https://datascience-project.eu/

Model de fișă de instruire

Titlu	ANALIZA DISCRIMINANT LINIARĂ
Cuvinte cheie (metaetichete)	analiza discriminant, clasificare, R, analiză bayesiană
Limba	Engleză
Obiective / Scopuri / Rezultate ale învățării	**obiectivul acestui modul este de a introduce și explica elementele de bază ale analizei discriminante liniare (LDA).** **La sfârșitul acestui modul vei fi capabil să:** - **Identificați situațiile în care LDA poate fi utilă** - **Calculați funcțiile LDA** - **Interpretarea rezultatelor produse de LDA descriptiv și predictiv**

Curs de pregatire:	
Cunoașterea științei datelor	
Modul de vizualizare a datelor și analiză vizuală	X
Introducere în știința datelor pentru științe umane și sociale	
Știința datelor pentru totdeauna	
Date de jurnalism și povestire	

Descriere	În acest modul de instruire veți fi introdus în utilizarea analizei discriminante lineare (LDA). LDA este o metodă de găsire a combinațiilor liniare de variabile care separă cel mai bine observațiile în grupuri sau clase și a fost dezvoltată inițial de Fisher (1936). Această metodă maximizează raportul dintre variația dintre clase și varianța în interiorul clasei în orice anumit set de date. Făcând acest lucru, variabilitatea între grupuri este maximizată, ceea ce are ca rezultat separabilitatea maximă. LDA poate fi folosit cu scopuri pur de clasificare, dar și cu obiective predictive.
Continut dispus pe 3 nivele	**1. INTRODUCERE: MOTIVAREA PRIN UN EXEMPLU ILUSTRATIV**

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Să presupunem că avem un eșantion de indivizi și observăm modul de transport (cu mașina, transportul public sau mersul pe jos) pe care îl folosesc de obicei pentru a se deplasa într-un oraș. Știm că alegerea modului de transport este parțial influențată de statutul lor economic și observăm date privind vârsta lor în ani și venitul anual al gospodăriei, împreună cu mijlocul de transport ales:

Dorim să știm cum aceste două covariante ajută la clasificarea (adică, la discriminarea) indivizilor, atribuindu-i unei categorii specifice de mod de transport. Putem observa că nu există o clasificare perfectă: persoanele cu venituri mari tind să folosească mașina mai des, dar există o mare suprapunere a categoriilor „mers pe jos” și „transport public” pentru cei cu venituri mai mici. Și există o suprapunere mai mare între categorii în ceea ce privește distribuția lor pe vârstă: persoanele în vârstă nu merg pe jos, dar la valori mai mici vârsta nu este un bun predictor al modului de transport. Aceasta este problema tipică pe care o abordează LDA.

2. LDA pentru clasificare

2.1. Formulare

Funcțiile LDA pot fi recuperate pentru a ajuta la clasificarea datelor pe baza unei matrice de covariantă $\mathbf{X}$. Similar analizei componentelor principale (PCA), funcțiile LDA urmăresc să găsească o combinație liniară a datelor originale ca:

$$\text{LDA} = \mathbf{u}^{\mathbf{T}}\mathbf{X}$$

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Funcțiile LDA pot fi recuperate pentru a ajuta la clasificarea datelor pe baza unei matrice de covariate $\mathbf{X}$. Similar analizei componentelor principale (PCA), funcțiile LDA urmăresc să găsească o combinație liniară a datelor originale ca:

$$u = \arg\max_{u} \frac{u^T B u}{u^T W u}$$

Coordonatele discriminante sunt obținute din vectorii proprii ai $W^{(-1)} B$.

2.2. Exemplu

Ca exemplu ilustrativ, rezolvăm problema de clasificare a modului de transport pe baza vârstei și venitului de către LDA în R. Acest lucru se poate face cu ușurință prin funcția „lda" din biblioteca „de masă". Pentru toată analiza prezentată aici, va trebui să instalăm și să încărcăm următoarele pachete R:

```
# LDA packages
install.packages("MVN")
install.packages("heplots")
install.packages("caret")
install.packages("MASS")
library(MVN)
library(heplots)
library(caret)
library(tidyverse)
library(MASS)
```

Datele studiate vin într-un fișier csv (numit „transnpor_example"), care poate fi importat cu ușurință în R rulând aceast cod:

```
# Get Data
transport <- read.csv(transport_example.csv)
View(transport)
transport <- as.data.frame(transport)
```

Pentru a avea o primă impresie asupra datelor, putem reprezenta un grafic eșantionul sub forma unui grafic de dispersie ca:

```
#scatterplots
ggplot(transport, aes(age, income)) +
  geom_point(aes(color = group))
```

Co-funded by the Erasmus+ Programme of the European Union

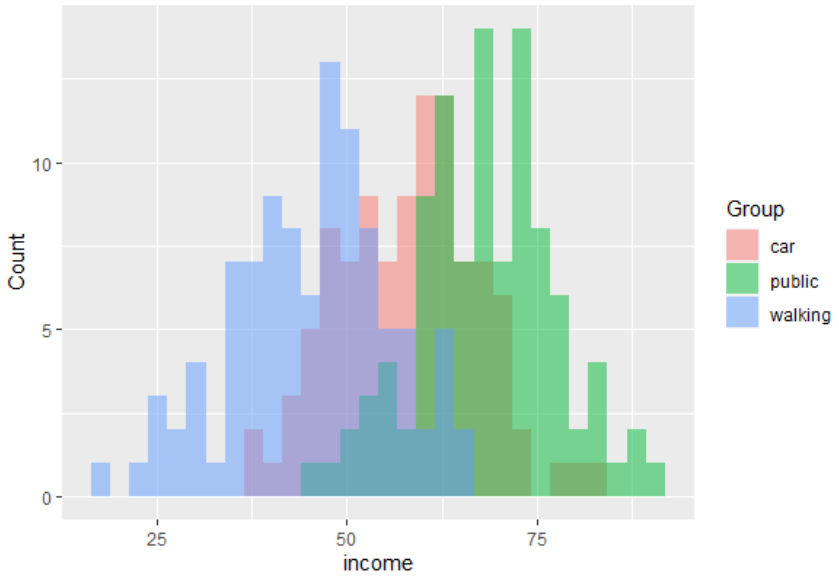
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Codurile de mai sus produc graficul de dispersie prezentat în secțiunea introductivă a celui de-al treilea document. Alternativ, am putea reprezenta datele ca o serie de histograme ca:

```
#histograms for income
ggplot(transport, aes(x = income, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "income", y = "Count", fill = "Group")

#or
ldahist(data = transport$income, g = transport$group)
```

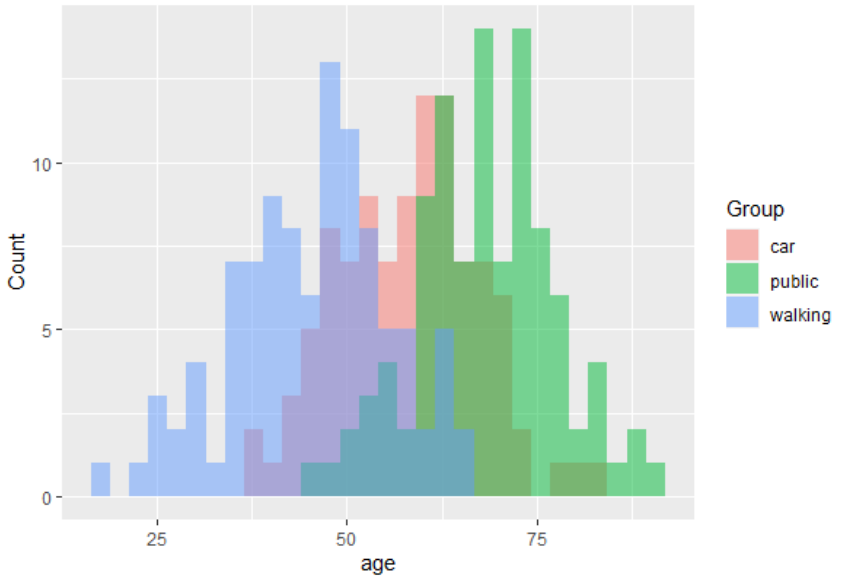
Prin rularea oricăreia dintre aceste două linii, putem avea o idee despre modul în care modul de transport se distribuie între valorile legate de vârstă și venit. De exemplu:



Or:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



LDA se realizează pur și simplu rulând:

```
#######################################
### Case Classification ###
#######################################
# Run the LDA using the lda function
output <- lda(group ~ ., transport)
output
```

Rezultatele clasice arată mediile inițiale pe grupe, coeficienții din proiecțiile LD și proporția dintre varianța (urmă) pe care o explică fiecare coordonată LD:

```
Group means:
           age income
car      58.32  89.44
public   68.40  49.82
walking  45.52  52.89

Coefficients of linear discriminants:
               LD1        LD2
age     -0.1177011 0.08844338
income   0.1376988 0.02050334

Proportion of trace:
   LD1    LD2
0.8997 0.1003
```

Co-funded by the Erasmus+ Programme of the European Union

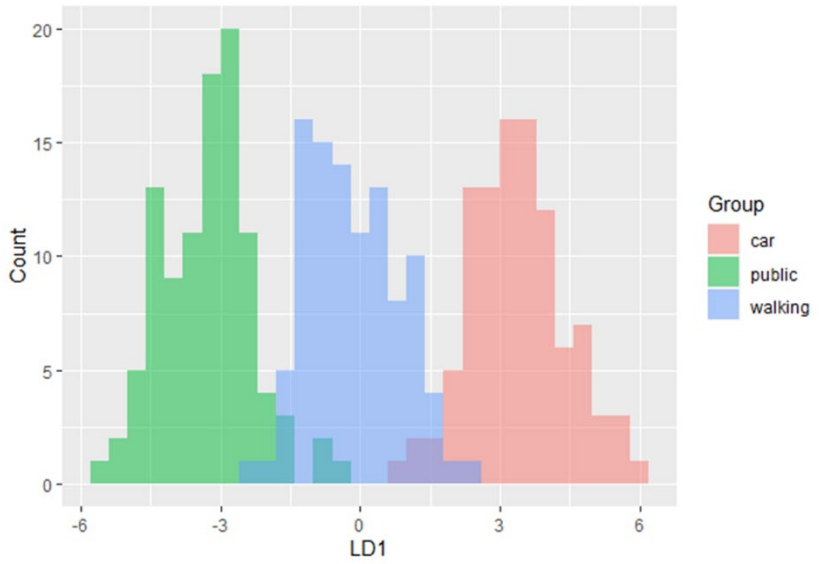
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

În exemplul nostru, prima coordonată LD este corelată pozitiv cu venitul și negativ cu vârsta și conține aproape 90% din variabilitatea dintre clase. A doua funcție LD prezintă o corelație pozitivă, dar mai slabă, cu ambele variabile și reprezintă doar aproximativ 10% din variabilitatea între variabile.

Noile coordonate sunt produse proiectând punctele de date originale cu coeficienții LDA prin expresia **uTX**. În aceste noi coordonate, observațiile sunt mai clar separate între grupuri. În exemplul nostru, avem două coordonate LD pentru fiecare individ, având în vedere vârsta și venitul acestuia. Coordonatele corespunzătoare primei funcție LD au puterea discriminantă mai mare. Putem vedea cu ușurință această putere discriminantă prin trasarea în R a unei histograme, punând acum primele coordonate LD în axa orizontală:

```
#histograms: first LDA
ggplot(lda.data, aes(x = LD1, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "LD1", y = "Count", fill = "Group")
```

Obținând:



Acest grafic arată modul în care cantitatea de suprapunere se reduce considerabil. Cu alte cuvinte, prima coordonată LD (rețineți că este un „compozit” care se corelează negativ cu vârsta și pozitiv cu venitul) discriminează în mod adecvat între categoriile de transport.

Co-funded by the Erasmus+ Programme of the European Union
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3. LDA predictivă

3.1 Procedura

LDA poate fi folosit nu numai în scopuri de clasificare (descriptive), ci și cu obiectivul de a prezice apartenența la clasă. De exemplu, să presupunem că avem date despre vârsta și venitul anual al gospodăriei pentru o persoană (în eșantion sau în afara eșantionului) și am dori să anticipăm modul de transport pe care este cel mai probabil să îl folosească această persoană. LDA poate fi de ajutor pentru a ne oferi o predicție, într-un mod similar cu modelele multinominale logit sau probit.

În acest scop predictiv, sunt necesare câteva ipoteze:

- - grupurile sunt multivariate normale
- - variațe-covariațe egale între grupuri

Formularea LDA predictivă este legată de formularea teoremei lui Bayes pentru actualizarea probabilităților: Fie g numărul de grupuri și qi probabilitatea anterioară (frecvențele relative de obicei observate) pentru grupul i. Fie $\mathbf{x}$ un vector de observații ale covariatelor pentru un individ. Probabilitatea (posterior) de a face parte din grupul Gi condiționată de $\mathbf{x}$, P(Gi $|\mathbf{x}$), poate fi exprimată ca:

$$\mathrm{P}(G_i|\mathbf{x}) = \frac{q_i \mathrm{P}(\mathbf{x}|G_i)}{\sum_{j=1}^{g} q_j \, \mathrm{P}(\mathbf{x}|G_j)}$$

Aceasta este o abordare bayesiană care actualizează probabilitățile anterioare qi pe baza probabilităților condiționate P($\mathbf{x}|Gi$). În ipotezele de normalitate:

$$\mathrm{P}(\mathbf{x}|G_i) = (2\pi)^{(-p/2)} |\mathbf{W}|^{(-1/2)} e^{(-D_i^2/2)}$$

unde $|\mathbf{W}|$ este determinantul matricei de varianță în cadrul clasei și D_i este $D_i \; = (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})^T \mathbf{W}^{-1} \; (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})$. Conectand expresia de $\mathrm{P}(\mathbf{x}|G_i)$ în formula pentru $\mathrm{P}(G_i|\mathbf{x})$, avem:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$$P(G_i|\mathbf{x}) = \frac{q_i e^{(-D_i^2/2)}}{\sum_{j=1}^{g} q_j \, e^{\left(-D_j^2/2\right)}}$$

3.2. Exemplu folosind R

Rutina LDA în R poate produce probabilități posterioare pe baza ipotezelor și a formulării detaliate mai înainte. Funcțiile LDA permit prezicerea celui mai probabil membru al clasei pentru orice individ, având în vedere un vector de covarianță (vârsta și venitul gospodăriei în exemplu).

Ca o ilustrare, tabelul afișat mai jos arată probabilitățile prezise pentru fiecare grup pentru un subset de indivizi din eșantion. Se presupune că ponderile qi sunt identice pentru fiecare dintre cele trei moduri de transport ($q_i = 1/3$).

group	income	age	LD1	LD2	predclass	pred_car	pred_public	pred_walk
walking	26	47	1.349620208	-3.127883266	walking	1.983231e-03	2.965401e-07	9.980165e-01
walking	27	51	1.782714373	-2.957426532	walking	1.245493e-02	1.063176e-07	9.875450e-01
walking	28	35	-0.538167997	-3.197036576	walking	2.241897e-06	9.299867e-05	9.999048e-01
walking	29	34	-0.793567966	-3.129096536	walking	1.034985e-06	2.354290e-04	9.997635e-01
walking	30	45	0.603417987	-2.815116429	walking	2.575777e-04	5.608833e-06	9.997368e-01
walking	31	35	-0.891271423	-2.931706440	walking	1.062902e-06	4.699394e-04	9.995290e-01
walking	31	43	0.210319191	-2.767679728	walking	7.042531e-05	2.095366e-05	9.999086e-01
walking	32	42	-0.045080777	-2.699739689	walking	3.251705e-05	5.305263e-05	9.999144e-01
walking	34	45	0.132613419	-2.461342914	walking	9.528279e-05	4.866634e-05	9.998561e-01
walking	37	37	-1.322080621	-2.360039490	walking	6.786660e-07	5.490035e-03	9.945093e-01
walking	56	60	-0.391329301	-0.208038498	walking	1.019914e-03	2.021248e-02	9.787676e-01
walking	54	48	-1.808312938	-0.630965323	walking	1.821514e-06	4.269625e-01	5.730357e-01
walking	63	78	1.263341587	0.780125254	car	6.956557e-01	2.513904e-04	3.040929e-01
public	69	56	-2.4722395	0.859712069	public	4.567507e-08	9.909603e-01	9.039690e-03
public	87	70	-2.6630764	2.738739632	public	1.118083e-08	9.998736e-01	1.264240e-04
public	73	50	-3.7692370	1.090465551	public	8.209481e-12	9.998980e-01	1.019855e-04
public	46	33	-2.9321862	-1.646062437	public	2.043553e-09	7.715710e-01	2.284290e-01
public	62	64	-0.5467408	0.404635130	walking	1.713491e-03	1.000701e-01	8.982164e-01
public	68	42	-4.2823219	0.484221945	public	2.851843e-13	9.999323e-01	6.768416e-05
public	50	31	-3.6783884	-1.333295600	public	1.790602e-11	9.845625e-01	1.543752e-02
public	71	36	-5.4616183	0.626532048	public	1.118031e-16	9.999987e-01	1.298905e-06
public	56	43	-2.7322094	-0.556595260	public	8.609396e-09	9.387842e-01	6.121583e-02
public	60	45	-2.9276163	-0.161815068	public	2.388442e-09	9.839053e-01	1.609473e-02

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	Clasa prezisă corespunde celui mai mare P(Gi |x) pentru fiecare individ. Acestea sunt calculate prin aplicarea următoarei rutine în Rstudio:
	`#####################################`
	`### Predicting classifications ###`
	`#####################################`
	
	`# Get the posterior values and predicted classification for each case`
	`pred <- predict(output)`
	`# Posterior values for each class for each case`
	`posteriors <- pred$posterior`
	
	`# Predicted Class`
	`predclass <- pred$class`
	`# Putting Data (including actual class) next to predicted class and posterior values`
	`post_transport <- cbind(lda.data,predclass,posteriors)`
	`colnames(post_transport) <- c("group","income","age","LD1","LD2","predclass",`
	`"pred_car","pred_public","pred_walk")`
	
	În cele mai multe cazuri, LDA prezice corect grupul căruia îi aparține fiecare individ. Există însă unele cazuri pentru care LDA nu prezice corect. Aceste cazuri corespund observațiilor suprapuse care rămân încă în clasificarea LDA

Autoevaluare (interogări și răspunsuri cu variante multiple)	LDA este o tehnică statistică care permite: a) Clasificarea datelor pe grupe b) Prezicerea apartenenței la clasă **c) Ambele răspunsuri sunt adevărate** Ipotezele necesare pentru aplicarea LDA în scopuri predictive sunt: a) Normalitate multivariată între grupuri b) Variante-covariante între grupuri egale **c) Ambele răspunsuri sunt adevărate** LDA se bazează pe maximizarea raportului: **a) Variabilitatea între grupuri versus variabilitatea în cadrul clasei** b) Variabilitatea în cadrul grupurilor versus variabilitatea totală c) Variabilitatea totală versus variabilitatea în interiorul grupurilor
Resurse (videoclipuri, link de referință)	
Material aferent	
PPT conexe	
Bibliografie	Boedeker, P., & Kearns, N. T. (2019). Linear discriminant analysis for prediction of group membership: A user-friendly primer. Advances in Methods and Practices in Psychological Science, 2, 250-263.
Furnizat de	[Uniovi]

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union