https://datascience-project.eu/

Training Fiche Template

Titolo	ANALISI DISCRIMINANTE LINEARE
Parole chiave	Analisi discriminante, classificazione, R, analisi bayesiana
Lingua	Inglese
Obiettivi / Risultati di apprendimento	**L'obiettivo di questo modulo è introdurre e spiegare le basi dell'analisi discriminante lineare (LDA).** **Alla fine di questo modulo sarai in grado di:** **1. Identificare le situazioni in cui LDA può essere utile** **2. Calcolare le funzioni LDA** **3. Interpretare i risultati prodotti da LDA descrittivo e predittivo**

Training course:	
Data Science Literacy	
Data Visualisation and Visual Analytics Module	X
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	

Descrizione	In questo modulo di formazione verrai introdotto all'uso dell'analisi discriminante lineare (LDA). LDA è un metodo per trovare combinazioni lineari di variabili che meglio separa le osservazioni in gruppi o classi, ed è stato originariamente sviluppato da Fisher (1936). Questo metodo massimizza il rapporto tra la varianza tra classi e la varianza all'interno della classe in un particolare set di dati. In questo modo, la variabilità tra gruppi è massimizzata, il che si traduce in massima separabilità. LDA può essere utilizzato con scopi puramente di classificazione, ma anche con obiettivi predittivi.
Contenuti suddivisi in tre punti	**INTRODUZIONE: SPIEGAZIONE CON UN ESEMPIO ILLUSTRATIVO**

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.

Supponiamo di avere un campione di individui e di osservare la modalità di trasporto (in auto, con i mezzi pubblici o a piedi) che di solito impiegano per spostarsi all'interno di una città. Sappiamo che la scelta della modalità di trasporto è parzialmente influenzata dallo status economico, inoltre osserviamo i dati sulla loro età e il loro reddito annuo familiare, insieme al mezzo di trasporto scelto:

Vogliamo sapere come queste due covariate aiutano a classificare (cioè discriminare) gli individui assegnandoli a una specifica categoria di modalità di trasporto. Possiamo vedere che non esiste una classificazione perfetta: gli individui con reddito elevato tendono a usare le auto più frequentemente, ma c'è una grande sovrapposizione di categorie "a piedi" e "trasporto pubblico" per coloro che hanno redditi più bassi. Inoltre, c'è una maggiore sovrapposizione tra le categorie per quanto riguarda la loro distribuzione per età: gli individui più anziani non camminano. L'età, però, non è un buon predittore della modalità di trasporto. Questo è il tipico problema che LDA affronta.

2. LDA per la classificazione

2.1. Formulazione

Le funzioni LDA possono essere utilizzate per aiutare con la classificazione dei dati sulla base di una matrice di covariate X. Analogamente all' Analisi in Componenti Principali (PCA), le funzioni LDA mirano a trovare una combinazione lineare dei dati originali come:

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.

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

dove la varianza tra classi ($\mathbf{B}$) è massimizzata rispetto alla varianza all'interno della classe ($\mathbf{W}$), che può essere affrontata come un problema generalizzato agli autovalori:

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

Le coordinate discriminanti sono ottenute dagli autovettori di $\mathbf{W}$^($-\mathbf{1}$) $\mathbf{B}$.

2.2. Un esempio

Come esempio illustrativo, risolviamo il problema della classificazione della modalità di trasporto in base all'età e al reddito attraverso la tecnica LDA in R, utilizzando la funzione "lda" all'interno della libreria "mass". Per tutte le analisi qui presentate, dovremo installare e caricare i seguenti pacchetti 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)
```

I dati sono in formato csv (chiamato "trasnpor_example"), che può essere importato in R eseguendo i seguenti comandi:

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

Per avere una prima descrizione dei dati, possiamo tracciare il campione sotto forma di grafico a dispersione:

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

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

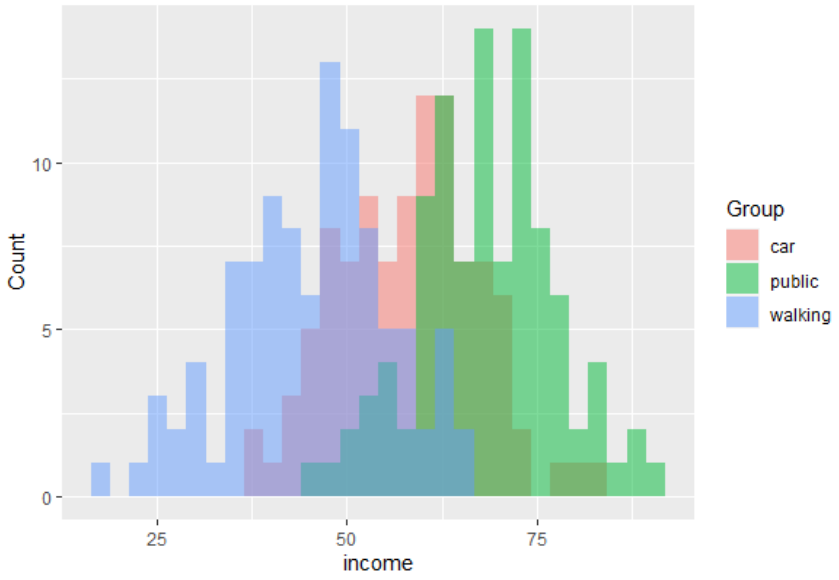
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Le righe di codice precedenti producono il grafico a dispersione mostrato nella sezione introduttiva del documento. In alternativa, potremmo tracciare i dati come una serie di istogrammi:

```
#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)
```

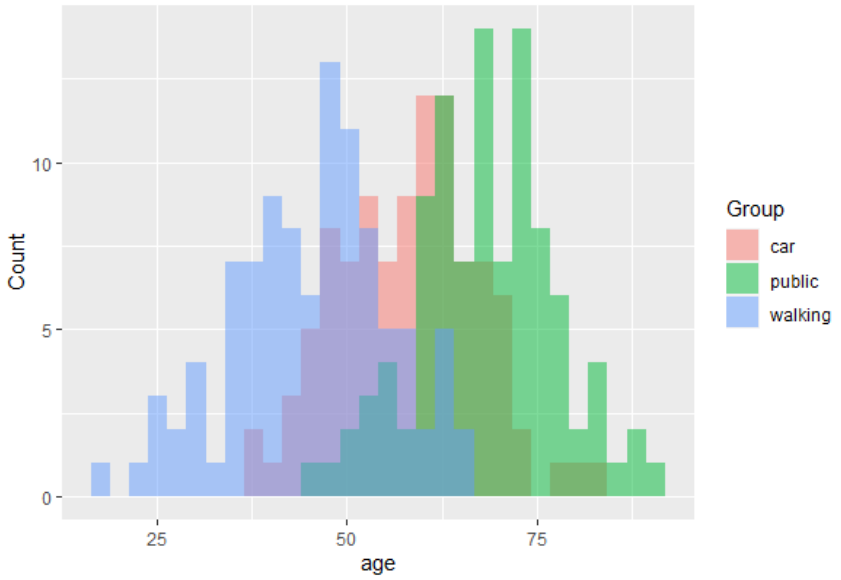
Eseguendo uno di questi comandi, possiamo avere un'idea di come la modalità di trasporto si distribuisce tra i valori dell'età e del reddito. Per esempio:



Oppure:

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 si ottiene eseguendo:

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

L'output tipico mostra le medie iniziali per gruppo, i coefficienti nelle proiezioni LD e la proporzione della varianza between che ogni coordinata LD spiega:

```
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

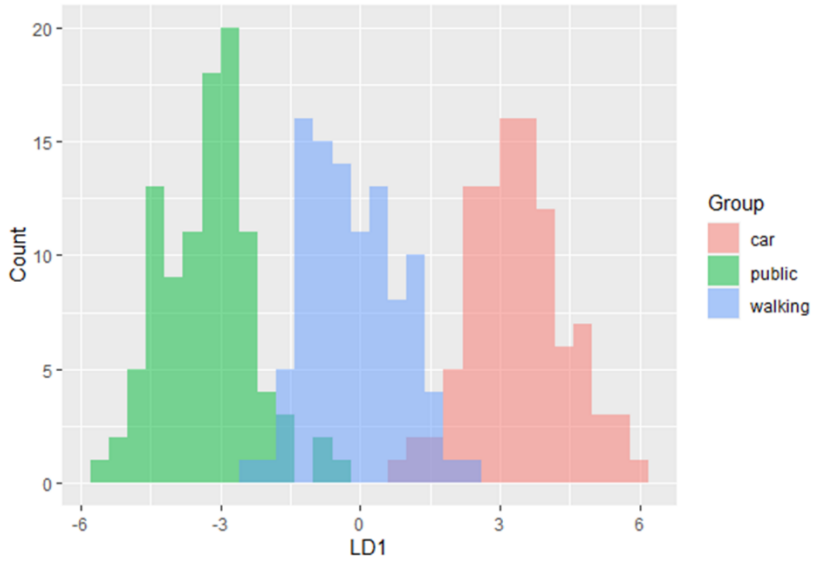
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Nel nostro esempio, la prima coordinata LD è correlata positivamente con il reddito e negativamente con l'età e contiene quasi il 90% della variabilità inte-class. La seconda funzione LD mostra una correlazione positiva, ma più debole con entrambe le variabili e rappresenta solo circa il 10% della variabilità between.

Le nuove coordinate vengono prodotte proiettando i punti originali con i coefficienti LDA, utilizzando l'espressione $\mathbf{u}^T\mathbf{X}$.. In queste nuove coordinate, le osservazioni sono separate più chiaramente tra i gruppi. Nel nostro esempio, abbiamo due coordinate LD per ogni individuo, data la loro età e reddito. Le coordinate corrispondenti alla prima funzione LD hanno la potenza discriminante maggiore. Possiamo facilmente vedere questa potenza discriminante tracciando in R un istogramma, mettendo ora le prime coordinate LD nell'asse orizzontale:

```
#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")
```

Obtaining:



Questo grafico mostra come la quantità di sovrapposizioni diminuisca considerevolmente. In altre parole, la prima coordinata LD (ricordiamo che è un "composito" che correla negativamente con l'età e

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.

positivamente con il reddito) discrimina adeguatamente tra le categorie di trasporto.

3. LDA predittivo

3.1 La procedura

LDA può essere utilizzato non solo per scopi di classificazione (descrittiva), ma anche con l'obiettivo di prevedere l'appartenenza alla classe. Ad esempio, supponiamo di avere dati sull'età e sul reddito familiare annuale per un individuo (in-sample o out-of-sample) e vorremmo prevedere la modalità di trasporto che questa persona è più probabile che utilizzi. LDA può essere utile per fornirci una previsione, in modo simile ai modelli logit o probit multinominali.

Per questo scopo predittivo, sono necessarie alcune ipotesi:

1. i gruppi sono multivariati
2. Varianze-covarianze uguali tra gruppi

La formulazione dell'LDA preventiva è correlata alla formulazione del teorema di Bayes per l'aggiornamento delle probabilità:

Sia $\mathbf{g}$ il numero di gruppi e q_i la probabilità precedente (frequenze relative) per il gruppo i. Sia $\mathbf{x}$ un vettore di osservazioni di covariate per un individuo. La probabilità (posteriore) di appartenere al gruppo G_i condizionata da $\mathbf{x}$, P(G_i |$\mathbf{x}$), può essere espressa come:

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

Questo è un approccio bayesiano che aggiorna le probabilità precedenti q_i basandosi sulle probabilità condizionali P(G_i|$\mathbf{x}$), Sotto le ipotesi di normalità:

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

dove|$\mathbf{W}$| è il determinante della matrice di varianza all'interno della classe e D_i è $D_i^2 = (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})^T \mathbf{W}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})$. Con l'inserimento dell'espressione di $\mathrm{P}(\mathbf{x}|G_i)$ nella formula per $\mathrm{P}(G_i|\mathbf{x})$, otteniamo:

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.

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

3.2. Un esempio con R

La routine LDA in R può produrre probabilità posteriori basate sulle ipotesi e sulla formulazione descritta in precedenza. Le funzioni LDA consentono di prevedere l'appartenenza alla classe più probabile per ogni individuo, dato un vettore di covariate (età e reddito familiare nell'esempio).

A titolo illustrativo, la tabella visualizzata di seguito mostra le probabilità previste per ciascun gruppo per un sottoinsieme di individui nel campione. A priori si assume che q_i siano identici per ciascuna delle tre modalità di trasporto ($q_i = 1/3$).

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

La classe prevista corrisponde alla più alta $P(G_i|\mathbf{x})$ per individuo. Sono calcolati applicando la seguente routine in Rstudio:

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.

	`` ##################################### ### 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") ``` Nella maggior parte dei casi, LDA prevede correttamente il gruppo a cui appartiene ogni individuo. Ci sono alcuni casi, tuttavia, per i quali LDA non riesce a fare una previsione corretta. Questi casi corrispondono alle osservazioni sovrapposte che rimangono ancora nella classificazione LDA.
Autovalutazione (domande a scelta multipla)	LDA è una tecnica statistica che consente la: 1. Classificazione dei dati in gruppi 2. Previsione dell'appartenenza alla classe 3. Entrambe le risposte sono vere Le ipotesi richieste per applicare LDA, per fini predittivi, sono: 1. Normalità multivariata tra i gruppi 2. Varianze-covarianze uguali tra gruppi 3. Entrambe le risposte sono vere LDA si basa sulla massimizzazione del rapporto: 1. Tra gruppi e variabilità all'interno della classe 2. All'interno dei gruppi rispetto alla variabilità totale 3. Variabilità totale rispetto a quella all'interno dei gruppi
Resources (videos, reference link)	
Related material	
Related PPT	
Bibliografia	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.
Provided by	[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.