7. ANALISI DISCRIMINANTE

7.1 Introduzione

Fra i primi studiosi che si occupò di analisi discriminante (AD) va ricordato Fisher, che la utilizzò per attribuire alcuni reperti fossili alla categoria dei primati o a quella degli umanoidi sulla base di diverse misurazioni effettuate sui reperti stessi.

In generale, l'obiettivo dell'AD consiste nell'individuare un insieme di variabili in grado di **discriminare** nel modo migliore le *n* unità statistiche tra due o più gruppi predefiniti, che risultano noti a priori. Generalmente si perviene ad una regola di decisione (detta **regola di classificazione**), funzione di un numero limitato di variabili (quantitative o qualitative), tramite la quale si attribuisce ciascuna delle unità a uno di questi gruppi predefiniti.

L'obiettivo dell'AD è comune a molti campi di ricerca:

- in botanica si può avere la necessità di determinare la specie a cui appartiene una pianta,
- in medicina si può voler individuare la malattia da cui è affetto un individuo sulla base dei sintomi che presenta,
- nelle operazioni bancarie si può voler decidere se un'azienda che richiede un prestito risulti affidabile o meno, sulla base delle sue caratteristiche contabili e di bilancio.

L'AD viene comunemente utilizzata anche per

- prevedere il rischio di valanghe o la presenza di particolari giacimenti nelle ricerche minerarie,
- per individuare quali bambini possono avere difficoltà nell'imparare a leggere e a scrivere,
- per poter consigliare l'indirizzo di studio agli studenti in base ai risultati di test psico-attitudinali.

L'AD tratta insiemi di osservazioni in cui:

• una variabile definisce i gruppi (o le classi di osservazioni)

• sono presenti una o più variabili quantitative o qualitative (utilizzate per effettuare la discriminazione fra gruppi)

Considerato un vettore $\boldsymbol{X}$ di variabili (qualitative o quantitative) ritenute utili per la classificazione, supponiamo che una unità statistica A, caratterizzata da una data determinazione $\boldsymbol{x}$ di $\boldsymbol{X}$, possa appartenere ad uno qualsiasi di g gruppi distinti $C_1, C_2, ..., C_g$ che costituiscono la popolazione.

L'AD si occupa di individuare la regola ottimale in base alla quale assegnare l'unità a uno dei g gruppi sulla base delle determinazioni $\boldsymbol{x}$ di $\boldsymbol{X}$.

Qualsiasi regola di decisione comporterà il rischio di attribuire l'individuo ad un gruppo errato, ossia comporterà il rischio di una **classificazione errata**, per cui il problema consisterà nell'individuare quella particolare regola che minimizzi questo rischio.

Indicata con N_j ($j = 1, 2, \ldots, g$), la numerosità del j-esimo gruppo, la probabilità iniziale che una unità A scelta in modo casuale appartenga al j-esimo gruppo potrebbe essere posta proporzionale alla numerosità del gruppo

$$P(A \in C_j) = \pi_j = N_j / N, \qquad j = 1, 2, ..., g.$$

dove

$$\sum_{j=1}^{g} N_j = N .$$

In questo caso le π_j, che rappresentano le cosiddette **probabilità a priori**, risultano proporzionali alle numerosità dei g gruppi, ma possono anche essere fissate sulla base di un qualunque criterio, anche soggettivo.

Le probabilità a priori vengono poi aggiornate tenendo presente i valori assunti dalle variabili ritenute utili per effettuare la classificazione, ottenendo le cosiddette **probabilità a posteriori**, che vengono utilizzate per decidere a quale gruppo vada assegnata ciascuna delle n osservazioni.

La migliore regola di assegnazione delle unità ad uno dei g gruppi C_j (ossia la regola che consente il minimo rischio di assegnazioni errate), consiste ovviamente nell'assegnare l'unità al gruppo per il quale risulta massima la corrispondente probabilità a posteriori.

Se l'individuo appartiene effettivamente al gruppo C_j si ha una **classificazione corretta**, mentre se appartiene al gruppo C_k (con $k \neq j$) si ha una **classificazione errata**.

L'**errore complessivo** corrisponde alla somma delle attribuzioni errate.

L'AD può essere effettuata sia in chiave parametrica, assumendo l'ipotesi di multinormalità di ciascuna delle sottopopolazioni (accompagnata sia dall'ipotesi di omoschedasticità sia di eteroschedasticità), sia sotto condizioni più generali, che non implicano assunzioni circa la distribuzione delle variabili usate nell'analisi.

7.2 Proc discrim (sotto ipotesi di multinormalità e omoschedasticità)

Una delle possibili procedure per effettuare l'AD con il SAS è la **`proc discrim`**.

La specificazione delle ipotesi sulla popolazione viene effettuata mediante l'opzione `method` che consente di effettuare l'analisi sia in chiave parametrica (ipotizzando la multinormalità delle variabili nelle sottopopolazioni considerate) sia in chiave non parametrica. Per default il SAS effettua l'analisi sotto condizioni di multinormalità, che può comunque essere resa esplicita utilizzando l'istruzione `method=normal`.

Per quanto riguarda le ipotesi sulla matrice di covarianza si utilizza l'opzione `pool` che viene posta uguale a:

- "yes" nel caso in cui si voglia assumere l'ipotesi di omoschedasticità,
- "no" in caso si assuma l'ipotesi di eteroschedasticità
- "test", nel caso in cui si voglia far eseguire al SAS il test di omoschedasticità.

Le probabilità a priori vengono invece esplicitate attraverso l'opzione `priors` a cui viene affiancata la keyword:

- `equal` quando si vuole attribuire a ciascun gruppo la stessa probabilità
- `proportional` (oppure `prop`) quando si vuole attribuire a ciascun gruppo una probabilità proporzionale al numero di osservazioni nei diversi gruppi.

Una volta ottenute le probabilità a posteriori e, di conseguenza, una volta che ciascuna unità sia stata attribuita a uno dei *g* gruppi, si procede alle stime delle probabilità di classificazione errata per ciascun gruppo. Per il *j*-esimo gruppo questa stima corrisponde al rapporto fra il numero di classificazioni errate per il *j*-esimo gruppo e la numerosità del gruppo stesso N_j ($j = 1, 2, \ldots, g$).

La stima complessiva del tasso di errore (ossia la stima generale della probabilità di attribuzioni errate) si può ottenere dalla media di queste *g* probabilità di classificazione errata, ciascuna ponderata con la probabilità a priori associata al gruppo.

Un metodo di stima migliore di questo tasso si ottiene con questo procedimento:

- si elimina dalle *n* osservazioni una singola osservazione (per esempio la prima)
- si costruisce la regola di attribuzione sulla base di queste *n*-1 osservazioni
- si attribuisce l'osservazione eliminata (la prima) ad uno dei *g* gruppi sulla base della regola di decisione ottenuta
- si verifica se l'attribuzione risulta corretta oppure errata
- si ripete questo procedimento per ciascuna osservazione, per cui ogni volta si ottiene una regola di decisione sulla base di *n*-1 osservazioni, mentre l'osservazione non considerata per determinare questa regola di decisione viene successivamente attribuita ad uno dei *g* gruppi sulla base delle probabilità calcolate con le *n*-1 osservazioni.
- si calcola infine la media di queste *g* probabilità di classificazione errata ponderate con la probabilità a priori associata a ciascun gruppo

Questo metodo di stima delle probabilità di attribuzioni errate (comunemente chiamato **cross validation**) può essere richiesto al SAS utilizzando l'opzione `crossvalidate` all'interno della **`proc discrim`**.

Esempio

Nel listato seguente è riportato un dataset relativo a 5 diverse specie di piante (coltura) per le quali sono state misurate 4 variabili quantitative (X_1, X_2, X_3 e X_4), con l'obiettivo di verificare se il valore assunto da queste variabili consente di distinguere le diverse specie di piante.

```
data a;
input coltura $ x1 x2 x3 x4;
cards;
grano 16 27 31 33
grano 15 23 30 30
grano 16 27 27 26
grano 18 20 25 23
grano 15 15 31 32
grano 15 32 32 15
grano 12 15 16 73
soia 20 23 23 25
soia 24 24 25 32
soia 21 25 23 24
soia 27 45 24 12
soia 12 13 15 42
soia 22 32 31 43
cotone 31 32 33 34
cotone 29 24 26 28
cotone 34 32 28 45
cotone 26 25 23 24
cotone 53 48 75 26
cotone 34 35 25 78
barbab 22 23 25 42
barbab 25 25 24 26
barbab 34 25 16 52
barbab 54 23 21 54
barbab 25 43 32 15
barbab 26 54  2 54
trifog 12 45 32 54
trifog 24 58 25 34
trifog 87 54 61 21
trifog 51 31 31 16
trifog 96 48 54 62
trifog 31 31 11 11
trifog 56 13 13 71
trifog 32 13 27 32
trifog 36 26 54 32
trifog 53  8  6 54
trifog 32 32 62 16
;
 proc discrim method=normal pool=yes
                list crossvalidate;
      priors prop;
      class coltura;
      var x1 x2 x3 x4;
   run;
```

In questo caso si sono esplicitate le ipotesi di multinormalità (`method=normal`) e di omoschedasticità (`pool=yes`), che sono comunque assunte per default dal SAS.

L'uso dell'opzione `list` consente di ottenere nella finestra di output i risultati della classificazione per ciascuna osservazione, mentre l'opzione `crossvalidate` fornisce le stime delle probabilità di errata classificazione mediante il metodo cross validation.

Con l'opzione `priors prop` le probabilità a priori sono state poste proporzionali alle numerosità dei singoli gruppi.

Nella keyword `class` va indicata la variabile che identifica i gruppi (nel caso esaminato i diversi tipi di piante), mentre la keyword `var` serve per specificare le variabili che devono essere utilizzate per determinare l'aggiornamento delle probabilità a priori, ottenendo le probabilità a posteriori che vengono utilizzate per la classificazione delle osservazioni.

Nelle tabelle successive sono riportati i principali risultati ottenuti dall'esecuzione della procedura.

Tabella 7.2.1
Scomposizione dei gradi di libertà

Dim. camp. totale	36	**Totale DF**	35
Variabili	4	**DF entro le classi**	31
Classi	5	**DF tra le classi**	4

Tabella 7.2.2
Osservazioni lette e utilizzate

Numero osservazioni lette	36
Numero osservazioni usate	36

Tabella 7.2.3
Probabilità a priori

coltura	Nome variabile	Frequenza	Peso	Proporzione	Probabilità a priori
barbab	barbab	6	6.0000	0.166667	0.166667
Cotone	cotone	6	6.0000	0.166667	0.166667
Grano	grano	7	7.0000	0.194444	0.194444
Soia	soia	6	6.0000	0.166667	0.166667
Trifog	trifog	11	11.0000	0.305556	0.305556

Una volta aggiornate le probabilità a priori in quelle a posteriori, sulla base dei valori assunti dalle 4 variabili quantitative, le diverse osservazioni vengono attribuite ai 5 diversi gruppi. La tabella successiva mostra, per ogni osservazione, il vero gruppo di appartenenza (Da coltura) e il gruppo di

classificazione (Classificata in coltura), evidenziando le errate attribuzioni mediante un asterisco (nella quarta colonna della tabella). Nelle colonne successive sono riportate le probabilità a posteriori (le caselle in grigio evidenziano le probabilità a posteriori che determinano le errate attribuzioni).

Tabella 7.2.4
Probabilità a posteriori ed errate classificazioni

Oss	Da coltura	Classificata in coltura		barbab	cotone	grano	soia	trifog
1	grano	grano		0.0897	0.1763	0.4054	0.2392	0.0894
2	grano	grano		0.0722	0.1421	0.4558	0.2530	0.0769
3	grano	grano		0.1157	0.1365	0.3422	0.3073	0.0982
4	grano	grano		0.0955	0.1078	0.3634	0.3281	0.1052
5	grano	grano		0.0398	0.1173	0.5754	0.2087	0.0588
6	grano	soia	*	0.1011	0.1318	0.3278	0.3420	0.0972
7	grano	grano		0.1083	0.1849	0.5238	0.1376	0.0454
8	soia	soia		0.1385	0.1176	0.2804	0.3305	0.1330
9	soia	soia		0.1502	0.1586	0.2483	0.2660	0.1768
10	soia	soia		0.1570	0.1200	0.2431	0.3318	0.1481
11	soia	barbab	*	0.3359	0.1016	0.0547	0.2721	0.2357
12	soia	grano	*	0.1013	0.0920	0.4749	0.2768	0.0549
13	soia	cotone	*	0.1448	0.2624	0.2606	0.1848	0.1474
14	cotone	trifog	*	0.1523	0.2377	0.1518	0.1767	0.2815
15	cotone	soia	*	0.1559	0.1529	0.1842	0.2549	0.2521
16	cotone	trifog	*	0.2091	0.2404	0.1023	0.1357	0.3125
17	cotone	soia	*	0.1780	0.1245	0.1809	0.3045	0.2121
18	cotone	trifog	*	0.0166	0.4384	0.0391	0.0223	0.4837
19	cotone	cotone		0.2548	0.3810	0.0794	0.0592	0.2256
20	barbab	grano	*	0.1381	0.1901	0.3066	0.2231	0.1421
21	barbab	soia	*	0.1667	0.1354	0.2050	0.2960	0.1969
22	barbab	barbab		0.3056	0.1665	0.0871	0.1479	0.2928
23	barbab	trifog	*	0.1845	0.1250	0.0194	0.0496	0.6215
24	barbab	soia	*	0.2191	0.1646	0.1135	0.2770	0.2258
25	barbab	barbab		0.7887	0.0521	0.0081	0.0661	0.0850
26	trifog	cotone	*	0.1789	0.3394	0.2663	0.1460	0.0693
27	trifog	barbab	*	0.4845	0.1680	0.0376	0.1452	0.1647
28	trifog	trifog		0.0165	0.0478	0.0003	0.0025	0.9328
29	trifog	trifog		0.1322	0.0872	0.0205	0.0959	0.6642
30	trifog	trifog		0.0173	0.0604	0.0002	0.0007	0.9215

Oss	Da coltura	Classificata in coltura		barbab	cotone	grano	soia	trifog
31	**trifog**	**barbab**	*	0.3588	0.0473	0.0402	0.3012	0.2525
32	**trifog**	**trifog**		0.2023	0.1226	0.0212	0.0408	0.6132
33	**trifog**	**trifog**		0.0943	0.1512	0.2616	0.2260	0.2669
34	**trifog**	**cotone**	*	0.0292	0.3495	0.2645	0.0918	0.2650
35	**trifog**	**trifog**		0.2392	0.0676	0.0237	0.0781	0.5914
36	**trifog**	**cotone**	*	0.0206	0.3327	0.3180	0.1125	0.2163

La tabella 7.2.5 fornisce una valutazione complessiva delle attribuzioni, dove solo i valori riportati sulla diagonale principale (evidenziati in colore verde) corrispondono ai casi di corretta attribuzione.

Tabella 7.2.5
Riepilogo della classificazione

Da coltura	barbab	cotone	grano	soia	trifog	Totale
barbab	2 33.33	0 0.00	1 16.67	2 33.33	1 16.67	6 100.00
cotone	0 0.00	1 16.67	0 0.00	2 33.33	3 50.00	6 100.00
grano	0 0.00	0 0.00	6 85.71	1 14.29	0 0.00	7 100.00
soia	1 16.67	1 16.67	1 16.67	3 50.00	0 0.00	6 100.00
trifog	2 18.18	3 27.27	0 0.00	0 0.00	6 54.55	11 100.00
Totale	5 13.89	5 13.89	8 22.22	8 22.22	10 27.78	36 100.00
Pr a priori	0.16667	0.16667	0.19444	0.16667	0.30556	

Nella tabella seguente sono riportate le stime delle errate attribuzioni, corrispondenti al rapporto fra il numero di attribuzioni errate per ciascuna coltura ed il numero di osservazioni corrispondenti. Così, per esempio il tasso delle errate attribuzioni per le barbabietole è 4/6, quello del trifoglio è 5/11.

Tabella 7.2.6
Stime conteggio errori per coltura

	barbab	**cotone**	**grano**	**soia**	**trifog**	**Totale**
Tasso	0.6667	0.8333	0.1429	0.5000	0.4545	0.5000
Pr a priori	0.1667	0.1667	0.1944	0.1667	0.3056	

In questo esempio i risultati ottenuti con l'AD sono deludenti: le piante di grano risultano generalmente classificate in modo corretto, ma solo una di quelle di cotone viene attribuita correttamente. In tutti gli altri casi il tasso di errata attribuzione si aggira intorno al 50%.
Le stesse informazioni delle tabelle 7.2.5 e 7.2.6 sono contenute nelle tabelle 7.2.7 e 7.2.8 quando si utilizza il metodo della cross validation, che porta ovviamente a risultati peggiori rispetto ai precedenti.

Tabella 7.2.7
Riepilogo della cross validation

Da coltura	**barbab**	**cotone**	**grano**	**soia**	**trifog**	**Totale**
barbab	1 16.67	0 0.00	1 16.67	2 33.33	2 33.33	6 100.00
cotone	1 16.67	0 0.00	0 0.00	2 33.33	3 50.00	6 100.00
grano	0 0.00	1 14.29	4 57.14	2 28.57	0 0.00	7 100.00
soia	1 16.67	1 16.67	1 16.67	3 50.00	0 0.00	6 100.00
trifog	3 27.27	1 9.09	3 27.27	0 0.00	4 36.36	11 100.00
Totale	6 16.67	3 8.33	9 25.00	9 25.00	9 25.00	36 100.00

Tabella 7.2.8
Stime conteggio errori per coltura (cross validation)

Tasso	0.8333	1.0000	0.4286	0.5000	0.6364	0.6667
Pr a priori	0.1667	0.1667	0.1944	0.1667	0.3056	

7.3 Proc discrim (sotto ipotesi di multinormalità e test di omoschedasticità)

Nel seguente listato SAS è riportato un file storico che venne usato da Fisher per distinguere tre diverse specie di iris (setosa, versicolor e virginica) sulla base di 4 diverse variabili quantitative (lunghezza dei sepali, larghezza dei sepali, lunghezza dei petali e larghezza dei petali).

In questo caso si è adottata l'ipotesi di multinormalità, ma si è richiesto al SAS di effettuare il test di omoschedasticità.

```
data a;
input lungsep largsep lungpet largpet iris $;
cards;
5.1 3.5 1.4 0.2 setosa
4.9 3.0 1.4 0.2 setosa
4.7 3.2 1.3 0.2 setosa
4.6 3.1 1.5 0.2 setosa
5.0 3.6 1.4 0.2 setosa
5.4 3.9 1.7 0.4 setosa
4.6 3.4 1.4 0.3 setosa
5.0 3.4 1.5 0.2 setosa
4.4 2.9 1.4 0.2 setosa
4.9 3.1 1.5 0.1 setosa
5.4 3.7 1.5 0.2 setosa
4.8 3.4 1.6 0.2 setosa
4.8 3.0 1.4 0.1 setosa
4.3 3.0 1.1 0.1 setosa
5.8 4.0 1.2 0.2 setosa
5.7 4.4 1.5 0.4 setosa
5.4 3.9 1.3 0.4 setosa
5.1 3.5 1.4 0.3 setosa
5.7 3.8 1.7 0.3 setosa
5.1 3.8 1.5 0.3 setosa
5.4 3.4 1.7 0.2 setosa
5.1 3.7 1.5 0.4 setosa
4.6 3.6 1.0 0.2 setosa
5.1 3.3 1.7 0.5 setosa
4.8 3.4 1.9 0.2 setosa
5.0 3.0 1.6 0.2 setosa
5.0 3.4 1.6 0.4 setosa
5.2 3.5 1.5 0.2 setosa
5.2 3.4 1.4 0.2 setosa
4.7 3.2 1.6 0.2 setosa
4.8 3.1 1.6 0.2 setosa
5.4 3.4 1.5 0.4 setosa
5.2 4.1 1.5 0.1 setosa
5.5 4.2 1.4 0.2 setosa
4.9 3.1 1.5 0.1 setosa
```

```
5.0 3.2 1.2 0.2 setosa
5.5 3.5 1.3 0.2 setosa
4.9 3.1 1.5 0.1 setosa
4.4 3.0 1.3 0.2 setosa
5.1 3.4 1.5 0.2 setosa
5.0 3.5 1.3 0.3 setosa
4.5 2.3 1.3 0.3 setosa
4.4 3.2 1.3 0.2 setosa
5.0 3.5 1.6 0.6 setosa
5.1 3.8 1.9 0.4 setosa
4.8 3.0 1.4 0.3 setosa
5.1 3.8 1.6 0.2 setosa
4.6 3.2 1.4 0.2 setosa
5.3 3.7 1.5 0.2 setosa
5.0 3.3 1.4 0.2 setosa
7.0 3.2 4.7 1.4 versicolor
6.4 3.2 4.5 1.5 versicolor
6.9 3.1 4.9 1.5 versicolor
5.5 2.3 4.0 1.3 versicolor
6.5 2.8 4.6 1.5 versicolor
5.7 2.8 4.5 1.3 versicolor
6.3 3.3 4.7 1.6 versicolor
4.9 2.4 3.3 1.0 versicolor
6.6 2.9 4.6 1.3 versicolor
5.2 2.7 3.9 1.4 versicolor
5.0 2.0 3.5 1.0 versicolor
5.9 3.0 4.2 1.5 versicolor
6.0 2.2 4.0 1.0 versicolor
6.1 2.9 4.7 1.4 versicolor
5.6 2.9 3.6 1.3 versicolor
6.7 3.1 4.4 1.4 versicolor
5.6 3.0 4.5 1.5 versicolor
5.8 2.7 4.1 1.0 versicolor
6.2 2.2 4.5 1.5 versicolor
5.6 2.5 3.9 1.1 versicolor
5.9 3.2 4.8 1.8 versicolor
6.1 2.8 4.0 1.3 versicolor
6.3 2.5 4.9 1.5 versicolor
6.1 2.8 4.7 1.2 versicolor
6.4 2.9 4.3 1.3 versicolor
6.6 3.0 4.4 1.4 versicolor
6.8 2.8 4.8 1.4 versicolor
6.7 3.0 5.0 1.7 versicolor
6.0 2.9 4.5 1.5 versicolor
5.7 2.6 3.5 1.0 versicolor
5.5 2.4 3.8 1.1 versicolor
5.5 2.4 3.7 1.0 versicolor
5.8 2.7 3.9 1.2 versicolor
6.0 2.7 5.1 1.6 versicolor
5.4 3.0 4.5 1.5 versicolor
6.0 3.4 4.5 1.6 versicolor
6.7 3.1 4.7 1.5 versicolor
6.3 2.3 4.4 1.3 versicolor
5.6 3.0 4.1 1.3 versicolor
5.5 2.5 4.0 1.3 versicolor
5.5 2.6 4.4 1.2 versicolor
6.1 3.0 4.6 1.4 versicolor
5.8 2.6 4.0 1.2 versicolor
5.0 2.3 3.3 1.0 versicolor
5.6 2.7 4.2 1.3 versicolor
5.7 3.0 4.2 1.2 versicolor
5.7 2.9 4.2 1.3 versicolor
```

```
6.2 2.9 4.3 1.3 versicolor
5.1 2.5 3.0 1.1 versicolor
5.7 2.8 4.1 1.3 versicolor
6.3 3.3 6.0 2.5 virginica
5.8 2.7 5.1 1.9 virginica
7.1 3.0 5.9 2.1 virginica
6.3 2.9 5.6 1.8 virginica
6.5 3.0 5.8 2.2 virginica
7.6 3.0 6.6 2.1 virginica
4.9 2.5 4.5 1.7 virginica
7.3 2.9 6.3 1.8 virginica
6.7 2.5 5.8 1.8 virginica
7.2 3.6 6.1 2.5 virginica
6.5 3.2 5.1 2.0 virginica
6.4 2.7 5.3 1.9 virginica
6.8 3.0 5.5 2.1 virginica
5.7 2.5 5.0 2.0 virginica
5.8 2.8 5.1 2.4 virginica
6.4 3.2 5.3 2.3 virginica
6.5 3.0 5.5 1.8 virginica
7.7 3.8 6.7 2.2 virginica
7.7 2.6 6.9 2.3 virginica
6.0 2.2 5.0 1.5 virginica
6.9 3.2 5.7 2.3 virginica
5.6 2.8 4.9 2.0 virginica
7.7 2.8 6.7 2.0 virginica
6.3 2.7 4.9 1.8 virginica
6.7 3.3 5.7 2.1 virginica
7.2 3.2 6.0 1.8 virginica
6.2 2.8 4.8 1.8 virginica
6.1 3.0 4.9 1.8 virginica
6.4 2.8 5.6 2.1 virginica
7.2 3.0 5.8 1.6 virginica
7.4 2.8 6.1 1.9 virginica
7.9 3.8 6.4 2.0 virginica
6.4 2.8 5.6 2.2 virginica
6.3 2.8 5.1 1.5 virginica
6.1 2.6 5.6 1.4 virginica
7.7 3.0 6.1 2.3 virginica
6.3 3.4 5.6 2.4 virginica
6.4 3.1 5.5 1.8 virginica
6.0 3.0 4.8 1.8 virginica
6.9 3.1 5.4 2.1 virginica
6.7 3.1 5.6 2.4 virginica
6.9 3.1 5.1 2.3 virginica
5.8 2.7 5.1 1.9 virginica
6.8 3.2 5.9 2.3 virginica
6.7 3.3 5.7 2.5 virginica
6.7 3.0 5.2 2.3 virginica
6.3 2.5 5.0 1.9 virginica
6.5 3.0 5.2 2.0 virginica
6.2 3.4 5.4 2.3 virginica
5.9 3.0 5.1 1.8 virginica
;
 proc discrim method=normal pool=test crossvalidate;
      priors prop;
      class iris;
      var lungsep largsep lungpet largpet;
   run;
```

I principali risultati sono elencati nelle tabelle seguenti

Tabella 7.3.1
Scomposizione dei gradi di libertà

Dim. camp. totale	150	**Totale DF**	149
Variabili	4	**DF entro le classi**	147
Classi	3	**DF tra le classi**	2

Tabella 7.3.2
Probabilità a priori

iris	Nome variabile	Frequenza	Peso	Proporzione	Probabilità a priori
setosa	setosa	50	50.0000	0.333333	0.333333
versicol	versicol	50	50.0000	0.333333	0.333333
virginic	virginic	50	50.0000	0.333333	0.333333

Tabella 7.3.3
Test di omoschedasticità

Chi-quadrato	DF	Pr > ChiQuadr
139.236945	20	<.0001

L'ipotesi di omoschedasticità viene quindi rifiutata, come si nota dalla seguente nota del SAS

Poiché il valore del chi-quadrato è significativo a livello 0.1, verranno usate matrici di covarianza interne nella funzione discriminante.
Riferimento: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Il numero di osservazioni per ciascun gruppo è sempre pari a 50, per cui le probabilità a priori sono tutte uguali a 1/3 sia se si utilizza il criterio di probabilità a priori tutte uguali, sia con il criterio proporzionale. Tali probabilità vengono poi aggiornate sulla base dei valori assunti dalle 4 variabili quantitative relative alla lunghezza e alla larghezza dei sepali e alla lunghezza e alla larghezza dei petali, ottenendo i valori delle probabilità a posteriori utilizzate per attribuire le 150 osservazioni ai tre diversi gruppi di iris.

In questo caso, a causa dell'elevato numero di osservazioni considerate, non si è chiesto al SAS di stampare le classificazioni per ogni singola osservazione, ma i risultati complessivi, riportati nella tabella successiva, mostrano come il tasso di errate attribuzioni sia molto prossimo a zero, per cui le 4 variabili considerate consentono effettivamente di stabilire un'ottima regola attribuzione delle osservazioni ai tre diversi sottogruppi.

Tabella 7.3.4
Riepilogo della classificazione

Numero di osservazioni e percentuale classificata in iris				
Da iris	**setosa**	**versicol**	**virginic**	**Totale**
setosa	50 100.00	0 0.00	0 0.00	50 100.00
versicol	0 0.00	48 96.00	2 4.00	50 100.00
virginic	0 0.00	1 2.00	49 98.00	50 100.00
Totale	50 33.33	49 32.67	51 34.00	150 100.00
Pr a priori	0.33333	0.33333	0.33333	

Tabella 7.3.5
Stime conteggio errori per iris

	setosa	**versicol**	**virginic**	**Totale**
Tasso	0.0000	0.0400	0.0200	0.0200
Pr a priori	0.3333	0.3333	0.3333	

Tutti gli iris setosa vengono classificati correttamente, solo un iris virginica (pari al 2%) viene classificato erroneamente negli iris versicolor, mentre 2 iris versicolor (pari al 4%) vengono erroneamente attribuiti agli iris virginica.

Applicando il metodo della cross validation si ottiene solo un lievissimo peggioramento dei risultati, come si osserva dai dati contenuti nelle due tabelle successive.

Tabella 7.3.6
Riepilogo della cross validation

Numero di osservazioni e percentuale classificata in iris				
Da iris	**setosa**	**versicol**	**virginic**	**Totale**
setosa	50 100.00	0 0.00	0 0.00	50 100.00
versicol	0 0.00	47 94.00	3 6.00	50 100.00
virginic	0 0.00	1 2.00	49 98.00	50 100.00
Totale	50 33.33	48 32.00	52 34.67	150 100.00
Pr a priori	0.33333	0.33333	0.33333	

Tabella 7.3.7
Stime conteggio errori per iris (cross validation)

	setosa	**versicol**	**virginic**	**Totale**
Tasso	0.0000	0.0600	0.0200	0.0267
Pr a priori	0.3333	0.3333	0.3333	

La `proc discrim` può essere anche utilizzata sotto condizioni più generali, che non ipotizzano la multinormalità dei dati, utilizzando l'opzione `method=npar`. Tuttavia la trattazione di questo caso è abbastanza complessa e va oltre gli scopi di queste dispense.

7.4 Proc stepdisc

Questa procedura effettua un'analisi discriminante con il metodo **stepwise**, ossia consente di selezionare le variabili in base alla loro capacità discriminante. Questa procedura esegue un'analisi discriminante "graduale" in modo da selezionare il miglior sottoinsieme delle variabili quantitative da utilizzare nella discriminazione. Anche in questo caso si ipotizza che l'insieme di variabili abbia una distribuzione normale multivariata con una matrice di covarianza comune.

```
PROC STEPDISC <options> ;
        CLASS variable ;
        VAR variables ;
```

La selezione può avvenire in due modi:

- forward selection; si parte con nessuna variabile e si aggiungono quelle con il maggior potere discriminante.
- backward elimination: si parte da tutte le variabili e vengono rimosse quelle con il minor potere discriminante;

Nel primo caso (opzione **METHOD=FORWARD | FW** all'interno della proc stepdisc) la procedura inizia partendo con nessuna variabile nel modello e successivamente, a ogni passo,

- inserisce la variabile che contribuisce maggiormente al potere discriminatorio del modello (calcolando un opportuno test *F*, individuando la variabile per cui tale test assume il valore più elevato e verificando se vada inserita o meno, a seconda della significatività associata a tale valore)
- effettua un diverso test *F* per verificare se tale variabile vada rimossa dal modello o vada mantenuta al suo interno.

Quando tutte le variabili nel modello soddisfano il criterio di permanenza e nessuna delle altre variabili soddisfa il criterio di inserimento, il processo si interrompe. Questo è il metodo predefinito per la selezione delle variabili nella procedura SAS.

Va notato che il SAS effettua automaticamente un test per valutare la dipendenza lineare fra le variabili, in modo da escludere automaticamente dal modello le eventuali variabili che dipendono linearmente da altre variabili già inserite nel modello.

Come con qualsiasi procedura stepwise, è importante ricordare che quando vengono eseguiti molti test, ciascuno ad un livello di significatività α, la probabilità complessiva di rifiutare almeno un'ipotesi nulla vera è molto più grande di α. Per questo motivo, se si desidera impedire l'inclusione di variabili che non contribuiscono significativamente al potere discriminatorio del modello, può essere opportuno utilizzare un livello di significatività molto piccolo.

Bisogna tenere presente che la procedura inserisce una sola variabile nel modello in ogni fase e che il processo di selezione non tiene conto delle relazioni tra variabili che non sono state ancora selezionate. Pertanto alcune variabili potrebbero venire escluse nel processo e si potrebbe arrivare a un sottoinsieme non ottimale delle variabili utili per la discriminazione. Inoltre le statistiche utilizzate non sempre sono le migliori per misurare il potere discriminatorio delle variabili. Tuttavia la PROC STEPDISC risulta generalmente utile nella selezione delle variabili, per cui può risultare utile effettuarla prima di una procedura DISCRIM.

Con il metodo backward elimination (opzione **METHOD=BACKWARD | BW** all'interno della proc stepdisc), invece, la procedura inizia con tutte le variabili nel modello eccetto quelle che dipendono linearmente dalle altre variabili elencate nell'istruzione VAR. A ogni passo viene rimossa la variabile che contribuisce meno al

potere discriminatorio del modello. Il processo di eliminazione si arresta quando tutte le variabili soddisfano il criterio per rimanere nel modello.

Esempio

Si applichi la procedura Stepwise al dataset relativo agli iris utilizzando il metodo forward selection.

Il listato può assumere la forma seguente in cui non è necessario specificare l'opzione method=forward, dato che questo metodo di selezione viene usato per default dal software

```
proc stepdisc ;
   class iris;
   var lungsep largsep lungpet largpet;
run;
```

L'output assume la forma seguente

Il metodo di selezione delle variabili è STEPWISE			
Dim. camp. totale	150	**Variabile/i nell'analisi**	4
Livelli di classi	3	**Var. che saranno incluse**	0
		Livello sign. per entrare	0.15
		Livello sign. per restare	0.15

Numero osservazioni lette	150
Numero osservazioni usate	150

Informazioni sui livelli di classicazione				
iris	**Nome variabile**	**Frequenza**	**Peso**	**Proporzione**
setosa	setosa	50	50.0000	0.333333
versicol	versicol	50	50.0000	0.333333
virginic	virginic	50	50.0000	0.333333

Dopo queste informazioni generali iniziano i diversi passi: al primo passo si ottengono i seguenti risultati

La procedura STEPDISC
Selezione stepwise: Passo 1

Statistiche di entrata, DF = 2, 147				
Variabile	**R-quadro**	**Valore F**	**Pr > F**	**Tolleranza**
lungsep	0.6187	119.26	<.0001	1.0000

Statistiche di entrata, DF = 2, 147				
Variabile	**R-quadro**	**Valore F**	**Pr > F**	**Tolleranza**
largsep	0.3919	47.36	<.0001	1.0000
lungpet	0.9413	1179.03	<.0001	1.0000
largpet	0.9288	959.32	<.0001	1.0000

La variabile lungpet entrerà.

A ogni primo passo di una procedura STEPWISE la Tolleranza è sempre pari a 1, in quanto non c'è alcuna variabile nel modello. La scelta della variabile lungpet come variabile da far entrare nel modello è determinata dal test *F* che per tale variabile assume il valore più alto (pari a 1179.03) e che risulta altamente significativo (con un *p*-valore <0.0001).

Variabili immesse
Lungpet

Statistiche multivariate					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.058681	1179.03	2	147	<.0001
Traccia di Pillai	0.941319	1179.03	2	147	<.0001
Correlazione canonica quadrata media	0.470659				

Nello step successivo, Passo 2, la variabile lungpet viene inserita nel modello e viene effettuato un test *F* per verificare se vada rimossa dal modello prima di procedere alla selezione di una nuova variabile.

Selezione stepwise: Passo 2

Statistiche per rimozione, DF = 2, 147			
Variabile	**R-quadro**	**Valore F**	**Pr > F**
lungpet	0.9413	1179.03	<.0001

Nessuna variabile può essere rimossa.

La variabile non deve essere rimossa e la procedura verifica quale variabile possa essere aggiunta sulla base del test *F*

Statistiche di entrata, DF = 2, 146				
Variabile	**R-quadro parziale**	**Valore F**	**Pr > F**	**Tolleranza**
lungsep	0.3191	34.21	<.0001	0.2400
largsep	0.3663	42.19	<.0001	0.8232
largpet	0.2530	24.72	<.0001	0.0731

La variabile largsep entrerà.

Viene scelta la variabile largsep che presenta il maggior valore del test F, sempre altamente significativo, e un valore della tolleranza (0.8232) che non è al di sotto del valore soglia.

Variabili immesse	
largsep	lungpet

Statistiche multivariate					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.037187	305.55	4	292	<.0001
Traccia di Pillai	1.117177	93.01	4	294	<.0001
Correlazione canonica quadrata media	0.558589				

Si procede nel medesimo modo anche nei successivi passi, 3 e 4, ottenendo i risultati riportati di seguito

La procedura STEPDISC
Selezione stepwise: Passo 3

Statistiche per rimozione, DF = 2, 146			
Variabile	**R-quadro parziale**	**Valore F**	**Pr > F**
largsep	0.3663	42.19	<.0001
lungpet	0.9388	1120.78	<.0001

Nessuna variabile può essere rimossa.

Statistiche di entrata, DF = 2, 145				
Variabile	**R-quadro parziale**	**Valore F**	**Pr > F**	**Tolleranza**
lungsep	0.1450	12.30	<.0001	0.1330
largpet	0.3271	35.24	<.0001	0.0663

La variabile largpet entrerà.

Variabili immesse		
largsep	lungpet	largpet

Statistiche multivariate					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.025025	257.20	6	290	<.0001
Traccia di Pillai	1.185274	70.80	6	292	<.0001
Correlazione canonica quadrata media	0.592637				

La procedura STEPDISC
Selezione stepwise: Passo 4

Statistiche per rimozione, DF = 2, 145			
Variabile	**R-quadro parziale**	**Valore F**	**Pr > F**
largsep	0.4291	54.50	<.0001
lungpet	0.3449	38.18	<.0001
largpet	0.3271	35.24	<.0001

Nessuna variabile può essere rimossa.

Statistiche di entrata, DF = 2, 144				
Variabile	**R-quadro parziale**	**Valore F**	**Pr > F**	**Tolleranza**
lungsep	0.0599	4.59	0.0117	0.0318

La variabile lungsep entrerà.

Tutte le variabili sono entrate.

Statistiche multivariate					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.023525	198.71	8	288	<.0001
Traccia di Pillai	1.187207	52.95	8	290	<.0001
Correlazione canonica quadrata media	0.593603				

Dato che non è possibile aggiungere o rimuovere altre variabili dal modello, la procedura si interrompe al punto 5 e visualizza un riepilogo del processo di selezione.

La procedura STEPDISC
Selezione stepwise: Passo 5

Statistiche per rimozione, DF = 2, 144			
Variabile	**R-quadro parziale**	**Valore F**	**Pr > F**
lungsep	0.0599	4.59	0.0117
largsep	0.2323	21.78	<.0001
lungpet	0.3263	34.87	<.0001
largpet	0.2601	25.31	<.0001

Nessuna variabile può essere rimossa.

Non sono possibili ulteriori passi.

La procedura STEPDISC

Riepilogo della selezione stepwise										
Passo	**Numero in**	**Immesso**	**Rimosso**	**R-quadro parziale**	**Valore F**	**Pr > F**	**Lambda di Wilks**	**Pr < Lambda**	**Correlazione canonica quadrata media**	**Pr > ASCC**
1	**1**	**lungpet**		0.9413	1179.03	<.0001	0.05868103	<.0001	0.47065949	<.0001
2	**2**	**largsep**		0.3663	42.19	<.0001	0.03718672	<.0001	0.55858862	<.0001
3	**3**	**largpet**		0.3271	35.24	<.0001	0.02502460	<.0001	0.59263711	<.0001
4	**4**	**lungsep**		0.0599	4.59	0.0117	0.02352545	<.0001	0.59360338	<.0001

Se si utilizzasse lo stesso dataset con il metodo backward elimination

```
proc stepdisc method=bw;
      class iris;
```

```
      var lungsep largsep lungpet largpet;
```

il SAS si fermerebbe al proimo step, in quanto i test porterebbero a concludere che tutte e quattro le variabili devono essere mantenute.

Se, infine, si utilizzasse la proc stepdisc sul dataset relativo alle colture, il metodo forward selection

```
proc stepdisc method=fw;
      class coltura;
      var x1 x2 x3 x4;
```

porta a selezionare la sola variabile x1, mentre tutte le altre sarebbero non soddisfano il criterio di immissione.