ponovljenim mjerenjima pruža jednostavan, jasan (ali zamoran) test poboljšanja. Slučajevi se ručno tabeliraju jedan po jedan prema tome jesu li točno ili netočno klasificirani prije koraka i nakon koraka u kojem se dodaju prediktori.

Slučajevi koji imaju isti rezultat u oba koraka (bilo ispravno klasificirani, na slici 1.2 polje A, ili netočno klasificirani, polje D, zanemaruju se jer se oni ne mijenjaju. Dakle, χ^2 za promjenu je

		Raniji korak klasifikacije: Točno	Raniji korak klasifikacije: Netočno
Kasniji korak klasifikacije	Točno	A	B
	Netočno	C	D

Slika 1.4: Slovima A, B, C, D označeni su brojevi točno, odnosno netočno klasificiranih sučajeva u ponovljenom mjerenju.

$$\chi^2 = \frac{(|B - C| - 1)^2}{B + C} \qquad st.sl. = 1. \tag{1.10}$$

Obično istraživača zanima samo poboljšanje u χ^2, odnosno u situacijama kada je $B > C$ jer je tada više slučajeva ispravno klasificirano nakon dodavanja prediktora. Kada je $B > C$ i χ^2 veći od 3.84 (kritična vrijednost χ^2 s jednim stupnjem slobode za $\alpha = 0.05$), dodani prediktori pouzdano poboljšavaju klasifikaciju.

Uz vrlo velike uzorke, ručno tabeliranje slučajeva nije razumno. Alternativan, ali možda manje poželjan, postupak je ispitivanje značajnosti razlike između dvije lambde.[3] Wilksova lambda iz koraka s većim brojem prediktora, Λ_2, je podijeljena lambdom iz koraka s manje prediktora, Λ_1, da bi se dobila Wilksova lambda za testiranje značajnosti razlike između dvije lambde, Λ_D,

$$\Lambda_D = \frac{\Lambda_2}{\Lambda_1}. \tag{1.11}$$

Poglavlje 2

Primjena linearne diskriminantne analize na primjeru daljinskih istraživanja usjeva

Za ilustraciju primjene linearne diskriminantne analize, koristit ćemo podatke daljinskih istraživanja usjeva, dostupne u SAS dokumentaciji. Daljinska istraživanja (engl. remote sensing) predstavljaju metodske postupke prikupljanja i interpretacije podataka o objektima i pojavama iz daljine (zrakoplov, satelit i slično), pri čemu mjerni instrumenti ne dolaze u dodir s objektima istraživanja. [6] U ovom skupu podataka, 36 opservacija grupirano je u pet usjeva: Clover (djetelina), Corn (kukuruz), Cotton (pamuk), Soybeans (soja) i Sugarbeets (šećerna repa). Četiri mjere nazvane $x1$, $x2$, $x3$ i $x4$ čine deskriptivne varijable, čiji točni nazivi nam nisu poznati. Podaci su analizirani statističkim softverom SAS.

2.1 Deskriptivna statistika

Tablica 2.1 prikazuje deskriptivnu statistiku nezavisnih varijabli (prediktora) $x1$ - $x4$. N predstavlja broj opservacija, $Mean$ aritmetičku sredinu podataka, $Std\ Dev$ standardnu devijaciju, te $Minimum$, $Median$ i $Maximum$ redom, podatak s najmanjom vrijednošću u skupu, medijan podataka, te podatak s najvećom vrijednošću.

Tablica 2.1: Deskriptivna statistika nezavisnih varijabli (SAS ispis)

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Median	Maximum
x1	36	31.5556	19.2865	12.0000	26.0000	96.0000
x2	36	29.6944	12.5921	8.0000	26.5000	58.0000
x3	36	28.8611	15.3347	2.0000	25.5000	75.0000
x4	36	35.8611	17.6912	11.0000	32.0000	78.0000

Iz tablice 2.1 možemo primjetiti da su sve varijable prediktora jednakih veličina, tj. da nema podataka koji nedostaju.

Provjerimo i korelaciju varijabli prediktora pomoću procedure PROC CORR i naredbe PEARSON, koja će nam dati Pearsonov koeficijent korelacije za svaki par varijabli $x1$-$x4$..

Tablica 2.2: Pearsonov koeficijent korelacije varijabli prediktora (SAS ispis)

Pearson Correlation Coefficients, N = 36 Prob > |r| under H0: Rho=0				
	x1	x2	x3	x4
x1	1.00000	0.29731 0.0782	0.41529 0.0118	0.18504 0.2800
x2	0.29731 0.0782	1.00000	0.41511 0.0118	-0.15025 0.3817
x3	0.41529 0.0118	0.41511 0.0118	1.00000	-0.29454 0.0812
x4	0.18504 0.2800	-0.15025 0.3817	-0.29454 0.0812	1.00000

Iz tablice 2.2 iščitavamo da su, na razini značajnosti $\alpha = 0.05$, statistički značajno korelirane varijable $x1$ i $x3$, te $x2$ i $x3$.

Kod u SAS-u kojim provodimo linearnu diskriminantnu analizu

```
title2 'Linearna diskriminantna analiza';
proc discrim data=crops distance anova manova pool=yes
             list crossvalidate;
   class Crop;
   priors prop;
   id xvalues;
   var x1 x2 x3 x4;
run;
```

PROC DISCRIM procedura započinje prikazom informacija o varijablama koje sudjeluju u analizi, a iz kojeg možemo iščitati broj opservacija, broj kvantitativnih varijabli te broj klasa (grupa) u klasifikacijskoj varijabli. Također, dostupne su i informacije o frekvenciji pojedine klase, njezinoj težini, proporciji u ukupnom uzorku, te vjerojatnost pojedine klase prije analize (engl. *Prior Probability*), koja je proporcionalna veličini klase u ukupnom uzorku.

Iz tablice 2.3 vidimo da je u analizi sudjelovalo svih 36 opservacija, od kojih najveću frekvenciju ima djetelina, slijedi je kukuruz, a najmanju frekvenciju imaju pamuk, soja i šećerna repa. Najveću vjerojatnost klasificiranja prije analize ima grupa s najvećom frekvencijom, tj. djetelina (0.305556).

Tablica 2.3: Deskriptivne informacije (SAS ispis)

Linearna diskriminantna analiza

The DISCRIM Procedure

Total Sample Size	36	**DF Total**	35
Variables	4	**DF Within Classes**	31
Classes	5	**DF Between Classes**	4

Number of Observations Read	36
Number of Observations Used	36

Class Level Information

Crop	Variable Name	Frequency	Weight	Proportion	Prior Probability
Clover	Clover	11	11.0000	0.305556	0.305556
Corn	Corn	7	7.0000	0.194444	0.194444
Cotton	Cotton	6	6.0000	0.166667	0.166667
Soybeans	Soybeans	6	6.0000	0.166667	0.166667
Sugarbeets	Sugarbeets	6	6.0000	0.166667	0.166667

2.2 Provjera pretpostavki linearne diskriminantne analize

2.2.1 Normalnost

Prije samog izvođenja diskriminacijskih funkcija, provjerit ćemo imaju li varijable prediktora $x1$-$x4$ normalne distribucije. Normalnost ćemo provjeriti PROC UNIVARIATE NORMAL PLOT procedurom.

Za varijable $x1$-$x4$ dobivamo sljedeće rezultate.

Tablica 2.4: Test normalnosti za varijablu $x1$ (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.804303	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.227349	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.366988	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.137024	**Pr > A-Sq**	<0.0050

Tablica 2.5: Test normalnosti za varijablu $x2$ (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.936981	**Pr < W**	0.0409
Kolmogorov-Smirnov	**D**	0.177362	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.173126	**Pr > W-Sq**	0.0112
Anderson-Darling	**A-Sq**	0.940496	**Pr > A-Sq**	0.0168

Tablica 2.6: Test normalnosti za varijablu $x3$ (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.862656	**Pr < W**	0.0004
Kolmogorov-Smirnov	**D**	0.254728	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.403089	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.104922	**Pr > A-Sq**	<0.0050

Tablica 2.7: Test normalnosti za varijablu $x4$ (SAS ispis)

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.92932	Pr < W	0.0239
Kolmogorov-Smirnov	D	0.18078	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.151713	Pr > W-Sq	0.0221
Anderson-Darling	A-Sq	0.869487	Pr > A-Sq	0.0235

Navedena procedura izvodi četiri testa normalnosti podataka, Shapiro-Wilksov test, Kolmogorov-Smirnovljev, Cramer von Misesov te Anderson-Darlingov test. Nulta hipoteza ovih testova je da podaci dolaze iz normalne distribucije,a odbacuje se za p-vrijednosti manje od 0.05.

Na razini značajnosti $\alpha = 0.05$ možemo odbaciti normalnost za sve varijable prediktora. Iz gornjih tablica, vidimo da su varijable $x2$ i $x4$ najbliže normalnoj distribuciji.

2.2.2 Homogenost matrice kovarijance

U proceduri koju koristimo, pretpostavljamo da su varijance usjeva jednake pa upotrebljavamo naredbu *POOL=YES*, koja koristi združenu matricu kovarijance kao osnovu za računanje generaliziranih kvadrata udaljenosti. Tom naredbom dobivamo diskriminacijske funkcije koje su linearne, dok bi naredbom *POOL=NO* dobili kvadratne diskriminacijske funkcije.

Treća opcija za naredbu *POOL* je *POOL=YES*, koja testira homogenost matrice kovarijance unutar grupa, pri čemu osnovna i alternativna hipoteza glase:

H_0 : Matrica kovarijance unutar grupa je homogena
H_1 : Matrica kovarijance unutar grupa nije homogena.

Ukoliko matrica nije homogena, već heterogena, *PROC DISCRIM POOL=TEST* računa kvadratne diskriminacijske funkcije i pritom upotrebljava zasebne matrice kovarijance, a ne združenu matricu kovarijance kao kod računanja linearnih diskriminacijskih funkcija.

Provjerimo homogenost matrice kovarijance za naš primjer. PROC DISCRIM procedurom u SAS-u dobivamo tablicu 2.8.

Tablica 2.8: Test homogenosti matrice kovarijance (SAS ispis)

Chi-Square	DF	Pr > ChiSq
98.022966	40	<.0001

Test je proveden na razini značajnosti $\alpha = .10$, a budući da je p-vrijednost manja od .0001 odbacujemo nultu hipotezu o homogenosti matrice.

Iako pretpostavka o homogenosti naše matrice kovarijance "pada u vodu", analizu ćemo nastaviti pod pretpostavkom da je matrica homogena kako bismo mogli ilustrirati primjer linearne analize. U suprotnom, procedura će nam dati kvadratne diskriminacijske funkcije, što više nije dio linearne diskriminantne analize.

2.3 Udaljenost između grupa

Naredbom *DISTANCE* dobivamo kvadriranu Mahalanobisovu udaljenost između grupa, tj. njihovih centroida.

Tablica 2.9: Kvadrirana Mahalanobisova udaljenost između grupa usjeva

Squared Distance to Crop					
From Crop	Clover	Corn	Cotton	Soybeans	Sugarbeets
Clover	0	4.25308	0.86617	2.58313	1.48910
Corn	4.25308	0	1.88446	0.73031	2.89043
Cotton	0.86617	1.88446	0	1.43467	1.29556
Soybeans	2.58313	0.73031	1.43467	0	1.07646
Sugarbeets	1.48910	2.89043	1.29556	1.07646	0

Iz tablice 2.9 vidimo da je najveća udaljenost između usjeva kukuruza i djeteline, tj. te dvije grupe se najviše razlikuju, pa očekujemo manju pogrešku klasifikacije podatka iz

jedne od tih grupu u onu drugu. S druge strane, najmanja je udaljenost između usjeva soje i kukuruza, te pamuka i djeteline. Grupe s manjom udaljenošću, manje se razlikuju i veća je vjerojatnost pogrešne klasifikacije među njima.

Ista naredba, provodi i F test značajnosti udaljenosti između dvaju usjeva.

Tablica 2.10: p-vrijednosti za test značajnosti kvadriranih Mahalanobisovih udaljenosti među usjevima

Prob > Mahalanobis Distance for Squared Distance to Crop					
From Crop	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**
Clover	1.0000	0.0096	0.5605	0.0874	0.2921
Corn	0.0096	1.0000	0.2679	0.7127	0.1063
Cotton	0.5605	0.2679	1.0000	0.4385	0.4898
Soybeans	0.0874	0.7127	0.4385	1.0000	0.5796
Sugarbeets	0.2921	0.1063	0.4898	0.5796	1.0000

Iz tablice 2.10 vidimo da se za razinu značajnosti $\alpha = .10$ značajno razlikuju djetelina i kukuruz te djetelina i soja, dok se na razini značajnosti $\alpha = .05$ značajno razlikuju jedino djetelina i soja.

Budući da većina grupnih centroida usjeva nije značajno udaljena, tj. usjevi se značajno ne razlikuju, očekujemo dosta pogrešaka u klasifikaciji.

2.4 Statistička značajnost

Naredbe *ANOVA* i *MANOVA* daju nam uvid u rezultate testova značajnosti utjecaja pojedinih prediktora u klasifikaciji, te test ukupne značajnosti modela. Zanima nas koje varijable imaju značajan utjecaj u modelu.

Tablica 2.11: Test značajnosti varijabli prediktora u klasifikaciji (SAS ispis)

Univariate Test Statistics							
F Statistics, Num DF=4, Den DF=31							
Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
x1	19.2865	16.0960	13.1596	0.3831	0.6210	4.81	0.0039
x2	12.5921	12.6748	4.4468	0.1026	0.1144	0.89	0.4837
x3	15.3347	15.0642	6.4427	0.1453	0.1699	1.32	0.2855
x4	17.6912	18.3787	4.0962	0.0441	0.0461	0.36	0.8367

Iz tablice 2.11 vidimo da jedino varijabla $x1$ značajno doprinosi klasifikaciji podataka u grupe. Za standardne razine značajnosti zaključujemo da varijable $x2$, $x3$ i $x4$ ne utječu značajno na klasifikaciju podataka.

Tablica 2.12: Test ukupne značajnosti klasifikacije (SAS ispis)

Multivariate Statistics and F Approximations					
S=4 M=-0.5 N=13					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.47687044	1.48	16	86.179	0.1271
Pillai's Trace	0.61336282	1.40	16	124	0.1506
Hotelling-Lawley Trace	0.91548442	1.55	16	50.286	0.1195
Roy's Greatest Root	0.67419307	5.22	4	31	0.0025
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

Budući da dizajn našeg istraživanja nije idealan, kao kriterij ukupne statističke značajnosti koristimo Pillaijev trag. Iz tablice 2.12 iščitavamo da klasifikacija nije statistički značajna, tj. da varijable koje sudjeluju u klasifikaciji, ne utječu značajno na samu klasifikaciju podataka.

2.5 Diskriminacijske funkcije

Konačno, PROC DISCRIM procedura daje nam koeficijente linearnih diskriminacijskih funkcija za usjeve, prikazane u tablici 2.13.

Tablica 2.13: Koeficijenti linearnih diskriminacijskih funkcija za usjeve

Linear Discriminant Function for Crop					
Variable	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**
Constant	-10.98457	-7.72070	-11.46537	-7.28260	-9.80179
x1	0.08907	-0.04180	0.02462	0.0000369	0.04245
x2	0.17379	0.11970	0.17596	0.15896	0.20988
x3	0.11899	0.16511	0.15880	0.10622	0.06540
x4	0.15637	0.16768	0.18362	0.14133	0.16408

Iz tablice 2.13 možemo iščitati i klasifikacijske jednadžbe za grupe.

Klasifikacijski rezultat za svaku grupu nalazimo pomoću jednadžbe 1.4, tj. množenjem originalne vrijednosti svakog prediktora odgovarajućim koeficijentom diskriminacijske, odnosno klasifikacijske funkcije, te zbrajanjem tih produkata sa konstantnim koeficijentom. Budući da u našem primjeru imamo 5 grupa usjeva, dobit ćemo i 5 klasifikacijskih jednadžbi.

Klasifikacijske jednadžbe za pet grupa usjeva:

$$C_{djetelina} = -10.98457 + 0.08907 \cdot X_1 + 0.17379 \cdot X_2 + 0.11899 \cdot X_3 + 0.15637 \cdot X_4 \ ,$$

$$C_{kukuruz} = -7.72070 + (-0.04180) \cdot X_1 + 0.11970 \cdot X_2 + 0.16511 \cdot X_3 + 0.16768 \cdot X_4 \ ,$$

$$C_{pamuk} = -11.46537 + 0.02462 \cdot X_1 + 0.17596 \cdot X_2 + 0.15880 \cdot X_3 + 0.18362 \cdot X_4 \ ,$$

$$C_{soja} = -7.28260 + 0.0000369 \cdot X_1 + 0.15896 \cdot X_2 + 0.10622 \cdot X_3 + 0.14133 \cdot X_4 \ ,$$

$$C_{secerna\ \ repa} = -9.80179 + 0.04245 \cdot X_1 + 0.20988 \cdot X_2 + 0.06540 \cdot X_3 + 0.16408 \cdot X_4\ ,$$

gdje X_1, X_2, X_3 i X_4 predstavljaju vrijednosti prediktora $x1$-$x4$ za svaki pojedini usjev u svakoj grupi.

Na primjer, želimo li izračunati klasifikacijski rezultat za jedan usjev kukuruza, čije su vrijednosti prediktora $x1$-$x4$ redom 16, 27, 31 i 33, onda klasifikacijski rezultat za taj usjev računamo tako da vrijednosti prediktora uvrstimo u svaku od prethodnih 5 jednadžbi te taj usjev svrstamo u grupu za koju ima najveći klasifikacijski rezultat.

$$C_{djetelina} = -10.98457 + 0.08907 \cdot 16 + 0.17379 \cdot 27 + 0.11899 \cdot 31 + 0.15637 \cdot 33 = 3.98178$$

$$C_{kukuruz} = -7.72070 + (-0.04180) \cdot 16 + 0.11970 \cdot 27 + 0.16511 \cdot 31 + 0.16768 \cdot 33 = 5.49425$$

$$C_{pamuk} = -11.46537 + 0.02462 \cdot 16 + 0.17596 \cdot 27 + 0.15880 \cdot 31 + 0.18362 \cdot 33 = 4.66173$$

$$C_{soja} = -7.28260 + 0.0000369 \cdot 16 + 0.15896 \cdot 27 + 0.10622 \cdot 31 + 0.14133 \cdot 33 = 4.96662$$

$$C_{sec.\ repa} = -9.80179 + 0.04245 \cdot 16 + 0.20988 \cdot 27 + 0.06540 \cdot 31 + 0.16408 \cdot 33 = 3.98621$$

Uvrštavanjem prediktora u klasifikacijske jednadžbe, dobivamo da ovaj usjev kukuruza ima najveći rezultat klasifikacije za grupu kukuruz pa taj usjev svrstavamo u grupu kukuruz. Ispravna klasifikacija u ovom slučaju.

Postupak se nastavlja za svaki pojedini usjev, te SAS procedurom PROC DISCRIM i naredbom *LIST* dobivamo tablicu, čiji jedan dio je prikazan u tablici 2.14,u kojoj je za svaki usjev prikazana grupa kojoj taj usjev stvarno pripada i grupa u koju je svrstan nakon postupka klasifikacije, te klasifikacijske vjerojatnosti. Usjev je svrstan u onu grupu za koju ima najveću klasifikacijsku vjerojatnost. Zvjezdicom su označeni usjevi koji su klasificirani u pogrešnu grupu.

Tablica 2.14: Ispis dijela tablice klasifikacije usjeva uporabom klasifikacijskih jednadžbi (SAS ispis)

Posterior Probability of Membership in Crop								
xvalues	From Crop	Classified into Crop		Clover	Corn	Cotton	Soybeans	Sugarbeets
15 32 32 15	Corn	Soybeans	*	0.0972	0.3278	0.1318	0.3420	0.1011
12 15 16 73	Corn	Corn		0.0454	0.5238	0.1849	0.1376	0.1083
20 23 23 25	Soybeans	Soybeans		0.1330	0.2804	0.1176	0.3305	0.1385
24 24 25 32	Soybeans	Soybeans		0.1768	0.2483	0.1586	0.2660	0.1502
21 25 23 24	Soybeans	Soybeans		0.1481	0.2431	0.1200	0.3318	0.1570
27 45 24 12	Soybeans	Sugarbeets	*	0.2357	0.0547	0.1016	0.2721	0.3359
12 13 15 42	Soybeans	Corn	*	0.0549	0.4749	0.0920	0.2768	0.1013
22 32 31 43	Soybeans	Cotton	*	0.1474	0.2606	0.2624	0.1848	0.1448
31 32 33 34	Cotton	Clover	*	0.2815	0.1518	0.2377	0.1767	0.1523
29 24 26 28	Cotton	Soybeans	*	0.2521	0.1842	0.1529	0.2549	0.1559
34 32 28 45	Cotton	Clover	*	0.3125	0.1023	0.2404	0.1357	0.2091
26 25 23 24	Cotton	Soybeans	*	0.2121	0.1809	0.1245	0.3045	0.1780

* Misclassified observation

Točan broj i postotak usjeva svrstanih u pojedinu grupu možemo vidjeti u tablici 2.15.

Tablica 2.15: Broj i postotak opservacija klasificiranih u pojedinu grupu usjeva te ukupna greška klasifikacije (SAS ispis)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CROPS
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Crop						
From Crop	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Clover	6 54.55	0 0.00	3 27.27	0 0.00	2 18.18	11 100.00
Corn	0 0.00	6 85.71	0 0.00	1 14.29	0 0.00	7 100.00
Cotton	3 50.00	0 0.00	1 16.67	2 33.33	0 0.00	6 100.00
Soybeans	0 0.00	1 16.67	1 16.67	3 50.00	1 16.67	6 100.00
Sugarbeets	1 16.67	1 16.67	0 0.00	2 33.33	2 33.33	6 100.00
Total	10 27.78	8 22.22	5 13.89	8 22.22	5 13.89	36 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop						
	Clover	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Rate	0.4545	0.1429	0.8333	0.5000	0.6667	0.5000
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Iz tablice 2.15 iščitavamo da je od 11 usjeva djeteline, njih 6 svrstano u grupu djetelina, 3 u grupu pamuk i 2 u grupu šećerna repa. Kod kukuruza, 6 ih je svrstano u grupu kukuruz, a 1 u grupu soja. Za pamuk, 3 u grupu djetelina, 1 u grupu pamuk i 2 u grupu soja. Za soju, 1 usjev svrstan je u grupu kukuruz, 1 u grupu pamuk, 3 u grupu soja i 1 u grupu šećerna repa. Kod šećerne repe, 1 usjev svrstan je u grupu djetelina, 1 u kukuruz, 2 u soju i 2 u grupu šećerna repa.

Ukupno je 10 slučajeva svrstano u grupu djetelina, 8 u grupu kukuruz, 5 u grupu pamuk, 8 u grupu soja i 5 u grupu šećerna repa. Prisjetimo se, grupe su na početku imale sljedeće frekvencije: djetelina 11, kukuruz 7, pamuk 6, soja 6 i šećerna repa 6.

Možemo primjetiti da je 18 od 36 usjeva svrstano u pogrešnu grupu, tj. 50% usjeva pogrešno je klasificirano i 50% ispravno. Ako bismo ručno izračunali očekivani broj i postotak slučajno ispravno razvrstanih usjeva na način opisan u poglavlju 1.10, onda bi taj broj iznosio 7.69 usjeva, odnosno 21.36% od ukupnog broja usjeva. Da bi jednadžbe bile korisne, postotak ispravno klasificiranih usjeva mora biti znantno veći od postotka očekivane slučajne točnosti. U ovom slučaju, klasifikacijske jednadžbe nisu od koristi.

U drugoj tablici vidimo da ukupna pogreška klasifikacije za usjeve iznosi 50.00%

U tablici 2.16 prikazana je procjena broja i postotka klasificiranih usjeva krosvalidacijom, koja služi ispitivanju točnosti s kojom klasifikacijske funkcije predviđaju pripadnost grupi, te procjena pogreške krosvalidacije.

Rezultat je još veća stopa pogrešne klasifikacije (66.67%), pa zaključujemo da klasifikacijske funkcije loše klasificiraju slučajeve u grupe.

Naredba *OUTSTAT=option* pohranjuje kalibracijske informacije u novi skup podataka za testiranje budućih opservacija.

Drugi dio PROC DISCRIM procedure koristi ove informacije za klasifikaciju skupa testnih podataka.

Tablica 2.16: Broj i postotak opservacija klasificiranih u pojedinu grupu usjeva krosvalidacijom (SAS ispis)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CROPS
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Crop						
From Crop	**Clover**	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Clover	4 36.36	3 27.27	1 9.09	0 0.00	3 27.27	11 100.00
Corn	0 0.00	4 57.14	1 14.29	2 28.57	0 0.00	7 100.00
Cotton	3 50.00	0 0.00	0 0.00	2 33.33	1 16.67	6 100.00
Soybeans	0 0.00	1 16.67	1 16.67	3 50.00	1 16.67	6 100.00
Sugarbeets	2 33.33	1 16.67	0 0.00	2 33.33	1 16.67	6 100.00
Total	9 25.00	9 25.00	3 8.33	9 25.00	6 16.67	36 100.00
Priors	0.30556	0.19444	0.16667	0.16667	0.16667	

Error Count Estimates for Crop						
	Clover	**Corn**	**Cotton**	**Soybeans**	**Sugarbeets**	**Total**
Rate	0.6364	0.4286	1.0000	0.5000	0.8333	0.6667
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

2.6 Klasifikacija testnih podataka

Za klasifikaciju skupa testnih podataka možemo koristiti kalibracijske informacije, pohranjene u skupu podataka *Cropstat*.

Kod u SAS-u kojim zadajemo i klasificiramo skup testnih podataka

```
data test;
   input Crop $ 1-10 x1-x4 xvalues $ 11-21;
   datalines;
Corn       16 27 31 33
Soybeans   21 25 23 24
Cotton     29 24 26 28
Sugarbeets54 23 21 54
Clover     32 32 62 16
;
proc discrim data=cropstat testdata=test testout=tout testlist;
   class Crop;
   testid xvalues;
   var x1-x4;
run;
```

Naredba *TESTLIST* ispisuje rezultate klasifikacije za svaku od opservacija iz skupa testnih podataka.

Tablica 2.17: Klasifikacijski rezultati testnih podataka (SAS ispis)

The DISCRIM Procedure
Classification Results for Test Data: WORK.TEST
Classification Results using Linear Discriminant Function

Posterior Probability of Membership in Crop								
xvalues	From Crop	Classified into Crop		Clover	Corn	Cotton	Soybeans	Sugarbeets
16 27 31 33	Corn	Corn		0.0894	0.4054	0.1763	0.2392	0.0897
21 25 23 24	Soybeans	Soybeans		0.1481	0.2431	0.1200	0.3318	0.1570
29 24 26 28	Cotton	Soybeans	*	0.2521	0.1842	0.1529	0.2549	0.1559
54 23 21 54	Sugarbeets	Clover	*	0.6215	0.0194	0.1250	0.0496	0.1845
32 32 62 16	Clover	Cotton	*	0.2163	0.3180	0.3327	0.1125	0.0206

*** Misclassified observation**

Koristeći kalibracijske informacije, tj. informacije klasifikacijskih funkcija iz prvog dijela PROC DISCRIM procedure, iz tablice 2.17 iščitavamo da je 3 od 5 usjeva iz skupa testnih podataka pogrešno klasificirano. Usjev kukuruza ispravno je klasificiran u grupu

kukuruz, također i usjev soje u grupu soja, ali usjev pamuka pogrešno je svrstan u grupu soja, usjev šećerne repe u grupu djetelina, a usjev djeteline u grupu pamuk.

Tablica 2.18: Greška klasifikacije testnih podataka (SAS ispis)

Error Count Estimates for Crop						
	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	1.0000	0.0000	1.0000	0.0000	1.0000	0.6389
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Iz tablice 2.18 iščitavamo da ukupna greška klasifikacije testnih podataka iznosi 0.6389, tj. 63.89% usjeva pogrešno je klasificirano.

2.7 Poboljšanje klasifikacije

Budući da smo analizu u ovom primjeru započeli s veoma lošim pretpostavkama, a ni klasifikacija nam nije dala zadovoljavajuće rezultate, pokušat ćemo najprije transformacijom varijabli prediktora dobiti bolju klasifikaciju.

2.7.1 Transformacija varijabli

U svrhu poboljšanja klasifikacije transformacijom podataka, proveli smo linearnu diskriminantnu analizu s logaritmiranim i korjenovanim vrijednostima varijabli prediktora. Nijedna od navedenih transformacija nije nam dala značajno bolje rezultate klasifikacije, već približno iste onima s originalnim podacima.

Procjene ukupnih pogrešaka klasifikacije dobivene krosvalidacijom u analizi s transformiranim varijablama dane su u tablicama 2.19 i 2.20 i iznose 0.5556 i 0.6111, što znači da je s logaritmiranim prediktorima 55.56% usjeva pogrešno klasificirano, a s korjenovanim 61.11%.

Tablica 2.19: Greška klasifikacije krosvalidacijom u analizi s logaritmiranim prediktorima (SAS ispis)

Error Count Estimates for Crop						
	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.4545	0.1429	1.0000	0.5000	0.8333	0.5556
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Tablica 2.20: Greška klasifikacije krosvalidacijom u analizi s korjenovanim prediktorima (SAS ispis)

Error Count Estimates for Crop						
	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.6364	0.1429	1.0000	0.5000	0.8333	0.6111
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

2.7.2 Kvadratna diskriminantna analiza

Budući da ni transformacijom podataka nismo dobili bolje rezultate, ostalo nam je još analizirati podatke kvadratnom diskriminantnom analizom.

Ovim korakom, izlazimo van granica linearne diskriminantne analize, no cilj nam je ustanoviti je li uzrok ovako lošoj klasifikaciji kršenje pretpostavki analize, ili diskriminantna analiza uopće nije dobar model za klasificiranje ovog skupa podataka.

Kvadratnu diskriminantnu analizu izvodimo SAS DISCRIM procedurom i naredbom *POOL=NO* koja daje kvadratne diskriminacijske funkcije za razdvajanje grupa usjeva.

Kod u SAS-u kojim izvodimo kvadratnu diskriminatnu analizu za podatke o usjevima

```
proc discrim data=crops method=normal pool=no crossvalidate;
class Crop;
priors prop;
id xvalues;
var x1-x4;
run;
```

Budući da nam je cilj značajno smanjiti pogrešku klasifikacije, iz SAS ispisa kvadratne analize izdvojili smo samo tablice s pogreškama klasifikacije.

U tablici 2.21 dana je pogreška klasifikacije uporabom kvadratnih diskriminacijskih funkcija i ona iznosi 0.1111, što je puno bolji rezultat, nego kod linearne analize (0.5000).

Tablica 2.21: Ggreška klasifikacije usjeva kvadratnom diskriminantnom analizom (SAS ispis)

Error Count Estimates for Crop						
	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.1818	0.0000	0.0000	0.0000	0.3333	0.1111
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

Međutim, pogreška klasifikacije krosvalidacijom dana je u tablici 2.22 i iznosi 0.5556, što znači da se ni klasifikacijske jednadžbe dobivene kvadratnom diskriminantnom analizom ne generaliziraju dobro za klasifikaciju novih uzoraka.

Tablica 2.22: Greška klasifikacije krosvalidacijom (SAS ispis)

Error Count Estimates for Crop						
	Clover	Corn	Cotton	Soybeans	Sugarbeets	Total
Rate	0.1818	0.7143	0.6667	0.6667	0.8333	0.5556
Priors	0.3056	0.1944	0.1667	0.1667	0.1667	

2.8 Kanonička diskriminantna analiza

U korist lošoj klasifikaciji ide i velika raspršenost podataka pa ćemo za kraj analize ovog primjera, provesti analizu dobivenih diskriminacijskih funkcija da bismo utvrdili koliko one pridonose razdvajanju grupa usjeva.

PROC CANDISC procedurom u SAS-u proveli smo kanoničku diskriminantnu analizu, koja nam daje broj izvedenih diskriminacijskih funkcija, njihovu značajnost te matricu korelacije prediktora sa diskriminacijskim funkcijama.

Tablica 2.23: Kanoničke korelacije koje objašnjavaju broj i značaj diskriminacijskih funkcija (SAS ispis)

The CANDISC Procedure

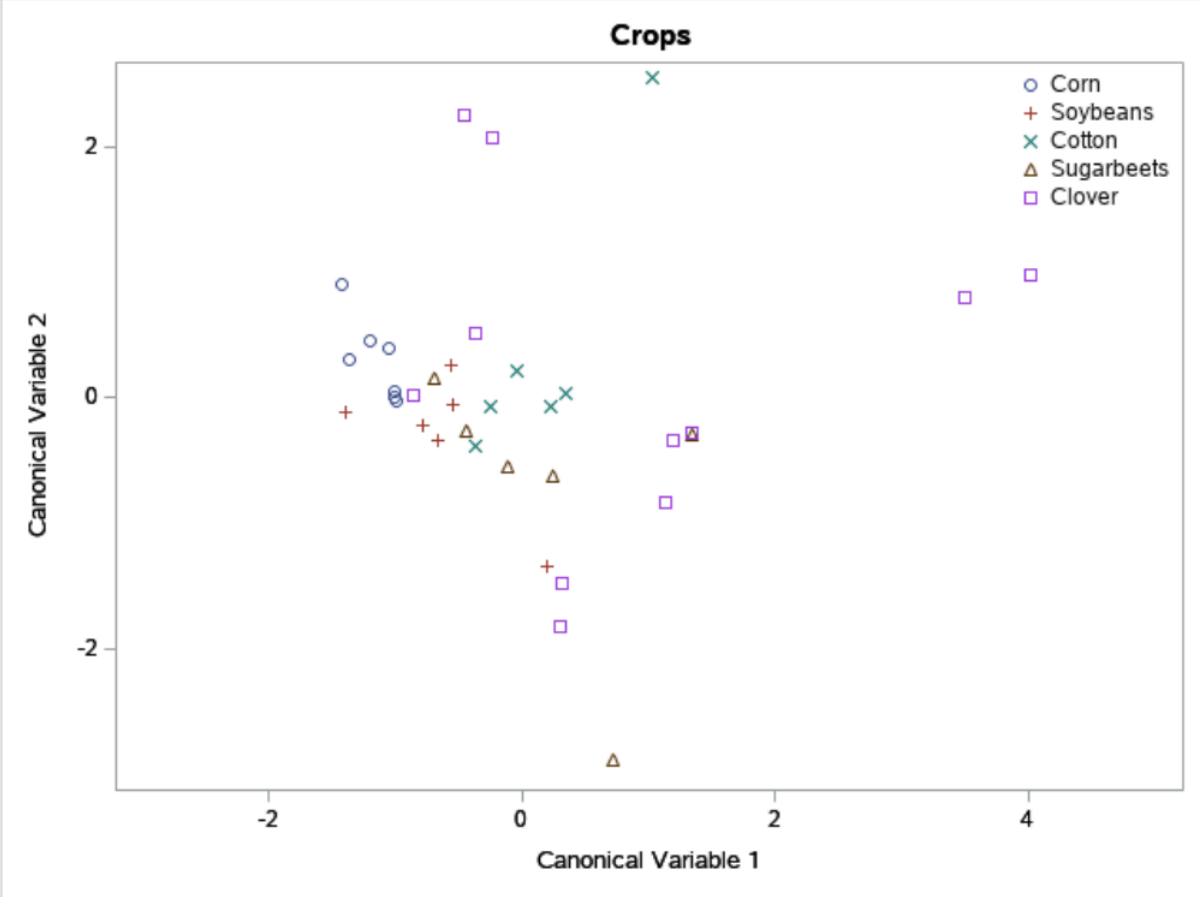
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.634584	0.546841	0.100963	0.402697	0.6742	0.4925	0.7364	0.7364
2	0.392116	0.268638	0.143042	0.153755	0.1817	0.1289	0.1985	0.9349
3	0.223852	0.147462	0.160561	0.050110	0.0528	0.0459	0.0576	0.9925
4	0.082467	.	0.167881	0.006801	0.0068		0.0075	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.47687044	1.48	16	86.179	0.1271
0.79837318	0.76	9	70.729	0.6515
0.94343017	0.44	4	60	0.7769
0.99319917	0.21	1	31	0.6482

Zbog bolje preglednosti, SAS ispis tablice 2.23 podijelili smo u dva dijela.

Prema tablici 2.23, za primjer podataka o usjevima izveden je maksimalni broj diskriminacijskih funkcija (4), od kojih nijedna nije značajna (p-vrijednosti veće od 0.10), odnosno nijedna statistički značajno ne dopinosi razdvajanju grupa usjeva.

Razdvajanje grupa pomoću prvih dviju diskriminacijskih funkcija, koje najviše doprinose razdvajanju, prikazano je na slici 2.1.



Slika 2.1: Prikaz prvih dviju diskriminacijskih funkcija za razdvajanje grupa usjeva (SAS ispis)

Na slici 2.1 osi prikazuju diskriminacijske funkcije, označene *Canonical Variable 1* i *Canonical Variable 2*. Vidimo da ni prva ni druga funkcija nema značajnu diskriminacijsku moć, jer se grupe usjeva međusubno jako preklapaju uzduž osi.

Tablica 2.24: Matrica strukture (SAS ispis)

Pooled Within Canonical Structure				
Variable	**Can1**	**Can2**	**Can3**	**Can4**
x1	0.950499	0.244477	-0.129788	-0.141199
x2	0.381259	-0.117118	0.420160	0.815102
x3	0.256786	0.825158	0.026957	0.502442
x4	0.158665	-0.059382	0.690471	-0.703241

Iz tablice 2.24 iščitavamo korelacije prediktora sa diskriminacijskim funkcijama. Primjećujemo da većina korelacija nije značajna (< 0.33), pa zaključujemo da prediktori ne doprinose značajno diskriminantnoj moći funkcija.

2.9 Zaključak

Na primjeru podataka daljinskih istraživanja usjeva, vidjeli smo da loše pretpostavke analize mogu značajno utjecati na rezultate klasifikacije.

Analizom smo pokazali da samo jedan od četiri prediktora značajno doprinosi klasifikaciji podataka, no ni to nije dovoljno da bi ukupna značajnost klasifikacije bila značajna.

Analizom diskriminacijskih funkcija utvrdili smo da nijedna od 4 izvedene funkcije ne razdvaja grupe usjeva statistički značajno.

Zaključujemo da podaci o usjevima nisu najbolji izbor za ilustraciju linearne diskriminantne analize, odnosno klasifikacije.

Poglavlje 3

Primjena linearne diskriminantne analize u istraživanju svojstava zrna triju sorti pšenice

U ovom primjeru, ispitivanu skupinu čine zrna triju različitih sorti pšenice (grupe) (Kama, Rosa i Canadian). Nasumično je odabrano po 70 opservacija za svaku sortu pšenice. Visokokvalitetna vizualizacija unutarnje strukture zrna detektirana je tehnikom meke rendgenske snimke (enlg. *soft X-ray technique*). Skup prediktora čini 7 kontinuiranih varijabli, *area* (površina), *perimeter* (opseg), *compactness* (kompaktnost), *length of kernel* (duljina zrna), *width of kernel* (širina zrna), *asymmetry coefficient* (koeficijent asimetrije) i *length of kernel groove* (duljina utora zrna), dobivenih mjerenjem geometrijskih svojstava zrna. Istraživanja su provedena na kombajnom požnjevenim zrnima pšenice s pokusnih polja, istraženih na Institutu za agrofiziku Poljske akademije znanosti u Lublinu, a podaci su dostupni na `http://archive.ics.uci.edu/ml/datasets/seeds`.

3.1 Deskriptivna statistika

Zbog lakše notacije i preglednosti ispisa, varijable *length of kernel*, *width of kernel*, *asymmetry coefficient* i *length of kernel groove* nazvali smo, redom, *KernelLength*, *KernelWidth*, *AsymmetryCoefficient* i *KernelGrooveLength*.

Iz tablice 3.1 vidimo da je učitano svih 210 opservacija. Broj nezavisnih varijabli (prediktora) je 7, a broj grupa (sorti) je 3. Frekvencije sorti pšenice su jednake, kao što smo i ranije naveli, za svaku sortu po 70 opservacija. Vjerojatnosti pripadanja pojedinoj sorti prije analize su proporcionalne veličinama grupa, pa su one jednake za sve sorte i

iznose 0.333333.

Tablica 3.1: Deskriptivna statistika (SAS ispis)

Total Sample Size	210	**DF Total**	209
Variables	7	**DF Within Classes**	207
Classes	3	**DF Between Classes**	2

Number of Observations Read	210
Number of Observations Used	210

Class Level Information					
Seed	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
Canadian	Canadian	70	70.0000	0.333333	0.333333
Kama	Kama	70	70.0000	0.333333	0.333333
Rosa	Rosa	70	70.0000	0.333333	0.333333

Tablica 3.2: Deskriptivna statistika varijabli prediktora (SAS ispis)

Variable	**N**	**Mean**	**Variance**	**Std Dev**	**Minimum**	**Median**	**Maximum**
Area	210	14.8475	8.4664	2.9097	10.5900	14.3550	21.1800
Perimeter	210	14.5593	1.7055	1.3060	12.4100	14.3200	17.2500
Compactness	210	0.8710	0.000558	0.0236	0.8081	0.8735	0.9183
KernelLength	210	5.6285	0.1963	0.4431	4.8990	5.5235	6.6750
KernelWidth	210	3.2586	0.1427	0.3777	2.6300	3.2370	4.0330
AsymmetryCoefficient	210	3.7002	2.2607	1.5036	0.7651	3.5990	8.4560
KernelGrooveLength	210	5.4081	0.2416	0.4915	4.5190	5.2230	6.5500

Prema tablici 3.2, za svaki od prediktora učitano je svih 210 vrijednosti, tj. nemamo podataka koji bi nedostajali. Varijabla *Area* ima najveću varijancu, dok varijabla *Compactness* ima najmanju.

Tablica 3.3: Korelacija prediktora (SAS ispis)

Pearson Correlation Coefficients, N = 210 Prob > |r| under H0: Rho=0							
	Area	**Perim**	**Compact**	**KL**	**KW**	**AsymCoeff**	**KGL**
Area	1.00000	0.99434 <.0001	0.60829 <.0001	0.94999 <.0001	0.97077 <.0001	-0.22957 0.0008	0.86369 <.0001
Perim	0.99434 <.0001	1.00000	0.52924 <.0001	0.97242 <.0001	0.94483 <.0001	-0.21734 0.0015	0.89078 <.0001
Compact	0.60829 <.0001	0.52924 <.0001	1.00000	0.36792 <.0001	0.76163 <.0001	-0.33147 <.0001	0.22682 0.0009
KL	0.94999 <.0001	0.97242 <.0001	0.36792 <.0001	1.00000	0.86041 <.0001	-0.17156 0.0128	0.93281 <.0001
KW	0.97077 <.0001	0.94483 <.0001	0.76163 <.0001	0.86041 <.0001	1.00000	-0.25804 0.0002	0.74913 <.0001
AsymCoeff	-0.22957 0.0008	-0.21734 0.0015	-0.33147 <.0001	-0.17156 0.0128	-0.25804 0.0002	1.00000	-0.01108 0.8732
KGL	0.86369 <.0001	0.89078 <.0001	0.22682 0.0009	0.93281 <.0001	0.74913 <.0001	-0.01108 0.8732	1.00000

Za ispis tablice 3.3, zbog preglednosti, koristili smo skraćene nazive prediktora (*Perim* za *Perimeter*, *Compact* za *Compactness*, *KL* za *KernelLength*, *KW* za *KernelWidth*, *AsymCoeff* za *AsymmetryCoefficient* i *KGL* za *KernelGrooveLength*).

Možemo primijetiti da su sve varijable prediktora međusobno značajno korelirane (s vrijednostima korelacija između 0.75 i 0.99), osim varijabli *AsymmetryCoefficient* i *Compactness*, koje su manje od ostalih korelirane s preostalim varijablama. No, to nas ne iznenađuje, budući da su to jedine dvije varijable koje nisu povezane s veličinom zrna. Varijable *Area* i *Perimeter* u najvećoj su (pozitivnoj) linearnoj korelaciji.

3.2 Provjera pretpostavki linearne diskriminantne analize

3.2.1 Normalnost

PROC UNIVARIATE procedurom ispitujemo normalnost varijabli prediktora.

Tablica 3.4: Test normalnosti varijable *Area* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.932594	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.106806	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.658401	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.448231	Pr > A-Sq	<0.0050

Tablica 3.5: Test normalnosti varijable *Perimeter* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.936162	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.106475	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.684765	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.379391	Pr > A-Sq	<0.0050

Tablica 3.6: Test normalnosti varijable *Compactness* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.973042	Pr < W	0.0005
Kolmogorov-Smirnov	D	0.062502	Pr > D	0.0443
Cramer-von Mises	W-Sq	0.196862	Pr > W-Sq	0.0058
Anderson-Darling	A-Sq	1.359826	Pr > A-Sq	<0.0050

Tablica 3.7: Test normalnosti varijable *KernelLength* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.943799	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.119154	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.655958	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.891567	Pr > A-Sq	<0.0050

Tablica 3.8: Test normalnosti varijable *KernelWidth* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.960624	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.068865	Pr > D	0.0163
Cramer-von Mises	W-Sq	0.271372	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.052553	Pr > A-Sq	<0.0050

Tablica 3.9: Test normalnosti varijable *AsymmetryCoefficient* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.983622	Pr < W	0.0154
Kolmogorov-Smirnov	D	0.048299	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.089059	Pr > W-Sq	0.1602
Anderson-Darling	A-Sq	0.635265	Pr > A-Sq	0.0973

Tablica 3.10: Test normalnosti varijable *KernelGrooveLength* (SAS ispis)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.924941	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.164777	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	1.303409	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	6.739782	**Pr > A-Sq**	<0.0050

Iz tablica 3.4 - 3.10 vidimo da na razini značajnosti $\alpha = 0.05$ možemo odbaciti pretpostavku o normalnosti za sve varijable prediktora, osim za varijablu *AsymmetryCoefficient*.

3.2.2 Homogenost matrice kovarijance

Naredbom *POOL=TEST* ispitujemo homogenost matrice kovarijance unutar grupa.

Tablica 3.11: Test homogenosti matrice kovarijance (SAS ispis)

Chi-Square	DF	Pr > ChiSq
674.117330	56	<.0001

Budući da iz tablice 3.11 iščitavamo p-vrijednost manju od razine značajnosti na kojoj je test proveden ($\alpha = 0.10$), zaključujemo da matrica kovarijance nije homogena.

Testiranjem uvjeta linearne diskriminantne analize, zaključujemo da pretpostavke nisu zadovoljene, kao ni u prethodnom primjeru. Analizu ćemo nastaviti pretpostavljajući zadovoljenu normalnost i homogenost, jer ukoliko dobijemo zadovoljavajuće rezultate klasifikacije, kršenje pretpostavki neće nas previše brinuti.

3.3 Udaljenost između grupa

Naredbom *DISTANCE* ispitujemo značajnost udaljenosti između grupnih centroida.

Tablica 3.12: Kvadrirana Mahalanobisova udaljenost između grupnih centroida (SAS ispis)

Squared Distance to Seed			
From Seed	**Canadian**	**Kama**	**Rosa**
Canadian	0	18.38956	35.38755
Kama	18.38956	0	27.42076
Rosa	35.38755	27.42076	0

Za ovaj primjer podataka, u tablici 3.12 vidimo da su centroidi sorti pšenice znatno udaljeniji, nego što je to bio slučaj kod usjeva u prethodnom primjeru (tablica 2.9).

Tablica 3.13: Test značajnosti kvadrirane Mahalanobisove udaljenosti između grupnih centroida (SAS ispis)

Prob > Mahalanobis Distance for Squared Distance to Seed			
From Seed	**Canadian**	**Kama**	**Rosa**
Canadian	1.0000	<.0001	<.0001
Kama	<.0001	1.0000	<.0001
Rosa	<.0001	<.0001	1.0000

Prema tablici 3.13 svi grupni centroidi statistički su značajno udaljeni, što nam daje naslutiti da će i klasifikacija biti puno bolja u odnosu na onu u poglavlju 2.

3.4 Statistička značajnost

Naredbom *ANOVA* analiziramo utjecaj pojedine varijable prediktora na klasifikaciju podataka u grupe.

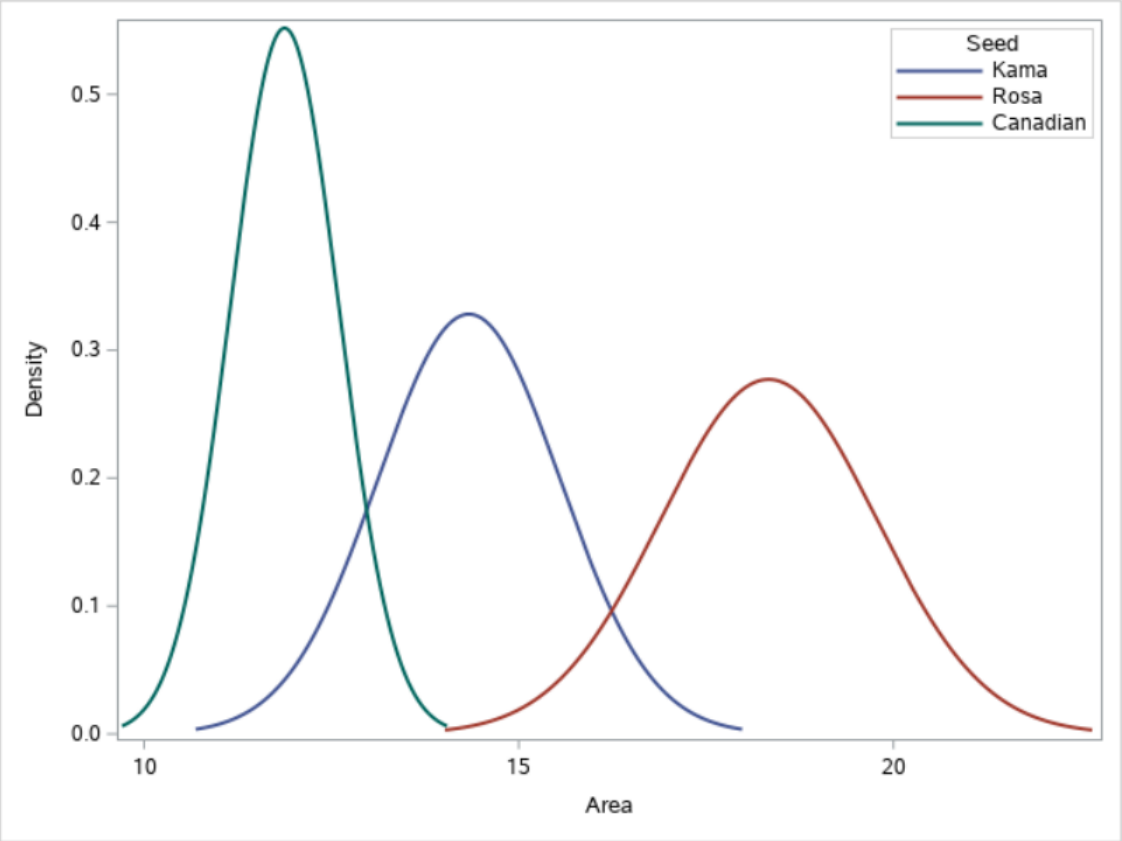
Prema tablici 3.14 i veličini F vrijednosti, varijabla *Area* najviše doprinosi klasifikaciji ($F = 548.19$). Slijede je varijable *Perimeter* ($F = 541.58$) i *KernelWidth* ($F = 406.30$). Klasifikaciji najmanje doprinosi varijabla *AsymmetryCoefficient* ($F = 51.89$), a slijedi je

Tablica 3.14: Test značajnosti varijabli prediktora u klasifikaciji (SAS ispis)

Univariate Test Statistics							
F Statistics, Num DF=2, Den DF=207							
Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
Area	2.9097	1.1652	3.2606	0.8412	5.2965	548.19	<.0001
Perimeter	1.3060	0.5256	1.4621	0.8396	5.2327	541.58	<.0001
Compactness	0.0236	0.0180	0.0188	0.4230	0.7330	75.87	<.0001
KernelLength	0.4431	0.2195	0.4710	0.7569	3.1129	322.19	<.0001
KernelWidth	0.3777	0.1710	0.4120	0.7970	3.9256	406.30	<.0001
AsymmetryCoefficient	1.5036	1.2330	1.0616	0.3339	0.5013	51.89	<.0001
KernelGrooveLength	0.4915	0.2311	0.5307	0.7809	3.5647	368.95	<.0001

varijabla *Compactness* ($F = 75.87$). Međutim, sve varijable prediktora statistički značajno doprinose klasifikaciji (p-vrijednosti manje od 0.001), pa zaključujemo kako je važno u modelu zadržati sve varijable da bismo dobili najtočniju klasifikaciju podataka.

Prikažimo grafički funkciju gustoće da bismo analizirali distribuciju varijable *Area* za svaku sortu pšenice.

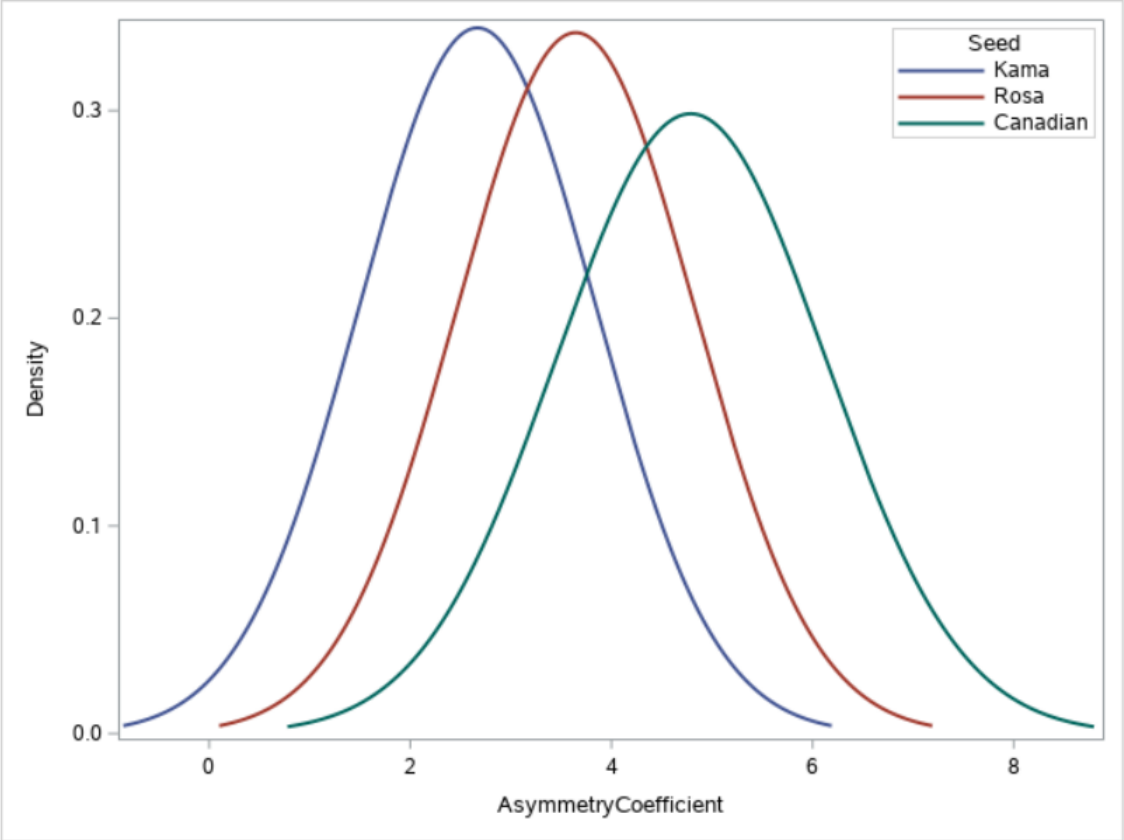


Slika 3.1: Funkcije gustoće za distribuciju varijable *Area* u sortama (SAS ispis)

Sa slike 3.1 vidimo da su središta distribucija za sve tri sorte znatno razmaknuta, no repovi distribucija se preklapaju za sorte Kama i Rosa, te Kama i Canadian. također, vidimo i minimalno preklapanje kod sorti Rosa i Canadian.

Kako je *Area* varijabla koja najviše razdvaja grupe, zaključujemo da se sorte Rosa i Canadian najviše međusobno razlikuju.

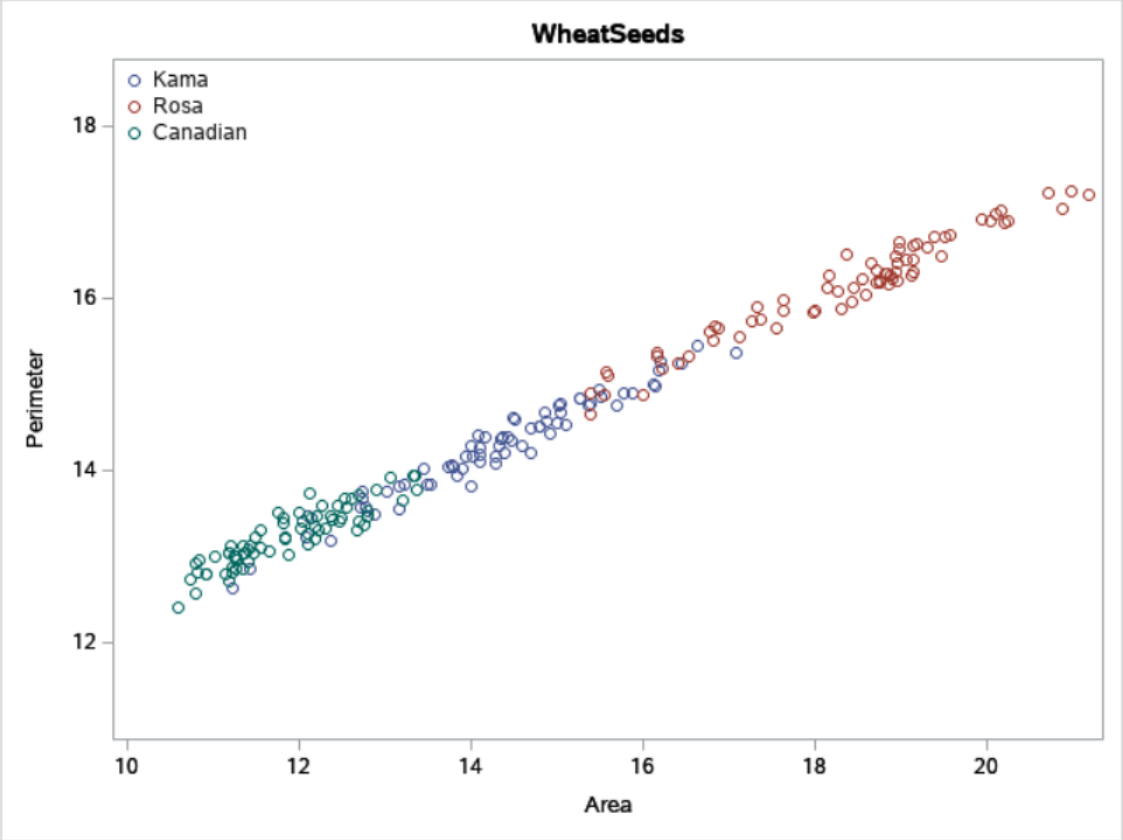
Analizirajmo i distribuciju varijable *AsymmetryCoefficient* za svaku od sorti.



Slika 3.2: Funkcije gustoće za distribuciju varijable *AsymmetryCoefficient* u sortama (SAS ispis)

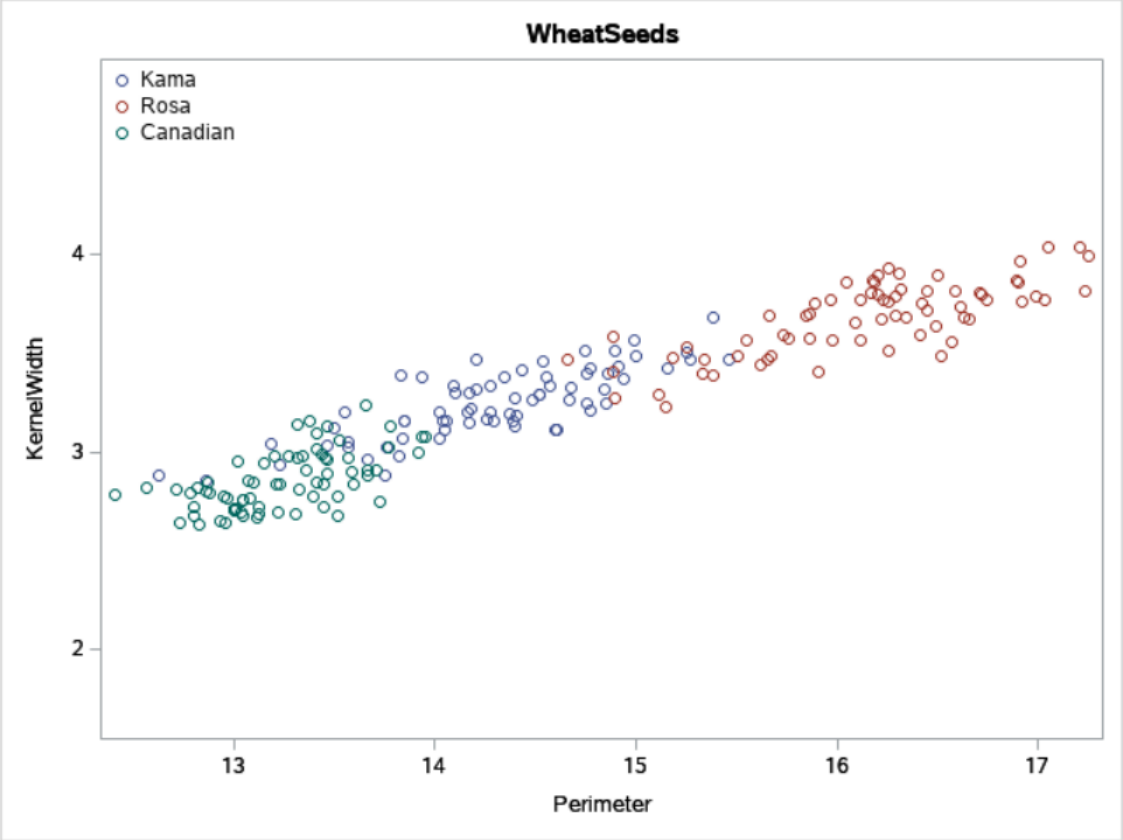
Za usporedbu sa slikom 3.1, na slici 3.2 vidimo da se distribucije varijable *AsymmetryCoefficient* jako preklapaju za sve tri sorte pšenice, no ipak, središta distribucija i dalje su razdvojena. Ovime opravdavamo značaj varijable *AsymmetryCoefficient* u budućoj klasifikaciji sorta pšenice.

Budući da varijable *Area*, *Perimeter* i *KernelWidth* najviše doprinose klasifikaciji podataka, nacrtajmo u parovima njihove zajedničke uzoračke distribucije u sortama. Na sljedećim grafovima, uzorci iz sorte Kama označeni su plavom bojom, iz sorte Rosa crvenom, a iz sorte Canadian zelenom.



Slika 3.3: Zajednička uzoračka distribucija varijabli *Area* i *Perimeter* u sortama pšenice (SAS ispis)

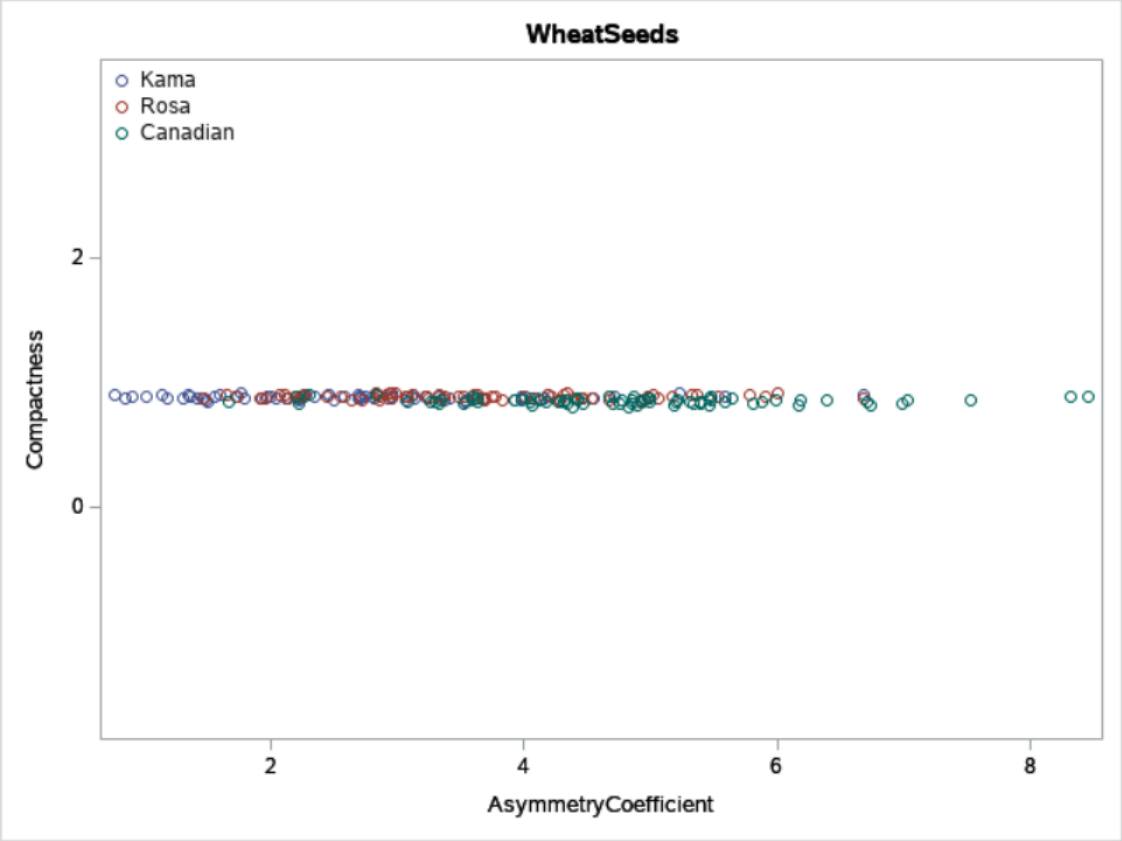
Prema slici 3.3, obje varijable, i *Area*, i *Perimeter*, značajno doprinose razdvajanju grupa.



Slika 3.4: Zajednička uzoračka distribucija varijabli *Perimeter* i *KernelWidth* u sortama pšenice (SAS ispis)

Iz slike 3.4 vidimo da varijabla *Perimeter* razdvaja sorte pšenice bolje od varijable *KernelWidth*.

Za usporedbu s varijablama koje najviše doprinose klasifikaciji, nacrtajmo i zajedničku uzoračku distribuciju varijabli *AsymmetryCoefficient* i *Compactness*, za koje smo ustanovili da najmanje doprinose klasifikaciji, no i dalje značajno. Očekujemo da se razdvajanje grupa neće vidjeti jasno kao na prethodnim grafovima.



Slika 3.5: Zajednička uzoračka distribucija varijabli *AsymmetryCoefficient* i *Compactness* u sortama pšenice (SAS ispis)

Na slici 3.5, pomoću distribucija varijabli *AsymmetryCoefficient* i *Compactness*, ne možemo razdvojiti sorte pšenice kao što smo to mogli s varijablama *Area*, *Perimeter* i *KernelWidth*.

U prijevodu, površina, opseg i širina zrna najviše razlikuju sorte pšenice Kama, Rosa i Canadian, dok se one najmanje razlikuju prema koeficijentu asimetrije i kompaktnosti.

Naredbom *MANOVA* testiramo ukupnu statističku značajnost klasifikacije. Za sva 4 kriterija navedena u tablici 3.15, možemo odbaciti nultu hipotezu i zaključiti da je klasifikacija dobivena pomoću svih 7 prediktora statistički značajna.

Tablica 3.15: Test ukupne statističke značajnosti klasifikacije (SAS ispis)

Multivariate Statistics and F Approximations					
S=2 M=2 N=99.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.03528718	124.14	14	402	<.0001
Pillai's Trace	1.60645126	117.79	14	404	<.0001
Hotelling-Lawley Trace	9.15273936	130.92	14	318.26	<.0001
Roy's Greatest Root	6.23679020	179.98	7	202	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

3.5 Rezultati klasifikacije

U tablici 3.16 su zapravo prikazani koeficijenti klasifikacijskih jednadžbi.

Broj značajnih diskriminacijskih funkcija i njihove koeficijente možemo iščitati iz ispisa procedure PROC CANDISC koja izvodi kanoničku diskriminantnu analizu. (Poglavlje 3.7).

Tablica 3.16: Koeficijenti linearnih diskriminacijskih funkcija za sorte pšenice (SAS ispis)

Linear Discriminant Function for Seed			
Variable	**Canadian**	**Kama**	**Rosa**
Constant	-40923	-41543	-41304
Area	-2445	-2463	-2453
Perimeter	5135	5175	5167
Compactness	50444	50804	50579
KernelLength	635.60886	658.39031	609.63796
KernelWidth	-1155	-1158	-1156
AsymmetryCoefficient	8.62597	7.26527	7.99451
KernelGrooveLength	-11.49046	-34.78367	-1.22101

Računanje klasifikacijskog rezultata za svaku pojedinu opservaciju izvodi se analogno

postupku opisanom u poglavlju 2.5.

U ovom primjeru imamo ukupno 3 različite sorte pšenice, odnosno 3 različite grupe, pa imamo i 3 klasifikacijske jednadžbe, čiji koeficijenti su dani u stupcima tablice 3.16.

Budući da u ovom primjeru imamo velik skup podataka, nećemo dati ispis klasifikacije za svaki pojedini slučaj, već samo ispis pogrešno klasificiranih sorti.

Tablica 3.17: Pogrešno klasificirane sorte pšenice i ukupna greška klasifikacije (SAS ispis)

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.WHEATSEEDS
Resubstitution Results using Linear Discriminant Function

Posterior Probability of Membership in Seed						
Obs	From Seed	Classified into Seed		Canadian	Kama	Rosa
9	Kama	Rosa	*	0.0001	0.1473	0.8526
24	Kama	Canadian	*	0.9177	0.0823	0.0000
61	Kama	Canadian	*	0.5588	0.4412	0.0000
62	Kama	Canadian	*	0.9898	0.0102	0.0000
198	Canadian	Kama	*	0.3603	0.6397	0.0000
200	Canadian	Kama	*	0.0878	0.9122	0.0000
202	Canadian	Kama	*	0.0617	0.9383	0.0000

*** Misclassified observation**

Error Count Estimates for Seed				
	Canadian	Kama	Rosa	Total
Rate	0.0429	0.0571	0.0000	0.0333
Priors	0.3333	0.3333	0.3333	

Prema tablici 3.17 klasifikacijske jednadžbe jako dobro klasificiraju podatke iz kojih su izvedene. Ukupno 7 od 210 zrna pšenice je klasificirano u pogrešnu sortu, što je greška od svega 3%.

Provjerimo i rezultate klasifikacije podataka metodom krosvalidacije.

U tablici 3.18 vidimo da su pogreške u klasifikaciji krosvalidacijom identične pogreškama klasifikacije na podacima iz kojih je model izveden, tj. i u ovom slučaju je svega 3% podataka pogrešno klasificirano.

Tablica 3.18: Pogrešno klasificirane sorte pšenice krosvalidacijom i ukupna greška klasifikacije (SAS ispis)

Classification Summary for Calibration Data: WORK.WHEATSEEDS
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Seed				
From Seed	**Canadian**	**Kama**	**Rosa**	**Total**
Canadian	67 95.71	3 4.29	0 0.00	70 100.00
Kama	3 4.29	66 94.29	1 1.43	70 100.00
Rosa	0 0.00	0 0.00	70 100.00	70 100.00
Total	70 33.33	69 32.86	71 33.81	210 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Seed				
	Canadian	**Kama**	**Rosa**	**Total**
Rate	0.0429	0.0571	0.0000	0.0333
Priors	0.3333	0.3333	0.3333	

Zaključujemo da klasifikacijske jednadžbe jako dobro klasificiraju sorte pšenica.

3.6 Klasifikacija testnih podataka

Za testiranje klasifikacije koristimo testni skup podataka koji se sastoji od 18 opservacija, po 6 za svaku sortu pšenice.

Tablica 3.19: Klasifikacija testnih podataka (SAS ispis)

The DISCRIM Procedure
Classification Results for Test Data: WORK.TEST
Classification Results using Linear Discriminant Function

Posterior Probability of Membership in Seed					
Obs	From Seed	Classified into Seed		Canadian	Kama	Rosa
1	Kama	Kama		0.0000	1.0000	0.0000
2	Kama	Kama		0.0000	1.0000	0.0000
3	Kama	Kama		0.0147	0.9853	0.0000
4	Kama	Kama		0.0001	0.9999	0.0000
5	Kama	Kama		0.0000	1.0000	0.0000
6	Kama	Kama		0.0003	0.9997	0.0000
7	Rosa	Rosa		0.0000	0.0004	0.9996
8	Rosa	Rosa		0.0000	0.0000	1.0000
9	Rosa	Rosa		0.0000	0.0000	1.0000
10	Rosa	Rosa		0.0000	0.0000	1.0000
11	Rosa	Rosa		0.0000	0.0000	1.0000
12	Rosa	Rosa		0.0000	0.0000	1.0000
13	Canadian	Canadian		1.0000	0.0000	0.0000
14	Canadian	Canadian		0.9887	0.0113	0.0000
15	Canadian	Canadian		1.0000	0.0000	0.0000
16	Canadian	Canadian		0.9989	0.0011	0.0000
17	Canadian	Canadian		0.9951	0.0042	0.0006
18	Canadian	Canadian		1.0000	0.0000	0.0000

Iz tablice 3.19 vidimo da su sva zrna pšenice iz skupa testnih podataka svrstana u ispravnu grupu, što znači da greška klasifikacije testnih podataka iznosi 0.0000 kao što

vidimo i u tablici 3.20.

Tablica 3.20: Greška klasifikacije testnih podataka (SAS ispis)

Number of Observations and Percent Classified into Seed				
From Seed	**Canadian**	**Kama**	**Rosa**	**Total**
Canadian	6 100.00	0 0.00	0 0.00	6 100.00
Kama	0 0.00	6 100.00	0 0.00	6 100.00
Rosa	0 0.00	0 0.00	6 100.00	6 100.00
Total	6 33.33	6 33.33	6 33.33	18 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Seed				
	Canadian	**Kama**	**Rosa**	**Total**
Rate	0.0000	0.0000	0.0000	0.0000
Priors	0.3333	0.3333	0.3333	

3.7 Kanonička diskriminantna analiza

PROC CANDISC procedurom izvodimo kanoničku diskriminantnu analizu s ciljem testiranja značajnosti diskriminacijskih funkcija.

Budući da u ovom primjeru imamo 3 grupe pšenice, možemo dobiti maksimalno dvije diskriminacijske funkcije.

Kanoničke korelacije koje dobivamo spomenutom procedurom, zapravo su veličine efekta diskriminacijskih funkcija. Broj kanoničkih korelacija odgovora broju diskriminacijskih funkcija.

Zbog bolje preglednosti, SAS ispis tablice 3.21 podijelili smo u dva dijela.

Tablica 3.21: Kanoničke korelacije koje opisuju diskriminacijske funkcije (SAS ispis)

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.928341	0.925530	0.009558	0.861817	6.2368	3.3208	0.6814	0.6814
2	0.862922	0.861152	0.017664	0.744634	2.9159		0.3186	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.03528718	124.14	14	402	<.0001
0.25536593	98.17	6	202	<.0001

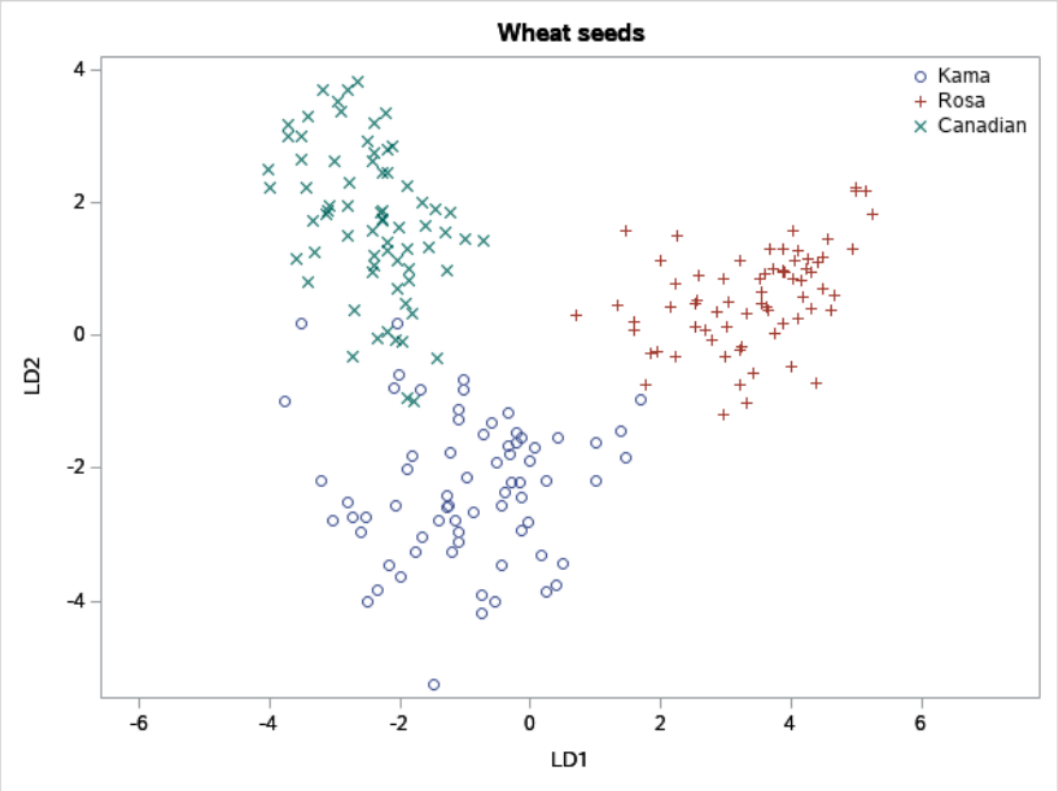
Prema tablici 3.21, postoje dvije diskriminacijske dimenzije (funkcije), od kojih obje statistički značajno razdvajaju sorte pšenice (p-vrijednosti manje od 0.0001). Kvadratne kanoničke korelacije za funkcije iznose redom 0.862 i 0.745, i to su zapravo veličine efekta svake diskriminacijske funkcije.

Tablica 3.22: Matrica korelacija prediktora i diskriminacijskih funkcija (SAS ispis)

Pooled Within Canonical Structure		
Variable	Can1	Can2
Area	0.912110	-0.192289
Perimeter	0.909132	-0.163354
Compactness	0.259645	-0.327407
KernelLength	0.705358	-0.058322
KernelWidth	0.768607	-0.287605
AsymmetryCoefficient	-0.080785	0.397454
KernelGrooveLength	0.728380	0.296229

Tablica 3.22 zapravo je matrica strukture u kojoj su sadržane korelacije prediktora i diskriminacijskih funkcija. Vidimo da su varijable *Area*, *Perimeter*, *KernelLength*, *KernelWidth* i *KernelGrooveLength* u jačoj korelaciji s prvom diskriminacijskom funkcijom, dok varijable *Compactness* i *AsymmetryCoefficient* jače koreliraju s drugom diskriminacijskom funkcijom. Budući da smo u poglavlju 1.7 naveli kako u obzir za interpretaciju uzimamo korelacije veće od 0.33, zaključujemo da sve varijable prediktora značajno doprinose diskriminaciji grupa. Pritom, varijablama koje su jače korelirane s prvom funkcijom, zajednička osobina je što su to mjere povezane s veličinom zrna, dok varijable koje jače koreliraju s drugom funkcijom nisu.

Pomoću rezultata kanoničke diskriminantne analize, lako možemo ucrtati naše opservacije po grupama duž diskiminacijskih osi.



Slika 3.6: Prikaz diskriminacijskih funkcija za razdvajanje sorti pšenice (SAS ispis)

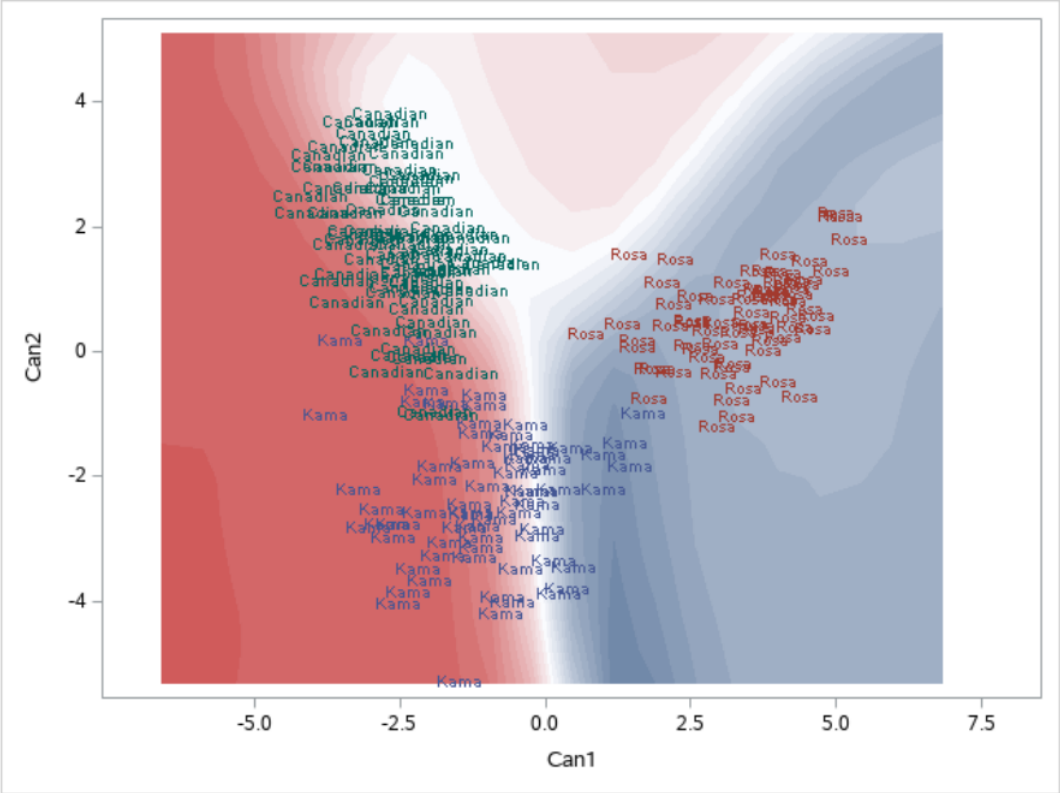
Na slici 3.6 prikazano je razdvajanje sorti pšenica pomoću dviju dobivenih diskriminacijskih funkcija. Funkcije su prikazane kao osi grafa, označene *LD1* (prva diskriminacijska funkcija) i *LD2* (druga diskriminacijska funkcija).

Primjećujemo da prva diskriminacijska funkcija potpuno razdvaja sorte Rosa i Canadian, dobro razdvaja sorte Kama i Rosa, dok sorte Kama i Canadian ne razdvaja baš najbolje.

S druge strane, druga diskriminacijska funkcija dobro razdvaja sorte Kama i Canadian, no manje upsješno razdvaja sortu Rosa od preostalih dviju.

Zaključujemo da se sorta Rosa od preostalih razlikuje najviše po svojstvima vezanim uz veličinu zrna, dok se sorte Kama i Canadian međusobno više razlikuju po kompaktnosti zrna i koeficijentu asimetrije.

Za kraj ove analize, prikažimo još i konturirani graf diskriminacijskih funkcija i sorti pšenice. Obojana područja prikazuju predviđene klasifikacije za svaku od sorti, a stvarni klasifikacijski rezultati podataka ispisani su za svaku sortu.



Slika 3.7: Prikaz klasifikacijskih rezultata za sorte pšenice duž diskriminacijskih funkcija (SAS ispis)

Na slici 3.7 diskriminacijske osi (funkcije) označene su *Can1* i *Can2*. Vidimo da dvije diskriminacijske funkcije dobro razdvajaju sorte pšenice; većina stvarnih klasifikacija sorti

Rosa i Kama spada unutar granica podudarne predviđene klasifikacije sorti. Također, primjećujemo i da je sorta Canadian teže odvojiva od sorte Kama.

3.8 Zaključak

U ovom primjeru i primjeru iz poglavlja 2, promatrani skup podataka krši uvjete pod kojima se izvodi optimalna linearna diskriminantna analiza, no klasifikacija podataka u drugom primjeru dala je puno bolje rezultate od one u prvom.

U stvarnoj primjeni diskriminantne analize, rijetko ćemo se susresti s podacima koji idealno zadovoljavaju sve uvjete, kao što je bio slučaj i u našim primjerima. U praksi, ukoliko su barem rezultati klasifikacije zadovoljavajući, kršenje pretpostavki i uvjeta ostavit ćemo po strani. No, u tom slučaju valja uzeti u obzir da dobiveni rezultati i doneseni zaključci nisu najtočniji, ali nam mogu poslužiti kao vodilja za daljnje analize i istraživanja, tj. daju nam barem neku, djelomično točnu, informaciju o podacima.

Dodatak A

Korišteni SAS kod

SAS kod za primjer iz poglavlja 2

```
   title 'Diskriminantna analiza podataka daljinskih istraživanja usjeva';
title2 'Deskriptiva za prediktore';
proc means data=crops n mean var std min median max fw=8;
var x1 x2 x3 x4;
run;

   title2 'Pearsonov koeficijent korelacije prediktora';
proc corr data=crops pearson vardef=df;
var x1 x2 x3 x4;
run;

   title2 'Provjera normalnosti prediktora';
proc univariate data=crops Normal plot;
var x1 x2 x3 x4;
run;

   title2 'Testiranje homogenosti matrice kovarijance';
proc discrim data=crops pool=test;
class Crop;
var x1 x2 x3 x4;
run;

   title2 'Linearna diskriminantna analiza';
proc discrim data=crops distance anova manova pool=yes
```

```
list crossvalidate;
class Crop;
priors prop;
id xvalues;
var x1 x2 x3 x4;
run;

   title2 'Klasifikacija testnih podataka';
data test;
input Crop $ 1-10 x1-x4 xvalues $ 11-21;
datalines;
Corn 16 27 31 33
Soybeans 21 25 23 24
Cotton 29 24 26 28
Sugarbeets54 23 21 54
Clover 32 32 62 16
;

   proc discrim data=cropstat testdata=test testout=tout testlist;
class Crop;
testid xvalues;
var x1-x4;
run;

   proc print data=tout;
title 'Diskriminantna analiza daljinskih istraživanja usjeva';
title2 'Output Classification Results of Test Data';
run;

   title2 'Kanonička diskriminantna analiza';
proc candisc data=crops out=outcan distance anova;
class crop;
var x1 x2 x3 x4;
run;

   title2 'Grafički prikaz kanoničkih varijabli';
proc template;
define statgraph scatter;
begingraph / attrpriority=none;
```

```
entrytitle 'Crops';
layout overlayequated / equatetype=fit
xaxisopts=(label='Canonical Variable 1')
yaxisopts=(label='Canonical Variable 2');
scatterplot x=Can1 y=Can2 / group=crop name='crop'
markerattrs=(size=7px);
layout gridded / autoalign=(topright topleft);
discretelegend 'crop' / border=false opaque=false;
endlayout;
endlayout;
endgraph;
end;
run;

   proc sgrender data=outcan template=scatter;
run;

   title2 'Diskriminantna analiza s korjenovanim prediktorima';
data crops; set crops;
sx1=sqrt(x1);
sx2=sqrt(x2);
sx3=sqrt(x3);
sx4=sqrt(x4);
run;

   proc discrim data=crops distance anova manova pool=yes
list crossvalidate;
class Crop;
priors prop;
id xvalues;
var sx1 sx2 sx3 sx4;
run;

   title2 'Diskriminantna analiza s logaritmiranim prediktorima';
data crops; set crops;
sx1=log(x1);
sx2=log(x2);
sx3=log(x3);
sx4=log(x4);
```

```
run;
proc discrim data=crops distance anova manova pool=yes
list crossvalidate;
class Crop;
priors prop;
id xvalues;
var sx1 sx2 sx3 sx4;
run;

   title2 'Kvadratna diskriminantna analiza';
proc discrim data=crops method=normal pool=no crossvalidate;
class Crop;
priors prop;
id xvalues;
var x1-x4;
run;
```

SAS kod za primjer iz poglavlja 3

```
   title2 'Deskriptiva za prediktore';
proc means data=crops n mean var std min median max fw=8;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength;
run;

   title2 'Pearsonov koeficijent korelacije prediktora';
proc corr data=wheatseeds pearson vardef=df plots=matrix;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength;
run;

   title2 'Provjera normalnosti prediktora';
proc univariate data=wheatseeds Normal;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength;
```

```
run;

   title2 'Provjera homogenosti matrice kovarijance';
proc discrim data=wheatseeds pool=test;
class Seed;
priors prop;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength ;
run;

   title2 'Testni podaci';
data test;
input Seed $ Area Perimeter Compactness KernelLength KernelWidth
AsymmetryCoefficient KernelGrooveLength;
datalines;
Kama 15.26 14.84 0.871 5.763 3.312 2.221 5.22
Kama 16.14 14.99 0.9034 5.658 3.562 1.355 5.175
Kama 14.28 14.17 0.8944 5.397 3.298 6.685 5.001
Kama 14.11 14.1 0.8911 5.42 3.302 2.7 5
Kama 13.84 13.94 0.8955 5.324 3.379 2.259 4.805
Kama 12.11 13.47 0.8392 5.159 3.032 1.502 4.519
Rosa 16.41 15.25 0.8866 5.718 3.525 4.217 5.618
Rosa 17.99 15.86 0.8992 5.89 3.694 2.068 5.837
Rosa 19.46 16.5 0.8985 6.113 3.892 4.308 6.009
Rosa 16.17 15.38 0.8588 5.762 3.387 4.286 5.703
Rosa 18.76 16.2 0.8984 6.172 3.796 3.12 6.053
Rosa 20.24 16.91 0.8897 6.315 3.962 5.901 6.188
Canadian 11.35 13.12 0.8291 5.176 2.668 4.337 5.132
Canadian 12.1 13.15 0.8793 5.105 2.941 2.201 5.056
Canadian 10.79 12.93 0.8107 5.317 2.648 5.462 5.194
Canadian 13.32 13.94 0.8613 5.541 3.073 7.035 5.44
Canadian 13.34 13.95 0.862 5.389 3.074 5.995 5.307
Canadian 12.22 13.32 0.8652 5.224 2.967 5.469 5.221
;

   title2 'Linearna diskriminantna analiza';
proc discrim data=wheatseeds distance anova manova pool=yes listerr
crossvalidate testdata=test testlist;
class Seed;
```

```
priors prop;
var Area Perimeter Compactness KernelLength KernelWidth
AsymmetryCoefficient KernelGrooveLength ;
run;

   title2 'Plot of Estimated Densities';
data plotdata;
do AsymmetryCoefficient=-2 to 11 by 0.5;
output;
end;
run;

   %macro plotden;
title3 'Plot of Estimated Densities';
data plotd2;
set plotd;
g = 'Kama'; Density = kama; output;
g = 'Rosa'; Density = rosa; output;
g = 'Canadian'; Density = canadian; output;
label AsymmetryCoefficient='AsymmetryCoefficient';
run;
proc sgplot data=plotd2;
series y=Density x=AsymmetryCoefficient / group=g;
discretelegend;
run;
%mend;

   %macro plotprob;
title3 'Plot of Posterior Probabilities';
data plotp2;
set plotp; g = 'Kama'; Probability = kama; output;
g = 'Rosa'; Probability = rosa; output;
g = 'Canadian '; Probability = canadian; output;
label AsymmetryCoefficient='AsymmetryCoefficient';
run;
proc sgplot data=plotp2;
series y=Probability x=AsymmetryCoefficient / group=g;
discretelegend;
run;
```

```
%mend;

   proc discrim data=wheatseeds method=normal pool=yes
testdata=plotdata testout=plotp testoutd=plotd
short noclassify crosslisterr;
class Seed;
var AsymmetryCoefficient;
run;
%plotden;
%plotprob;

   title2 'Zajedničke uzoračke distribucije';
proc template;
define statgraph scatter;
begingraph;
entrytitle 'WheatSeeds';
layout overlayequated / equatetype=fit;
scatterplot x=Perimeter y=Area /
group=Seed name='seed';
layout gridded / autoalign=(topleft);
discretelegend 'seed' / border=false opaque=false;
endlayout;
endlayout;
endgraph;
end;
run;
proc sgrender data=wheatseeds template=scatter;
run;

   title2 'Kanonička diskriminantna analiza';
proc candisc data=wheatseeds out=outcan distance anova;
class Seed;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength ;
run;

   title2 'Grafički prikaz kanoničkih varijabli';
proc template;
define statgraph scatter;
```

```
begingraph / attrpriority=none;
entrytitle 'Wheat seeds';
layout overlayequated / equatetype=fit
xaxisopts=(label='LD1')
yaxisopts=(label='LD2');
scatterplot x=Can1 y=Can2 / group=Seed name='seed'
markerattrs=(size=7px);
layout gridded / autoalign=(topright topleft);
discretelegend 'seed' / border=false opaque=false;
endlayout;
endlayout;
endgraph;
end;
run;

   proc sgrender data=outcan template=scatter;
run;

   proc stepdisc data=wheatseeds;
class Seed;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength;
run;

   title2 'Konturirani prikaz kanoničkih varijabli';

   data fakedata;
do Area = 10 to 25 by 0.5;
do Perimeter = 10 to 20 by 0.5;
do Compactness= 0 to 1 by 0.1;
do KernelLength = 0 to 10 by 0.5;
do KernelWidth = 0 to 5 by 0.5;
do AsymmetryCoefficient = 0 to 10 by 0.5;
do KernelGrooveLength = 10 to 25 by 0.5;
output;
end;
end;
end;
end;
```

```
end;
end;
end;
run;

   proc discrim data=wheatseeds testdata=fakedata testout=fake_out
out=outcan canonical;
class seed;
var Area Perimeter Compactness KernelLength KernelWidth AsymmetryCoefficient
KernelGrooveLength;
run;

   data plotclass;
merge fake_out outcan;
run;

   proc template;
define statgraph classify;
begingraph;
layout overlay; contourplotparm x=Can1 y=Can2 z=_into_ / contourtype=fill
nhint = 30 gridded = false;
scatterplot x=Can1 y=Can2 / group=seed includemissinggroup=false
markercharactergroup = seed;
endlayout;
endgraph;
end;
run;

   proc sgrender data = plotclass template = classify;
run;
```

Bibliografija

[1] A. L. Comrey i H. B. Lee, *A first course in factor analysis (2nd. ed.)*, Hillsdale, NJ: Erlbaum, 1992.

[2] M. C. Costanza i A. A. Afifi, *Comparison of stopping rules in forward stepwise discriminant analysis*, Journal of the American Statistical Association (1979), br. 74, 777–785.

[3] J. W. Frane, *The univariate approach to repeated measures- foundation, advantages, and caveats*, BMD Technical Report No. 69. Health Sciences Computing Facility, University of California, Los Angeles, 1980.

[4] K.V. Mardia, *The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model*, Biometrika (1971), br. 58, 105–121.

[5] C. L. Olson, *Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens*, Psychological Bulletin (1979), br. 86, 1350–1352.

[6] M. Oluić, *Snimanje i istraživanje Zemlje iz svemira*, GEOSAT d.o.o., 2001.

[7] B. G. Tabachnick i L. S. Fidell, *Using multivariate statistics*, Pearson, 2013.

[8] J. A. Woodward, D. G. Bonett i M. L. Brecht, *Introduction to linear models and experimental design*, San Diego, CA: Harcourt Brace Jovanovich, 1990.

Sažetak

U ovom radu promatramo diskriminantnu analizu kao metodu kojom možemo klasificirati objekte u odgovarajuće grupe te objasniti stupanj odnosa između grupa i određenih osobina objekata, tzv. prediktora. Koncentrirali smo se na linearnu diskriminantnu analizu (LDA), kojoj je cilj procijeniti linearnu kombinaciju varijabli prediktora koja najbolje diskriminira pripadnost individualnih objekata grupi. Prednost LDA u odnosu na ostale klasifikacijske metode jest što LDA dopušta kategorijsku zavisnu varijablu, a nezavisna varijabla ne mora biti dihotomna.

U prvom poglavlju, analizirali smo glavne uvjete koji moraju biti zadovoljeni za dobivanje optimalnih rezultata analize; multivarijatna normalnost prediktora, homogenost matrice varijance i kovarijance te linearni odnosi varijabli prediktora unutar svake grupe. Prikazali smo izvod i testiranje diskriminacijskih funkcija za separaciju grupa, te klasifikacijskih jednadžbi za klasificiranje objekata u grupe. Opisali smo i tri glavne vrste linearne diskriminantne analize; direktnu, sekvencijalnu i stepwise analizu, te kriterije za procjenu značajnosti. Kanoničkim varijablama objasnili smo značaj diskriminacijskih funkcija i grafički prikazali dobivene separacije grupa.

U drugom poglavlju, primijenili smo opisanu analizu na bazu podataka o daljinskim istraživanjima pet različitih usjeva. Baza se sastoji od 36 opservacija koje predstavljaju usjeve i 4 varijable koje predstavljaju mjerenja dobivena daljinskim istraživanjem. Analiza je pokazala da samo jedna varijabla značajno utječe na klasifikaciju, ali sveukupna dobivena klasifikacija nije značajna. Transformacijom varijabli prediktora, pokušali smo poboljšati rezultate klasifikacije, no bezuspješno. izvođenjem kvadratne diskriminantne analize ustanovili smo da korišteni skup podataka nije pogodan za ilustraciju diskriminantne analize.

U trećem poglavlju, analizirali smo podatke o svojstvima zrna triju različitih sorti pšenice. Baza se sastoji od 210 opservacija koje predstavljaju sorte pšenice, te 7 varijabli koje predstavljaju geometrijska svojstva zrna pšenice. Izvođenjem linearne diskriminantne analize na ovom skupu podataka, zaključili smo da sve varijable prediktora značajno doprinose separaciji grupa i klasifikaciji opservacija u grupe. Za kraj, grafovima smo prikazali separacije sorti pšenice, te smo zaključili da na razdvajanje grupa najviše utječe veličina zrna pšenice.

Summary

In this paper, we considered discriminant analysis as a method by which we can classify objects into groups and explain the degree of relationship between groups and certain properties of objects, the so-called, predictors. We concentrated on the linear discriminant analysis (LDA), which aims to estimate the linear combination of predictors that best discriminates the affiliation of individual objects to a group. The advantage of LDA over other classification methods is that LDA allows a categoric dependent variable, and the independent variable does not have to be dichotomous.

In the first chapter, we analyzed the main conditions that must be met to obtain optimal analysis results: multivariate normality of predictors, homogeneity of the matrix of variance and covariance, and linear relationships of predictors within each group. We presented a derivation and testing of discriminant functions for group separation, and classification equations for classifying objects in groups. We also described three main types of linear discriminant analysis: direct, sequential, and stepwise analysis, and the criteria for assessing the significance. We explained the importance of discriminant functions with canonical variables and graphically presented the obtained separation of groups.

In the second chapter, we applied the described analysis to a database of remote sensing data of five different crops. The database consists of 36 observations representing crops and 4 variables representing measurements obtained by remote sensing. The analysis showed that only one variable significantly affects the classification, but the overall obtained classification did not turn out to be significant. By transforming the predictor variables, we tried to improve the results of the classification, but without success. By performing a quadratic discriminant analysis , we found that the data set used, is not suitable for illustrating discriminant analysis.

In the third chapter, we analyzed data on kernel properties of three different wheat varieties. The database consists of 210 observations representing wheat varieties, and seven variables representing the geometric properties of wheat kernel. By performing a linear discriminant analysis on this data set, we concluded that all predictor variables significantly contribute to the separation of groups and to classifying observations into the correct group. Finally, we plotted the separations of wheat cultivars and concluded that the separation of groups is mostly influenced by the kernel size.

Životopis

Rođena sam 16. kolovoza 1995. godine u Zagrebu. Školovanje sam započela u osnovnoj školi u Maču, te nastavila upisom u opću gimnaziju srednje škole u Zlataru. Preddiplomski studij Matematike na matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu, upisala sam 2014., a završila 2019. godine, te time stekla titulu prvostupnice matematike. Na istom odsjeku, 2019. godine, upisala sam i diplomski studij Matematičke statistike. Kroz fakultetsko obrazovanje sudjelovala sam na Otvorenim danima Matematičkog odsjeka PMF-a, bila članica Studentskog zbora Sveučilišta u Zagrebu te stalna članica Udruge zagorskih studenata.