３章　判別分析

複数の変数によって、分類の変数を予想する手法を判別分析と言います。例えばいくつかの模擬試験の点数によって入試の合否を予想するなどは典型的な例です。以下の例を見てみましょう。

例

入学試験の合否と勉強時間・模擬試験の平均点のデータを求めたところ以下のような結果を得た（Samples¥判別分析１.txt)。合否を判定するための勉強時間と平均点の１次関数を求めよ。またこの関数によってこのデータを判別し、誤判別の確率を求めよ。

合否	勉強時間	平均点	合否	勉強時間	平均点
1	5.6	70.2	2	3.8	67.4
1	5.9	74.2	2	3.8	61.3
1	4.1	72.7	2	1.7	60.6
1	5.1	84.9	2	2.7	77.2
1	5.0	93.0	2	4.3	65.9
1	3.2	80.5	2	3.3	74.4
1	4.3	62.7	2	3.5	72.1
1	4.8	85.4	2	2.1	69.7
1	3.3	84.3	2	4.3	68.7
1	5.3	64.8	2	2.0	70.5
1	5.3	60.7	2	3.6	45.9
1	5.4	74.4	2	2.8	54.6
1	3.6	85.5	2	2.5	64.4
2	3.8	47.9	2	5.2	50.7
2	3.9	70.8	2	2.2	65.7

勉強時間と平均点で散布図を描いてみましょう。そのとき合格者を白丸、不合格者を黒丸で描いたとします。そうすると図 3.1 のような点が描かれたとしましょう（現実には勉強時間はあまり関係ないらしいですが)。群 1 は合格群で、平均点が高く、勉強時間も長い群です。群 2 は不合格群で平均点は低く、勉強時間も短い群です。これらの群を合格と不合格で２つに分けることを考えます。群分けには直線を使うものと仮定し、できるだけ誤判別がないようにと考えると、図 3.1 に描かれたような直線を引くことになります。

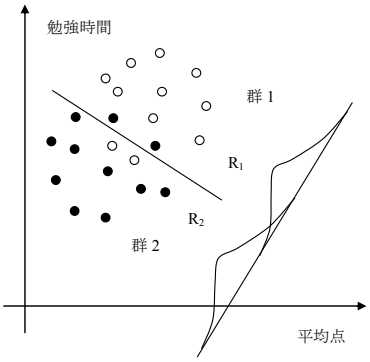


図 3.1 判別分析の概念図

2 次元平面上で直線は、以下のように表されます。

$$b_1 \times \text{勉強時間} + b_2 \times \text{平均点} + b_0 = 0 \qquad (y = ax + b \text{ はこの変形です})$$

特に$b_1 > 0$ とすると、この式の符号で領域が決まります。

$$z = b_1 \times \text{勉強時間} + b_2 \times \text{平均点} + b_0 \geq 0 \qquad (\text{領域 1})$$

$$z = b_1 \times \text{勉強時間} + b_2 \times \text{平均点} + b_0 < 0 \qquad (\text{領域 2})$$

このように直線（一般には平面）の式の符号を判別することで、２つの領域の判別ができることになります。この式を判別関数といいます。

実際に判別分析を見て行きましょう。メニュー［分析－多変量解析－判別分析］を選択すると図 3.2 のような分析メニューが表示されます。

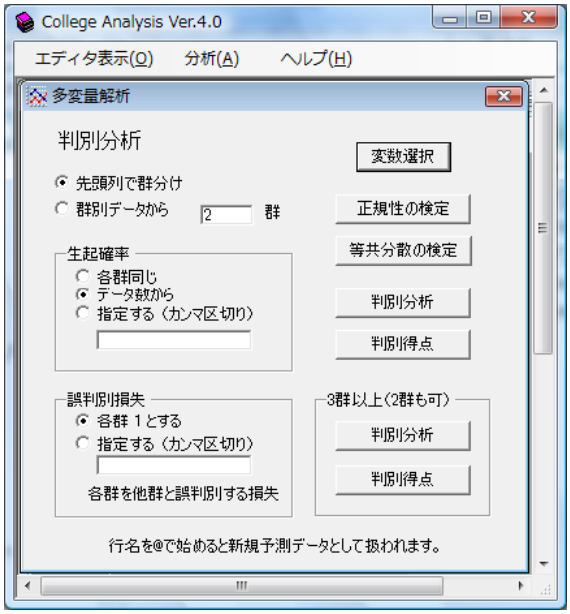


図 3.2 判別分析メニュー画面

変数は最初に群を分ける変数を選び、その後それを判別するのに利用する変数を選択します。「最初は分析名のボタンから」なので、「判別分析」ボタンをクリックすると、図 3.3 のような結果が表示されます。

図 3.3 判別分析結果

ここで判別関数の係数は一番上に表示されています。また、各変数を標準化して計算を実行した結果が、次の標準化係数です。これは判別関数における各変数の重要性を考える際に役立ちます。F 検定値、自由度、確率は判別関数の係数が 0 か否かの検定結果です。確率の値が有意水準以下なら 0 と異なるといえると判定します。

判別については、データを判別関数に代入して、0 以上なら 1 群（辞書順で前の群）、0 未満なら 2 群（辞書順で後の群）とします。

マハラノビスの距離以下は誤判別についての表示です。誤判別確率には、2 群の分布を多変量正規分布と仮定した場合の理論的な誤判別確率と実測データを分析にかけて求められた誤判別確率の 2 通りがあります。それぞれ「理論から」と「実測から」となっています。またその上にある「1 群を 2 群と」とは、本来 1 群であるデータを 2 群と誤判別する確率と解釈します。「2 群を 1 群と」はその逆です。マハラノビスの距離は各群のデータが多変量正規分布すると仮定した場合の 2 つの群の中心の距離の 2 乗で、どの程度 2 群が離れているかを表わす指標と考えればよいでしょう。表 3.1 にマハラノビスの距離と誤判別確率の値との関係を示します。

表 3.1 マハラノビスの距離と誤判別確率

マハラノビス距離	1	4	9	16	25
誤判別確率	0.309	0.159	0.067	0.023	0.006

次の誤判別の部分は左が実測の群、上が予測の群で、それぞれのデータがどこに判別されるか、そのデータ数を表示しています。その下は分類されたデータ出現の確率（割合）です。実測と異なった部分の確率に注意して下さい。
具体的な判別結果を見るには、「判別得点」ボタンをクリックします。実際の所属群と判別得点、それから予想した判別群が図 3.4 のように表示されます。

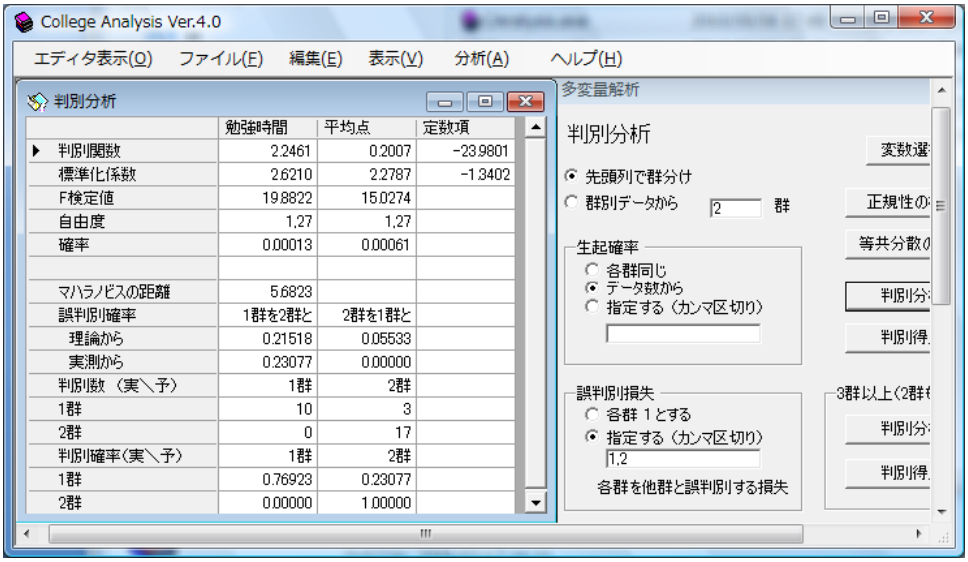
図 3.4 判別得点結果

次に、図 3.2 左側の生起確率と誤判別損失についてです。判別分析は元々２つの群の出現確率は等しいと仮定されています。しかしこの確率が大きく異なる場合は、生起確率を指定することができます。記述法は、群 1 から確率をカンマ区切りで書いて行きます。しかし実用にはデータをランダムに抽出して、生起確率がデータ数に比例するようにして、「データ数」からのラジオボタンを選択することです。デフォルトではそのような設定になっています。

誤判別損失については、以下の例を考えてみましょう。受験生に「あなたは不合格でしょう」と予測して合格になった場合と「あなたは合格でしょう」と予測して不合格になった場合とを比べてどちらが問題でしょうか。おそらく合格と言われて不合格になった方がダメージは大きいはずです。このように同じ誤判別でも損失の大きさが異なる場合に誤判別損失を指定します。例えば上の例では、合格群（不合格と判定）

と不合格群（合格と判定）に対して、1,2 などと指定します。

生起確率と誤判別損失は群の境界の平行移動を引き起こしますので、変化するのは判別関数の定数項の部分です。合格と不合格に対して誤判別損失を指定して判別分析を実行してみましょう。図 3.5 にその結果を示します。



	勉強時間	平均点	定数項
判別関数	2.2461	0.2007	-23.9801
標準化係数	2.6210	2.2787	-1.3402
F検定値	19.8822	15.0274	
自由度	1,27	1,27	
確率	0.00013	0.00061	
マハラノビスの距離	5.6823		
誤判別確率	1群を2群と	2群を1群と	
理論から	0.21518	0.05533	
実測から	0.23077	0.00000	
判別数（実＼予）	1群	2群	
1群	10	3	
2群	0	17	
判別確率(実＼予)	1群	2群	
1群	0.76923	0.23077	
2群	0.00000	1.00000	

図 3.5 誤判別損失を指定した判別分析結果

この結果を図 3.3 と比較してみると、まず判別関数の定数項が-23.2869 から-23.9801 に減っています。これにより 1 群（合格、判別関数が非負）と判別しにくくなります。それに伴い、誤判別確率も変わってきます。誤判別損失を考えると、実測からの 2 群（不合格）を 1 群（合格）と誤判別する確率は 0 になっていますし、理論の値も 0.09605 から 0.06533 と小さくなっています。但し、誤判別損失の大きさの比較は非常に人為的なので、個人的には各群 1 としておいても良いように思います。

以上 2 つの群に分ける場合を考えてきましたが、3 群以上に分ける場合も考えられます。以下の問題にもありますが、Samples¥判別分析 3.txt を開いて、変数すべてを読込み、分析メニューの 3 群以上のグループボックスで「判別分析」ボタンをクリックすると図 3.5 のような結果が表示されます。

図 3.5 3 群以上の判別分析結果

これはフィッシャーが利用した有名な 3 種類のあやめのデータで、いろいろな教科書でもよく利用されています。2 群に分ける場合と比べて、判別関数が 3 つになっています。判別はデータの変数値を代入したとき、これらの関数の中で最大となる群に所属すると判定します。方式がこれまでと全く違うように見えるので戸惑われるかも知れませんが、実は 2 群の判別の場合でも 2 つの判別関数で判別する方法もあります。この教科書で使った方法は、これら 2 つの判別関数の差を取って正と負の値で分けただけで、2 つの判別関数の大きい方と判定しても全く同じです。基本的な教科書には差を取る方法で紹介されている場合が多いので、両方の形を出力するようにしています。後で問題にもありますので見て下さい。

最後になりましたが、判別分析は分けた群がそれぞれ多変量正規分布し、それぞれが等共分散であることが仮定されています。ちなみに判別分析 1.txt のデータはこれらの条件を満たしています。但し正規性に関しては、College Analysis に多変量正規分布を検定する手法が含まれていないため、それぞれの変数についての正規性で代用しています。等共分散性については「等共分散の検定」ボタンで調べることができます。その結果を図 3.6 にその結果を示します。

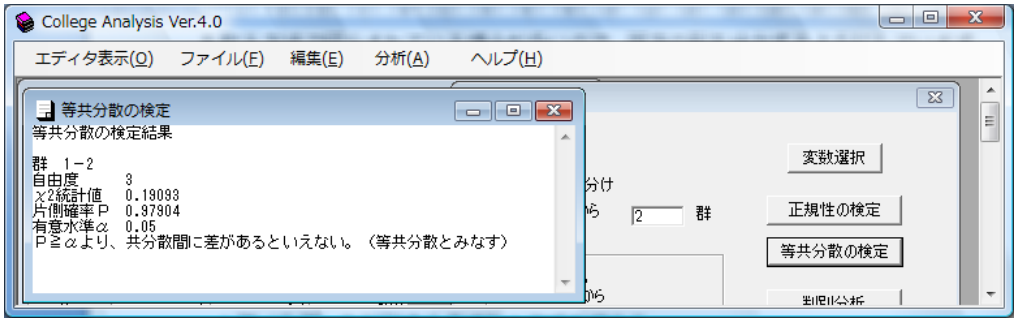


図 3.6 等共分散性の検定結果

これで係数が 0 かどうかの検定と理論的な誤判別損失の値とが安心して利用できます。しかし、これらの条件を満たしていなければ判別分析は使えないかというと、そうではなさそうです。上で述べたあやめのデータは、正規性も等共分散性も満たしていませんが、判別の精度は抜群です。判別分析の利用可能性は誤判別確率がカギになるようです。ただその際の係数の検定や理論的な誤判別確率の値はあまり信用できないと思わなければならないでしょう。

以下に判別分析の利用法をまとめておきましたので、参考にして下さい。

判別分析の目的　　２群（多群）を判別する最適な１次式を求める。

２群の場合　　判別得点＝b_1 勉強時間＋b_2 平均点＋b_0　　判別関数

判別の分点 0 より大きいか小さいかで１群と２群を分ける

２群以上の場合　判別得点＝b_1 勉強時間＋b_2 平均点＋b_0－判別の分点

判別得点が最大となる群に属すると判定する。

判別分析が有効に利用できる条件は？

→　正規性・等共分散性（等共分散の検定）

判別関数の係数は？　→　判別関数の欄

判別関数で群を分けるのは？

→　判別の分点 0（多群の場合値が最大の群）

判定に影響を与える変数は？　→　標準化係数の絶対値の大きい変数

各係数の有効性は？（要正規性・等共分散性）

→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　→　誤判別確率（実測と理論）（理論値は要正規性・等共分散性）

マハラノビス距離とは　→　どの程度 2 群が離れているかを表わす指標

マハラノビス距離	1	4	9	16	25
誤判別確率	0.309	0.159	0.067	0.023	0.006

データ毎の判別関数の値と判別状況　　→　判別得点

事象の生起確率とは？→ 合格・不合格の現れる確率が大きく異なっている場合の措置

各群同じかデータ数からが実用的

誤判別損失とは？→ 間違った判断をした場合の致命傷の程度

大きな差がない限り、各群１とするのが実用的

最後に判別分析 1.txt のデータを使った上の例を簡単な文章にまとめておきましょう。

まとめ

正規性の検定から、２群とも正規性があるとみなされ、等共分散の検定でも共分散に差があるとは言えなかった。以上から判別分析が適用可能であると判断した。

２群の生起確率を同じとし、誤判別損失を等しいとすると、判別分析によって、以下の判別関数が得られた。

y=2.2461*勉強時間+0.2007*平均点-23.0187

データはこの判別関数の値をもとに、判別の分点を 0 として、2 群に分けられる。

係数の有効性の検定では、勉強時間が p=0.00013、平均点が p=0.00061 のように、両方とも有意に 0 でないことが示された。このことから２つの変数とも有効であると思われる。

マハラノビス距離 5.6823 から、理論的な誤判別確率として p=0.117 が予想される。また、実際に判定を行うと、1 群を 2 群と間違える割合が 7.7%、その逆が 5.9%となる。これらの数値から、判別はかなりうまく行われたものと思われる。

ここで利用した理論は以下の通りです。

理論

群 1			群 2		
変数 1	…	変数 k	変数 1	…	変数 k
x^1_{11}	…	X^1_{k1}	x^2_{11}	…	x^2_{k1}
x^1_{12}	…	X^1_{k2}	x^2_{12}	…	x^2_{k2}
	…	⋮	⋮	…	⋮
$x^1_{1n_1}$	…	$x^1_{kn_1}$	$x^2_{1n_2}$	…	$x^2_{kn_2}$

判別分析の実行可能条件

分布が多変量正規分布

2 群の共分散が等しい

判別式

$$z = {}^t\mathbf{x}\mathbf{S}^{-1}(\mathbf{m}^{(1)} - \mathbf{m}^{(2)}) - \frac{1}{2}{}^t(\mathbf{m}^{(1)} + \mathbf{m}^{(2)})\mathbf{S}^{-1}(\mathbf{m}^{(1)} - \mathbf{m}^{(2)})$$

$$= b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

$$\mathbf{m}^{(a)} = \frac{1}{n_a}\sum_{\lambda=1}^{n_a}\mathbf{x}_\lambda^a \quad : \text{群 } a \text{ の各変数の平均}$$

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}\sum_{a=1}^{2}\sum_{\lambda=1}^{n_a}(\mathbf{x}_\lambda^a - \mathbf{m}^{(a)})^t(\mathbf{x}_\lambda^a - \mathbf{m}^{(a)}) \quad : \text{共分散行列}$$

判別方法

群 j を群 i と間違える損失　C_{ij}

群 i の要素が出現する確率　P_i

1 群に属する：$z - \log_e h \geq 0$

2 群に属する：$z - \log_e h < 0$　　$h = C_{12}P_2 / C_{21}P_1$

z の確率分布

x が群 1 に属する場合　　$N(D^2/2, D^2)$

x が群 2 に属する場合　　$N(-D^2/2, D^2)$

$D^2 = {}^t(\mathbf{m}^{(1)} - \mathbf{m}^{(2)})\mathbf{S}^{-1}(\mathbf{m}^{(1)} - \mathbf{m}^{(2)})$ ：マハラノビスの距離

誤判別の理論確率

群 1 を群 2 と誤判別　　$P_{21} = Z\left(\dfrac{\log_e h - D^2/2}{D}\right)$　網掛け部分

群 2 を群 1 と誤判別　　$P_{12} = 1 - Z\left(\dfrac{\log_e h + D^2/2}{D}\right)$

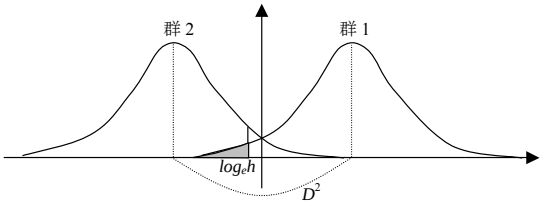


図　誤判別確率

問題 1

Samples¥判別分析 2.txt は、適性の有無の判定（有：1，無：2）と適性検査の結果と SPI の結果を与えたデータである。判定を適性検査と SPI で予測する判別分析を行い、結果を上のまとめにならって記述せよ。

問題２

Samples¥判別分析 2.txt は、適性の有無の判定（有：1，無：2）と適性検査の結果とＳＰＩの結果を与えたデータである。判定を適性検査とＳＰＩで予測する判別分析を行い、以下の問いに答えよ。但し、事象の生起確率はデータ数から、誤判別損失は２群とも１とすること。

１）このデータに判別分析は利用可能か？

正規性の検定　　正規性があると［みなす・いえない］

等共分散性　　検定確率［　　］，等共分散と［みなす・いえない］

判別分析は効率よく利用可能か。［利用可能・要注意］

２）判別関数を求めよ。

判別得点＝［　　］適性検査＋［　　］ＳＰＩ＋［　　］

３）どちらの変数が判定に影響があると思われるか。［適性検査・ＳＰＩ］

４）実測値から求めた誤判別の確率は？

適性有りを無しと［　　］　適性無しを有りと［　　］

５）厳選して新入社員を取ろうとする場合、上の誤判別でどちらの場合の損失が大きいと思われるか。［適性有りを無し・適正無しを有り］と誤判別する場合

６）上の方針に従って、大きな誤判別損失の値を 2、小さな誤判別損失の値を 1 としたとき、実測値から見た誤判別の確率はどうなるか。

適性有りを無しと［　　］　適性無しを有りと［　　］

７）上の方針で見ると、結果は改善されたか。［改善された・改善されていない］

８）誤判別損失を元に戻して、先頭（1 番）の人の判別得点はいくらか。［　　］

９）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　　判別得点［　　］　適性［有り・無し］

問題３

Samples¥判別分析 3.txt はあやめの種類をがくの長さと幅、花弁の長さと幅で３群に分類したデータである。あやめの群を他の変数の１次式で判別する３群以上の判別分析を行い、以下の問題に答えよ。

1）３つの判別得点の式を求めよ。

判別得点１＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点２＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点３＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

2）実測値から求めた誤判別確率はいくらか。

群１を他と［　　］群２を他と［　　］群３を他と［　　］

3）先頭のデータの３つの判別得点を求めよ。

判別得点１［　　］判別得点２［　　］判別得点３［　　］

4）がくの長さ 4.9、がくの幅 3.4、花弁の長さ 1.2、花弁の幅 0.3 のデータはどれに判定されるか。またそのときの最大の判別得点はいくつか。

判定［群１・群２・群３］　最大判別得点［　　］

5）もう１度 Samples¥判別分析 2.txt のデータを用いて、２群の方法の判別関数と３群以上の方法の判別関数の関係を考えよ。

２群の方法の係数は３群以上の方法の係数の［　　］になっている。