Capitolo 1

Analisi Discriminante

1.1 Introduzione

L'analisi discriminante viene condotta per definire una modalità di assegnazione dei casi a differenti gruppi, in funzione di una serie di variabili fra di loro correlata.

i gruppi sono già definiti al momento dell'analisi, pertanto l'interesse è rivolto a definire un modello che consenta di asseganre un nuovo caso ad un gruppo predefinito, in funzione di un certo numero di variabili.

L'analsi discriminante fu introdotta da R. A. Fisher nel 1936, che l'utilizzò per attribuire dei reperti fossili ad uno di due gruppi: scimmie e umanoidi.

Il file che utilizzeremo di seguito contiene i famosi dati sugli c raccolti da Fisher, e utilizzati per effettuare tests su diversi algoritmi di classificazione.

1.2 Un'analisi discriminante

1.2.1 Descrizione del dataset

Il dataset[1] consta di 50 osservazioni su ciascuna di tre specie di Iris:

1. iris setosa,

[1] Il file è reperibile nella sezione 'Materiale Didattico' sul sito del corso di laurea in Statistica (www.economia.unical.it/statistica/)

2. iris versicolor,

3. iris virginica.

le variabili osservate su tale dataset sono quattro:

- Lunghezza del Sepalo,
- Larghezza del Sepalo,
- Lunghezza del Petalo,
- Larghezza del Petalo.

L'applicazione consiste nel valutare se le variabili considerate possono considerarsi discriminanti per i tre tipi di iris considerati.

1.2.2 La procedura in SPSS

Prima di avviare la procedura che consente di effettuare l'anlisi discriminante, opportuno rappresentare graficamente i dati contunute nel dataset e verificare se le specie considerate sono separate. Ricorriamo al diagramma a dispersione (scatter plots): selezionare dal menù *Grafici→A dipsersione*, Figura 1.1.

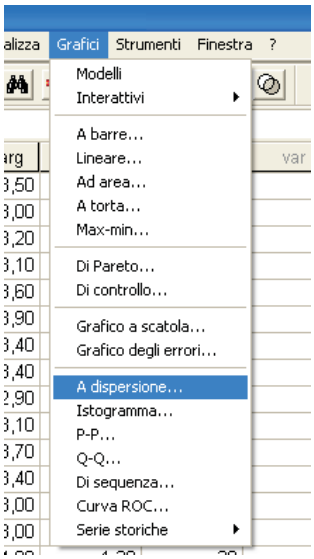

Figura 1.1: Grafico a dispersione

scelgiere *Matrice* e cliccare su *Definisci*.

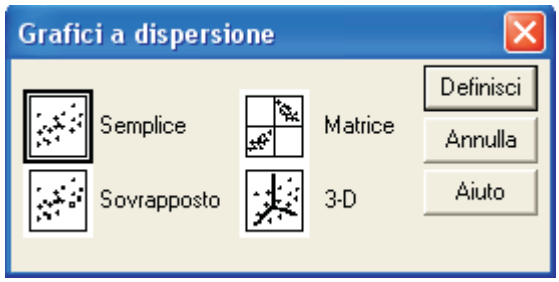

Figura 1.2: Selezione del grafico a dispersione in matrice

A questo punto selezionare la variabile *specie* per distinguere i casi, mentre le restanti variabili comporranno la matrice (Figura 1.3), infine cliccare su *OK*.

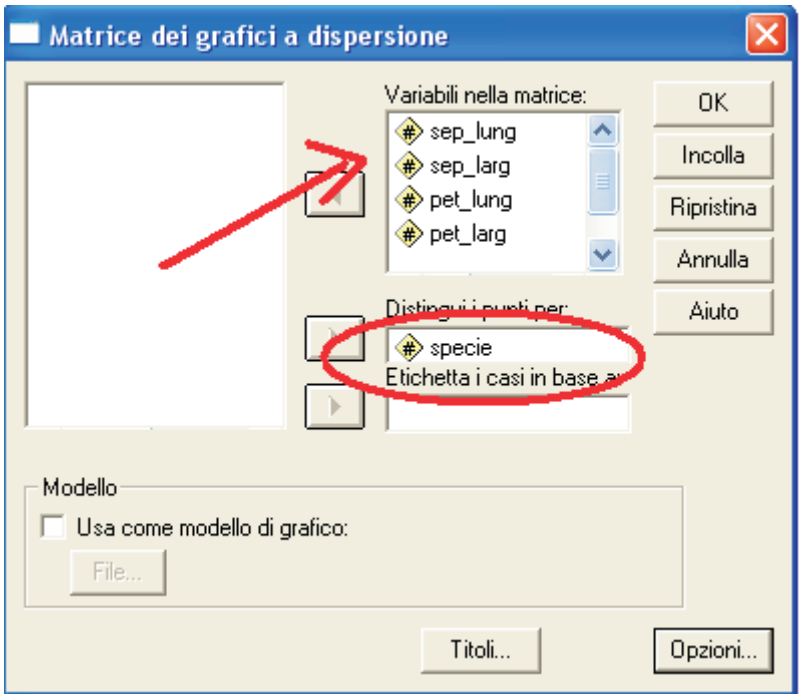

Figura 1.3: Selezione delle variabili per il grafico

Osservando il grafico a dispersione prodotto, si nota come la specie *setosa* risulta nettamente separata, mentre per le altre due specie si notano delle sovrapposizioni.

Le variabili considerate consentono, dunque, di distinguere la specie setosa dalle altre, e non distinguono in maniera univoca le altre due specie.

Si può pensare di procedere all'analisi discrimnante su tali variabili.

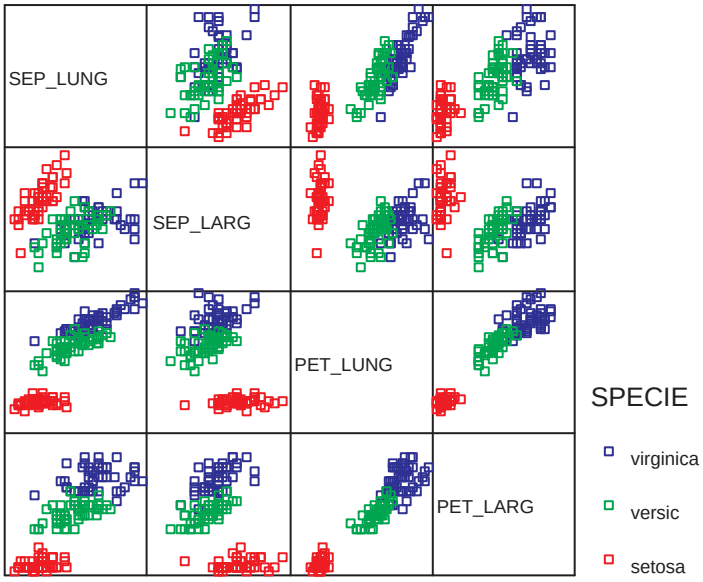

Figura 1.4: Il grafico a dispersione

Selezionare dal menù *Analizza→Classificazione→Discriminante*.

Al solito, è necessario definire le variabili che verranno processate nell'analisi, nel nostro caso la variabile di raggruppamento è *specie* e i valori minimo e massimo dell'intervallo sono 1 e 3 (vedi Figura 1.5), mentre le altre sono le variabili indipendenti.

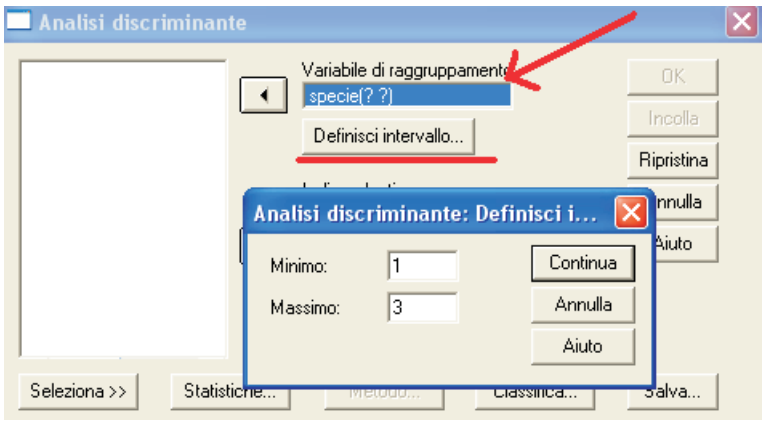

Figura 1.5: Analisi Discriminante: procedura

Dal sottomenù *Statistiche* selezionare:

- Medie
- ANOVA univariate
- correlazione entro gruppi
- covarianza entro gruppi

Dal sottomenù *Classifica* selezionare:

- (Come probabilità a priori), Tutti i gruppi uguali
- (Visualizza), Risultati per casi, Tabella riassuntiva, Classificazione autoesclusiva
- (Usa matrice di covarianza), Entro i gruppi
- (Grafici), Gruppi accorpati

Per eseguire la procedura cliccare su *OK*.

1.3 L'interpretazione dei risultati

L'analisi discriminante eseguita con SPSS restituisce un buon numero di tabelle che ora esaminaremo nel dettaglio. La prima tabella riguarda, semplicemente, le statistiche descrittive calcolate per ogni varaibile.

Dalla Figura 1.6, notiamo che i casi processati nell'analisi sono 50 per ogni specie, se fosse stato presente un dato mancante per una o più variabili, SPSS avrebbe escluso quel caso dall'analisi.

Osservando la tabella, è facile descrivere le differenze dei valori medi calcolati per ogni variabile nei tre gruppi di casi considerati. Infatti, notiamo subito che la lunghezza media del petalo risulta essere quella che presenta maggior differenza nei tre gruppi: si passa da 1,46 nella specie setosa a 4,26 per la specie versic e infien a 5,5 per virgicolor.

Si lasciano allo studente le considerazioni sulla standard deviation.

Statistiche di gruppo

			Media	Deviazione std.	Validi (listwise)	
					Non pesati	Pesati
SPECIE	setosa	SEP_LUNG	5,01	,35	50	50
		SEP_LARG	3,43	,38	50	50
		PET_LUNG	1,46	,17	50	50
		PET_LARG	,25	,11	50	50
	versic	SEP_LUNG	5,94	,52	50	50
		SEP_LARG	2,77	,31	50	50
		PET_LUNG	4,26	,47	50	50
		PET_LARG	1,33	,20	50	50
	virginica	SEP_LUNG	6,59	,64	50	50
		SEP_LARG	2,97	,32	50	50
		PET_LUNG	5,55	,55	50	50
		PET_LARG	2,03	,27	50	50
	Totali	SEP_LUNG	5,84	,83	150	150
		SEP_LARG	3,06	,44	150	150
		PET_LUNG	3,76	1,77	150	150
		PET_LARG	1,20	,76	150	150

Figura 1.6: Analisi Discriminante: Le statistiche descrittive

Dai valori medi visti in precedenza possiamo cominciare a intuire che la variabile *lunghezza del petalo* è quella che presenta un più alto potere discriminnate, rispetto alle altre variabili. Tuttavia essa riesce a distinguere meglio la specie setosa e un meno le altre specie (vedi anche Figura 1.4).

La seconda tabella che SPSS restituisce riguarda due test sull'uguaglianza delle medie nei gruppi, e si compone sulle righe le variabili e sulle colonne le statistiche test calcolate.

In particolare la prima colonna restituisce il valore della **Lambda di Wilks** che consiste in un test multivariato per valutare l'ipotesi che due o più gruppi provengano da popolazioni con la stessa media, detta anche statistica U. **Lambda** varia tra 0 ed 1. Valori elevati indicano che le medie di gruppo non sono molto differenti (è pari a 1 se le medie sono tutte uguali). Valori bassi indicano differenze nelle medie di gruppo.

I risultati del secondo test sono posti, nella seconda e ultima colonna. I vaolri provengono dall'ANOVA calcolata su ogni singola variabile. Il valore di **F** e la significatività ci danno ulteriori informazioni sulla diffrenza tra i gruppi.

Quando **F** alto e il livello di significatività è sufficientemente basso (tra 0,01 e 0,05),

l'ipotesi nulla può essere rifiutata. In altre parole, un basso livello di significatività indica che i risultati osservati nel campione non dipendono dal caso e che quindi è lecito attendersi che si ripetano uguali nella popolazione.

Test di uguaglianza delle medie di gruppo

	Lambda di Wilks	F	df1	df2	Sig.
SEP_LUNG	,381	119,265	2	147	,000
SEP_LARG	,599	49,160	2	147	,000
PET_LUNG	,059	1180,161	2	147	,000
PET_LARG	,071	960,007	2	147	,000

Figura 1.7: Analisi Discriminante: Test di uguaglianza delle medie

Dall'osservazione delle statistiche descrittive e dai test sull'uguaglianza delle medie si possono ordinare le variabili in ordine del loro potere discriminante:

1. lunghezza del petalo,
2. larghezza del petalo,
3. lunghezza del sepalo,
4. larghezza del sepalo.

Di seguito, si riportano le matrici di covarianza e correlazione entro i gruppi, ossia calcolate mediando le singole matrici di covarianza di tutti i gruppi.

Matrici globali entro gruppi[a]

		SEP_LUNG	SEP_LARG	PET_LUNG	PET_LARG
Covarianza	SEP_LUNG	,265	,093	,168	,038
	SEP_LARG	,093	,115	,055	,033
	PET_LUNG	,168	,055	,185	,043
	PET_LARG	,038	,033	,043	,042
Correlazione	SEP_LUNG	1,000	,530	,756	,365
	SEP_LARG	,530	1,000	,378	,471
	PET_LUNG	,756	,378	1,000	,484
	PET_LARG	,365	,471	,484	1,000

a. La matrice di covarianza ha 147 gradi di libertà.

Figura 1.8: Analisi Discriminante: Matrice globale entro i gruppi

Dopo queste prime analisi di tipo esplorativo su dati, SPSS mostra i risultati dell'analisi discriminane.

Dalla tabella degli autovalori (dove l'autovalore è il rapporto fra la somma dei quadrati fra gruppi e la somma dei quadrati entro gruppi) si vede che le funzioni discriminanti 'estratte' sono due e in particolare la prima presente un'autovalore molto alto che spiega da solo il $99,1\%$ della varianza totale. L'autovalore più alto corrisponde all'autovettore nella direzione della massima dispersione delle medie di gruppo.

Autovalori

		Autovalore	% di varianza	% cumulata	Correlazione canonica
Funzione	1	32,192[a]	99,1	99,1	,985
	2	,285[a]	,9	100,0	,471

a. Per l'analisi sono state usate le prime 2 funzioni discriminanti canoniche.

Figura 1.9: Analisi Discriminante: Gli autovalori

Lambda di Wilks è una misura del potere discriminantte di ogni funzione, ossia quanto bene le funzioni riescono a separare i casi in gruppi. Bassi valori di Lambda indicano un grande potere discriminante delle funzioni.

Nel caso in esame Lambda presenta un valore bassissimo (0,23) e congiuntamente il valore del Chi-quadrato uguale a 546,115, conferma un alta e significativa differenza tra i gruppi.

Mentre la seconda funzione, da sola, non presente lo stesso potere discriminante.

Lambda di Wilks

		Lambda di Wilks	Chi-quadrato	df	Sig.
Test di funzion	Da 1 a 2	,023	546,115	8	,000
	2	,778	36,530	3	,000

Figura 1.10: Analisi Discriminante: Lambda di Wilks

I coefficienti standardizzati non sono altro che i coefficienti delle due funzioni discriminanti appena calcolate. Si nota il forte potere discriminnate della variabile *lunghezza del petalo*.

Coefficienti standardizzati della funzione discriminante canonica

	Funzione	
	1	2
SEP_LUNG	-,427	,012
SEP_LARG	-,521	,735
PET_LUNG	,947	-,401
PET_LARG	,575	,581

Figura 1.11: Analisi Discriminante: I coefficienti

Questi coefficienti ci serviranno per scrivere le finzioni discriminanti.

La matrice di struttura mostra la correlazione esistente tra ogni variabile discriminante e le funzioni discrimanti determinante, in ordine della loro dimensione assolute della correlazione. L'asterisco evidenzia la correlazione assoluta più grande tra ciascuna variabile e qualsiasi funzione discriminante.

Matrice di struttura

	Funzione	
	1	2
PET_LUNG	,706*	,168
SEP_LARG	-,119	,864*
PET_LARG	,633	,737*
SEP_LUNG	,223	,311*

Figura 1.12: Analisi Discriminante: La matrice struttura

Le Figure 1.13 e 1.14 mostrano i punteggi (scores) delle funzioni discriminanti calcolate prima alle medie dei gruppi (1.13) e poi per ogni caso presente nel dataset (1.14).

In particolare, in quest'ultima tabella si può confrontare per ogni caso il gruppo di appartenza e il gruppo previsto dal modello, (=dall'analisi discriminante effettuata) e fornisce ulteriori informazioni sulla classificazione:

1. individua il 'gruppo più alto' di appartenenza di ogni caso,
2. fornisce la probalilità del punteggio osservato, assumendo l'appartenenza del caso al gruppo più probabile,
3. la probabilità posteriori di appartenenza di un caso al gruppo stimato,
4. la distanza di Mahalanobis di ogni caso dalla media del proprio gruppo; valori altri di questa distanza identificano un outlier,
5. i punteggi delle funzioni discriminati per ogni caso.

Funzioni ai baricentri di gruppo

		Funzione	
		1	2
SPECIE	setosa	-7,608	,215
	versic	1,825	-,728
	virginica	5,783	,513

Funzioni discriminanti canoniche non standardizzate valutate alle medie di gru

Figura 1.13: Analisi Discriminante: Le funzioni ai baricentri

Statistiche per casi

	Gruppo di appartenenza	Gruppo più alto: Gruppo previsto	Gruppo più alto: P(D=d | G=g) p	Gruppo più alto: P(D=d | G=g) df	Gruppo più alto: P(G=g | D=d)	Gruppo più alto: Distanza quadratica di Mahalanobis dal baricentro	Secondo gruppo più probabile: Gruppo	Secondo gruppo più probabile: P(G=g | D=d)	Secondo gruppo più probabile: Distanza quadratica di Mahalanobis dal baricentro	Punteggi discriminanti: Funzione 1	Punteggi discriminanti: Funzione 2
	1	1	,899	2	1,000	,214	2	,000	98,807	-8,062	,300
	1	1	,540	2	1,000	1,233	2	,000	80,173	-7,129	-,787
	1	1	,885	2	1,000	,245	2	,000	86,981	-7,490	-,265
	1	1	,493	2	1,000	1,416	2	,000	74,623	-6,813	-,671
	1	1	,833	2	1,000	,365	2	,000	100,692	-8,132	,514

Figura 1.14: Analisi Discriminante: Statistiche dei casi

Infine, il grafico 1.15 e la tabella 1.16 riassumo in sintesi i risultati della classificazione.

Il grafico mostra come le due funzioni discriminanti riescono a separare più nettamente i gruppi rispetto alle variabili originarie (si confronti con il grafico 1.4).

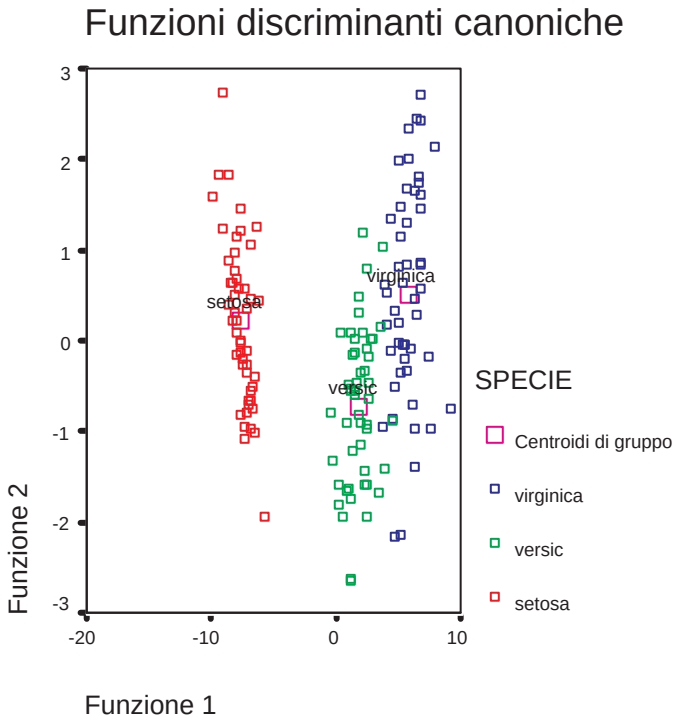

Figura 1.15: Analisi Discriminante: Il grafico delle funzioni discriminanti

Mentre la tabella mostra i risulatati della classicazione.

Risultati della classificazione[b,c]

				Gruppo di appartenenza previsto			
				setosa	versic	virginica	Totali
Originale	Conteggio	SPECIE	setosa	50	0	0	50
			versic	0	48	2	50
			virginica	0	1	49	50
	%	SPECIE	setosa	100,0	,0	,0	100,0
			versic	,0	96,0	4,0	100,0
			virginica	,0	2,0	98,0	100,0
Cross-validat[a]	Conteggio	SPECIE	setosa	50	0	0	50
			versic	0	48	2	50
			virginica	0	1	49	50
	%	SPECIE	setosa	100,0	,0	,0	100,0
			versic	,0	96,0	4,0	100,0
			virginica	,0	2,0	98,0	100,0

a. La cross-validazione viene eseguita solo per tali casi nell'analisi. Per la cross-validazione, ogni ca viene classificato dalle funzioni derivate da tutti gli altri casi.

b. 98,0% di casi raggruppati originali classificati correttamente.

c. 98,0% di casi raggruppati cross-validati classificati correttamente.

Figura 1.16: Analisi Discriminante: I risultati