VII-2. 教師あり学習としての判別分析

VII-2-1. ロジスティック回帰とシグモイド関数（等分散性からの解放）

VII-2-1-1. 距離と比率

表 55 のような体長と体重のデータがあったとします。異なる人種を含んでいて、データの分布を人種 A（赤)、B（青)、C（緑） ごとに色分けして図 97 に示しました。また、身長と体重の関係を表す回帰式を同じ色の直線として図に示しました。この直線が人種ごとの特徴を表しています。これは単回帰分析です。図 97 に示した散布図のグループ間に境界線を引いて、新たに得られた体格データが散布図のどこに位置するかで、それがどのグループに属するか（どの人種か）を判断できるようにすることを判別分析といいます。判別分析にはいろいろな方法があります。「VI-1-3.線形判別分析」では、境界となる平面の傾きを最適化して、平面を平行移動するという方法を、線形代数学的な説明とともに紹介しました。

表 55.異なる人種の体長（身長）と体重

A		B		C	
身長	体重	身長	体重	身長	体重
160.6	44.8	182.2	109.4	175.9	44.1
163.1	63.6	186.1	141.9	193.9	59.5
153.6	58.3	185.9	147.6	190.3	58.5
151.7	54.6	152.0	92.6	179.2	48.4
171.1	68.2	202.5	216.9	165.4	34.2
164.0	58.1	203.3	227.2	171.3	43.8
167.8	69.0	164.9	105.4	181.3	47.4
161.3	65.0	149.8	83.0	194.9	65.2
145.6	40.9	176.7	145.5	201.3	65.7
173.9	69.1	152.5	79.2	200.0	58.6
166.4	68.4	180.7	138.8	183.5	45.8
165.5	73.8	164.0	102.8	200.3	62.0
181.0	64.2	164.6	95.8	183.9	48.4
159.9	63.3	194.3	144.1	179.0	48.6
153.6	49.1	144.8	78.0	210.4	76.7
166.5	55.9	165.4	106.6	190.6	54.0
153.6	50.1	178.8	122.7	168.5	39.1
153.3	52.0			210.8	64.5
170.7	71.5			200.4	62.1
165.0	67.3			189.0	58.5
158.9	55.2			192.9	57.4
160.8	64.8			177.6	46.5
151.3	44.2			162.2	35.1
176.8	68.4			190.5	52.1
148.2	51.8				
168.3	74.8				
164.3	55.0				
172.8	69.1				
176.8	70.7				
173.7	78.0				

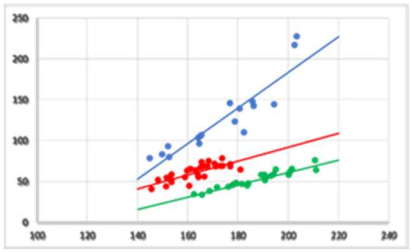


図 97. 体長と体重の散布図

表 56. 各人種のデータの分散

人種	分散		自由度
	体長	体重	
A	83.9	97.03	29
B	307.95	1769.9	18
C	177.94	110.42	26

線形判別分析は、個々のデータから境界平面の法線へのデータの写像が等分散（境界平面からの距離が等分散）であること前提にしています。しかし、この前提は常に成り立つわけではありません。表 56 に表 55 のデータの分散を示しました。分子自由度 18、分母自由度 29 の p=0.0 5 での F 比の臨界値は 1.94 です。分子自由度 18、分母自由度 26 の p=0.0 5 での F 比の臨界値は 2.05 です。A と B の体長、A と B、および B と C の体重の分散は明らかに等分散ではないでしょう。したがって、等分散を前提とした線形の判別分析をこのデータに使うことは不適切です。ここでは、等分散性という距離に依存した考え方を離れて、様々な場面に対応可能な汎用性が広い判別方法を考えなくてはならないでしょう。

１次元の方が考えやすいので、表 55 の人種 A と人種 B の体重による判別を例にします。こういう場合、普通の人は、体重を使ってデータを順番に並べ替えてみるでしょう。表 57 は、人種 A と B のデータを混合して、体重の軽い方から順位をつけて並べ替えた表です。表を見ただけで、順位 30 位と 31 位の間、78kg あたりに境界がありそうだ思うでしょう。これは直観的な判断ですが、それなりの根拠があり非論理的な判断とは言えないでしょう。

例えば、この表でのデータのもととなった人を、もう一度 47 人全員を集めて、体重だけ測って、どちらの人種に属するかを判断するということを考えます。もし、判別基準値を、体重 40.9kg より軽い値にして、それより重い人を人種 B と判定することにすると、47 人全員が B と判断されて、そのうち本当に B であるのは１ 7 人だから、正答率は 17/47=0.361 です。40.9kg と 4.2kg の間にすると、正答率は(17+1)/47=0.383 になりま

表 57. 人種 A と B の体重データ

順位	体重	人種	順位	体重	人種	順位	体重	人種	順位	体重	人種
1	40.94841	A	13	58.28133	A	25	69.12638	A	37	105.3962	B
2	44.23636	A	14	63.32097	A	26	70.65811	A	38	106.6235	B
3	44.82035	A	15	63.6093	A	27	71.49387	A	39	109.4391	B
4	49.11976	A	16	64.22131	A	28	73.77345	A	40	122.734	B
5	50.13258	A	17	64.77768	A	29	74.75967	A	41	138.8345	B
6	51.77903	A	18	64.97124	A	30	77.9867	B	42	141.8961	B
7	51.98684	A	19	67.25768	A	31	78.01168	A	43	144.0872	B
8	54.56979	A	20	68.17206	A	32	79.22857	B	44	145.5228	B
9	55.04781	A	21	68.41507	A	33	83.00537	B	45	147.6327	B
10	55.22369	A	22	68.43142	A	34	92.57853	B	46	216.8869	B
11	55.91289	A	23	69.01207	A	35	95.83666	B	47	227.2248	B
12	58.07964	A	24	69.09698	A	36	102.8453	B			

す。重い方に注目して、227.2kg 以上に境界を置いて、それ以上軽い人を A と判断すると全員が A と判断されて、そのうち本当に A なのは 30 人だから、正答率は 30/47=0.638 です。227.2kg と 216.9kg の間に置くと、一人が B と判断され、これは正答で、残り４６人が A と判断されますが、そのうち３０人が実際には人種 A だから、正答率は (1+30)/47=0.60 となります。この方法で正答率を体重に対してプロットすると、図 98 の青い折れ線が描かれます。この図から、体重 74.8 から 78.0 にピークがあり、この辺りに妥当な境界があることがうかがえます。しかし、データがあって人種だってわかっているのだから、それを改めて判別するなどということはあり得ません。人種が未知の人の新しいデータについて、データから人種を判断するというのが普通の使われ方でしょう。その場合、判定しようとするグループ内の人種の比がいつも 30 対 17 ということはないでしょう。例えば、宇宙人がランダムに地球人を捕まえて、人種を判断するのであれば、世界人口７０億で中国の人口は 14 億だから、なにもしなくても中国人だと判定すれば２割は正解となります。ここで考えている判別は、そういう状況を前提にしていません。確率２分の１でどちらかわからないという条件で考えています。そうすると、B の人種と同数の A

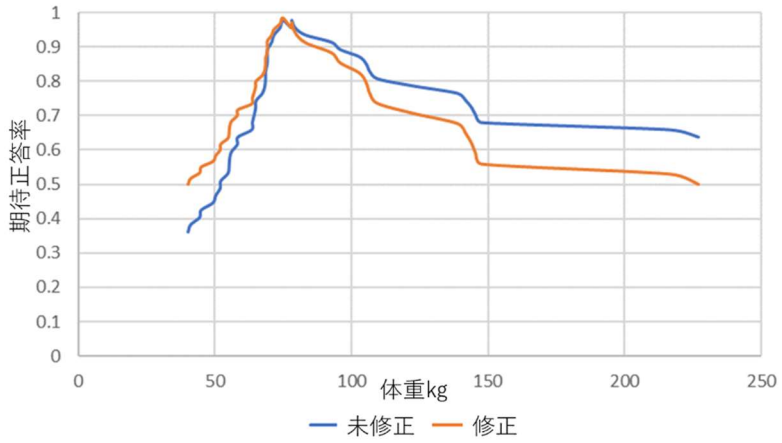


図 98.境界体重による期待正答率の変化

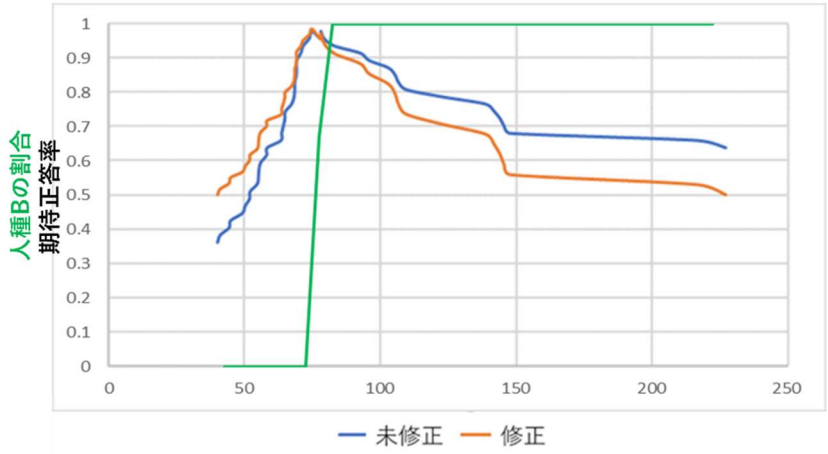


図 99. 体重と人種 B の割合との関係

表 58. 人種 A、B の体重データと確率

順位	w	t	確率	順位	w	t	確率	順位	w	t	確率	順位	w	t	確率
1	40.9	1	$1-p_n$	13	58.3	1	$1-p_n$	25	69.1	1	$1-p_n$	37	105.4	0	$1-p_n$
2	44.2	1	$1-p_n$	14	63.3	1	$1-p_n$	26	70.7	1	$1-p_n$	38	106.6	0	p_n
3	44.8	1	$1-p_n$	15	63.6	1	$1-p_n$	27	71.5	1	$1-p_n$	39	109.4	0	p_n
4	49.1	1	$1-p_n$	16	64.2	1	$1-p_n$	28	73.8	1	$1-p_n$	40	122.7	0	p_n
5	50.1	1	$1-p_n$	17	64.8	1	$1-p_n$	29	74.8	1	$1-p_n$	41	138.8	0	p_n
6	51.8	1	$1-p_n$	18	65.0	1	$1-p_n$	30	78.0	0	p_n	42	141.9	0	p_n
7	52.0	1	$1-p_n$	19	67.3	1	$1-p_n$	31	78.0	1	$1-p_n$	43	144.1	0	p_n
8	54.6	1	$1-p_n$	20	68.2	1	$1-p_n$	32	79.2	0	p_n	44	145.5	0	p_n
9	55.0	1	$1-p_n$	21	68.4	1	$1-p_n$	33	83.0	0	p_n	45	147.6	0	p_n
10	55.2	1	$1-p_n$	22	68.4	1	$1-p_n$	34	92.6	0	p_n	46	216.9	0	p_n
11	55.9	1	$1-p_n$	23	69.0	1	$1-p_n$	35	95.8	0	p_n	47	227.2	0	p_n
12	58.1	1	$1-p_n$	24	69.1	1	$1-p_n$	36	102.8	0	p_n				

の人種がいるという前提で考えることになります、B の一人分を、30/17 倍する必要があります。そのような条件で正答率の期待値をプロットしたのが図 98 の赤いラインです。このプロットでも、74.9kg にピークがあります。図 98 は非対称です。これは、人種 A と人種 B の分散が大きく異なるためです。このような考え方を発展させて、もう少し数学的に確率がどのように変化するのかを考えます。体重 40kg から 230kg まで、5kg ごとに人種 B の割合を調べると、図 99 の緑の線で示した折れ線が得られます。緑の線を S 字曲線と考えると、75kg 前後に変曲点があると推定できますが、これは正答率が最も高くなる境界体重にほぼ一致しています。表 58 も体重のデータを小さい方から順位をつけて並べたものですが、列 t は人種を表す列で、人種 A であれば、$t=1$、人種 B すなわち人種 A であれば$t=0$と表現しています。w_nであったときに$t=1$であるという条件付き確率を

$$\mathrm{P}(1|w_n)=1-p_n$$

データがw_nであったときに人種 A でないことの、という条件付きの確率は、

$$\mathrm{P}(0|w_i)=1-(1-p_n)=p_n$$

となります。これをクロネッカーのデルタを使って、場合分けをしないで書くと、

$$\mathrm{P}(\delta_{ij}|w_n)={p_n}^{1-\delta_{ij}}(1-p_n)^{\delta_{ij}}$$

（クロネッカーのデルタは$\delta_{ij}=\begin{cases}1\,(i=j)\\0\,(i\neq j)\end{cases}$はとなる二値変数、$n$は下から数えた順位、$i$は判別関数で識別して取り出そうとする人種（この場合は人種 A）、 j はそのデータの実際の人種）

となります。

$${p_n}^{1-1}=1、(1-p_n)^1=1-p_n、{p_n}^{1-0}=p_n、(1-p_n)^0=1$$

だから、

$$\mathrm{P}(0|w_n)=p_n$$

$$\mathrm{P}(1|w_n)=1-p_n$$

t は排反事象の 2 値変数で、クロネッカーのデルタそのものだから、δ_{ij}を t と書き換えて

$$\mathrm{P}(t|w_n)={p_n}^{1-t}(1-p_n)^t$$

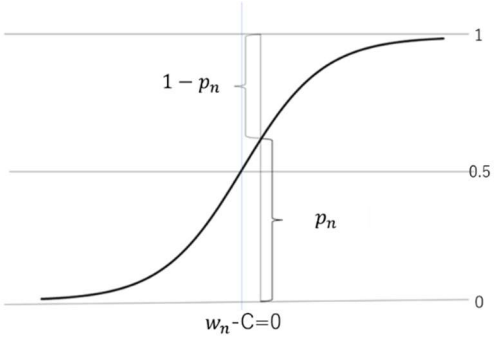


図 100. S 字曲線（例えばシグモイド曲線）と確率

となります。

p_nは累積確率ですから S 字曲線になりますが、S 字曲線（図 100）から上の長さが、変数w_n以下の場合に人種 A が含まれる割合ですが、これは、変数w_nが得られた時に人種が A である条件付き確率（尤度）でもあります。1 次元ですから判別境界は点になります。この点を$p_n = 0.5$の点にとって、その点の体重を定数 C とすれば体重を境界体重からの差として表すことが出来ます。境界体重からの差によって、確率p_nが決まるということです。p_n=$\varsigma(w_n - C)$となる関数を最尤法的に決めればよいということになります。関数$\varsigma(w)$の形を考えなければ、図 100 の横方向の縮尺を伸ばしたり縮めたりして、実際のデータの当てはまりをよくすることを考えれば良いだけでから、p_n=$\varsigma\big(a_1(w_n - C)\big)$と表して、

D= $a_1 w_n - a_1$C= $a_0 + a_1 w_n$　の$\boldsymbol{a} = (a_0 \quad a_1)$を最適化することになります。S 字曲線がどんな形をしているのかなど一般的にはわからないし、具体的な事例に当てはまりそうな S 字曲線を探すなどという神経質な作業をしても、そんなに予測精度が上がるとも思えないから、関数$\varsigma(w)$は単調増加する S 字曲線ならば何でも良いような気もします。代表的な S 字曲線としてすぐに思いつくのは、累積正規分布関数とシグモイド関数です。正規分布にこだわるならば累積正規分布関数を使うことも考えられます。いずれにしても、実際のデータ分布に合わせて確率を考えることによって等分散性の縛りを抜け出せます。せっかく正規分布から遠ざかったのに、また正規分布にこだわるのは少し残念です。正規分布関数をモデルとした判別分析をプロビット回帰といいますが、より一般的なのは、シグモイド関数を使ったロジスティック回帰（ロジット分析）です。次の項では、一旦、具体的な事例を離れて、シグモイド関数について説明します。

VII-2-1-2.　シグモイド関数

確率密度分布関数としては正規分布関数が良く知られていますが、正規分布関数以外にも確率水戸度関数として使える関数がたくさんあります。確率関数が持つべき必要条件は、1. 起こり得る確率の総和は 1 だから、マイナス無限大から無限大の積分値が 1 になること、2. 説明変数の大小関係が、関数によって得られる被説明変数の大小関係に維持されなければならないから、その積分関数が単調増加関数であること、この二つです。代表的なのは正規

分布関数ですから、まず正規分布について復習しておきます。
以下の形の関数をガウス関数といいます。

$$f(x) = ae^{\left(-\frac{(x-b)^2}{2C^2}\right)}$$

正規分布関数は、ガウス関数の特殊な例です。正規分布関数は以下の式で書き表せます。

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

$a = 1, b = 0, c = 1$の時、ガウス関数の無限積分は以下の通りになります。

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

$$\int_{-\infty}^{\infty} ae^{\left(-\frac{(x-b)^2}{2C^2}\right)} dx = a\sqrt{2C^2\pi}$$

この無限積分をガウス積分といいます。ガウス積分の求め方はいくつかありますが、わかりやすいのは、2 乗の形にして極座標変換する方法です。この方法、正規分布を誘導するときに説明したので、興味があればそちらを参考にしてください。いずれにしても、

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx = \frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi\sigma^2}} = 1$$

です。単調増加であることは式の形から明らかですから、正規分布関数は確率密度関数として使えます。
次にシグモイド関数について説明します。以下の関数をシグモイド関数といいます。

$$\varsigma_a(x) = \frac{1}{1+e^{-ax}} \qquad \text{式 81}$$

特に、$a = 1$の時のシグモイド関数を標準シグモイド関数といいます。

$$\varsigma_1(x) = \frac{1}{1+e^{-x}} \qquad \text{式 82}$$

シグモイド関数は便利な関数で様々な分野で使われています。数学的な解説の前に、水産学で使われている例として Russel のモデルを紹介します。具体的な例を挙げた方がわかりやすいと思うからでが、それ以上に、もう５０年ぐらい前に水産資源学でこのモデルを習ったのがシグモイド関数との最初の出会いだったという筆者の個人的な思い出があるからです。Russel のモデルは、水産資源の増殖カーブとして使われるロジスティック曲線を作るモデルです。水産資源量とは漁獲対象になるサイズの水産物の量だと思ってください。水産資源の量は、そのサイズの水産物として資源に加入する水産物の量と加入後の成長によって増え、自然死亡と漁獲によって減ると考えます。

$$\Delta B = R + G - V - Y$$

ΔB: *ある*漁期からある漁期までの資源量の変化
R:期間中の加入量
G: *資源全体の成長量*
V: *自然死亡量*
Y: *漁獲量*

ここから個体群の増殖曲線を作ります。まず、最も簡単な増殖モデルとして、環境制約がな

く無限に増殖する場合を考えます。もちろん、そんなことは絶対にあり得ませんが、環境制約がなく無限に個体数が増加すると考えれば、単位時間当たりの増加量は個体数だけに依存しまから，単位時間当たりの個体数の増加は次式になります。(rは内的自然増加率：多くのモデルでこれを一定と仮定している。)

$$\Delta B = rB\Delta t$$

したがって、

$$\frac{\Delta B}{\Delta t} = rB$$

これを偏微分方程式にすると、

$$\frac{\partial B}{\partial t} = rB$$

$$B\partial B = r\partial t$$

$$\log_e B = rt + C$$

$$C: constant$$

$$B = e^{rt+C}$$

$$B = e^C e^{rt}$$

定数 C について，

$t = 0$の時のB、すなわち、B_0を考えると、

$$B_{\,0} = e^C e^{r0} = e^C \cdot 1 = e^C$$

したがって、

$$B = B_0 e^{rt}$$

この式では個体群は無限に大きくなることになります。実際には、代謝に必要な物質の量や排せつ物の蓄積などの制約から環境抵抗が生まれ、それが個体数の増加とともに増加していくために環境収容力に限界が出来ます。制約がなければ、空間はあっという間に細菌などに満たされてしまうでしょう。環境制約がある場合には以下のモデルが一般的です。

個体群の増殖曲線（環境制約がある場合）

$$\frac{\partial B}{\partial t} = r\left(1 - \frac{B}{K}\right)B$$

このモデルを変形して、モデルの意味を解説します。

$$\frac{\partial B}{\partial t} = \frac{r}{K}(K - B)B \qquad \text{i}$$

$$\frac{\partial B}{\partial t} = -\frac{r}{K}(B^2 - KB)$$

$$\frac{\partial B}{\partial t} = -\frac{r}{K}\left\{\left(B - \frac{K}{2}\right)^2 - \frac{K^2}{4}\right\} \qquad \text{ii}$$

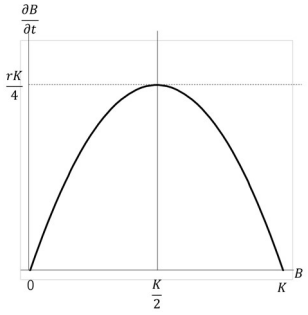


図 101. $\frac{\partial B}{\partial t} = -\frac{r}{K}\left\{\left(B - \frac{K}{2}\right)^2 - \frac{K^2}{4}\right\}$の軌跡

捕食や漁獲などの外的な減耗要因がなければ増加量$\frac{\partial B}{\partial t}$は負の値を取りません。式 i の変形により、B は 0 から K の値で変動するモデルであり、K は B の最大値B_{max}であることがわかります。また式 ii より。図 101 に示したように、このモデルは上に凸の二次式であり、$B = \frac{B_{max}}{2}$の時に、最大の増殖速度、$\frac{rK}{4}$となります。この点が極値です。
時間 t に対する B の変化をさらに明瞭にするために、$B = f(t)$の積分形にします。
式 i より

$$\frac{\partial B}{\partial t} = \frac{r}{K}(K - B)B$$

$$\frac{1}{(K - B)B}\partial B = \frac{r}{K}\partial t$$

$$\frac{1}{K}\left\{\frac{1}{B} - \frac{1}{K - B}\right\}\partial B = \frac{r}{K}\partial t$$

$$\left\{\frac{1}{B} - \frac{1}{K - B}\right\}\partial B = r\partial t$$

$$\log_e B - \log_e(K - B) = rt + \mathrm{C}$$

$$\log_e \frac{B}{(K - B)} = rt + \mathrm{C}$$

積分定数 C を決めます。図 101 に示したように、この関数は$\frac{B_{max}}{2} = \frac{K}{2}$となる現時点で極値となります。この点を$t = 0$として、C を求めます。

$$\log_e \frac{\frac{K}{2}}{\left(K - \frac{K}{2}\right)} = r \cdot 0 + \mathrm{C}$$

$$\log_e 1 = \mathrm{C}$$

$$C = 0$$

したがって

$$\log_e \frac{B}{(K - B)} = rt$$

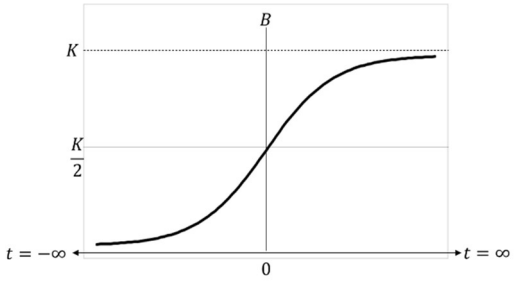


図 102. ロジスティック曲線

これをBについて解くと

$$\frac{B}{(K-B)} = e^{rt}$$

$$B = Ke^{rt} - Be^{rt}$$

$$B(1+e^{rt}) = Ke^{rt}$$

$$B = \frac{Ke^{rt}}{1+e^{rt}}$$

となりますが、この表現は一般的ではありません。一般に使われるのは、分母分子をe^{rt}で割って、分母分子の意味をより明瞭に表現した以下の式です(図 102 参照)。

$$B = \frac{K}{1+e^{-r}}$$

この式で K=1 としたものがシグモイド関数です。この説明で分かるように、シグモイド関数は、

$$\frac{\partial f(x)}{\partial x} = f(x)\big(1-f(x)\big)$$

となる関数です。

K=1 ならば、

$$\varsigma_a(x) = \frac{1}{1+e^{-a}}$$

$$x \to -\infty \quad \varsigma_a(x) = 0$$

$$x \to \infty \quad \varsigma_a(x) = 1$$

となります。

シグモイド関数は単調増加であり、その極限値は 1 です。「関数の積分が単調増加で、マイナス無限大から無限大の積分値が 1 になるものは確率密度関数として使える。」のですから、シグモイド関数の導関数は確率密度関数として使えますし、シグモイド関数も累積確率密度関数として使えます。$\boldsymbol{x}$が$\boldsymbol{x}_{limit}$よりも小さければ XX だというのは累積確率的な表現ですから、シグモイド関数は判別分析でしばしば使われます。$\boldsymbol{x}$が与えられたときに、XX である確率を次の式で表します。

$$P(\mathrm{XX}|\boldsymbol{x})$$

これが標準シグモイド関数になるのならば、

$$P(\mathrm{XX}|\boldsymbol{x}) = \varsigma(x) = \frac{1}{1+e^{-x}}$$

となります。念のために、これを微分してみます。

入れ子型の合成関数の微分の公式、$\frac{\partial f(x)}{\partial x} = \frac{\partial f(x)}{\partial g(x)} \cdot \frac{\partial g(x)}{\partial x}$を使って、

$$\varsigma(x) = \frac{1}{1+e^{-x}}$$

$$f(x) = \frac{1}{g(x)}$$

$$g(x) = 1 + e^{-x}$$

$$\frac{\partial f(x)}{\partial g(x)} = -\frac{1}{\big(g(x)\big)^2} = -\frac{1}{(1+e^{-x})^2}$$

$$\frac{\partial g(x)}{\partial x} = -e^{-x}$$

$$\frac{\partial f(x)}{\partial x} = \frac{\partial f(x)}{\partial g(x)} \cdot \frac{\partial g(x)}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2}$$

部分分数に分解して単純化します。

$$\frac{e^{-x}}{(1+e^{-x})^2} = \frac{1(1+e^{-x}-1)}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})} \cdot \frac{(1+e^{-x})-1}{(1+e^{-x})}$$

$$= \frac{1}{(1+e^{-x})} \cdot \left\{1 - \frac{1}{(1+e^{-x})}\right\}$$

ここで

$$\frac{1}{(1+e^{-x})} = f(x)$$

ですから、確かに

$$\frac{\partial f(x)}{\partial x} = \frac{1}{(1+e^{-x})} \cdot \left\{1 - \frac{1}{(1+e^{-x})}\right\} = f(x)\big(1-f(x)\big)$$

となります。

確率の総和は 1 ですから、$1 - P(A|x)$はその排反事象、xであったときにAが起こらない確率です。

$$\frac{\partial f(x)}{\partial x} = P(A|x)\, P(not\ A|x)$$

となります。$P(A|x)$が増加すれば、その分だけ$P(not\ A|x)$が減るのだから、計算するまでもなく$P(not\ A|x)$の微分は直観的に

$$\frac{\partial P(not\ A|x)}{\partial x} = -P(A|x)P(not\ A|x)$$

となります。

次にシグモイド関数の逆関数を考えてみます。

$$P(A|x) = \frac{1}{1+e^{-x}}$$

の式を$x = f\big(P(A|x)\big)$の形に書き換えるということです。

$$1 + e^{-x} = \frac{1}{P(A|x)}$$

$$e^{-x} = \frac{1}{P(A|x)} - 1 = \frac{1 - P(A|x)}{P(A|x)}$$

$$-x = \log_e \frac{1 - P(A|x)}{P(A|x)}$$

$$x = \log_e \frac{P(A|x)}{1 - P(A|x)} = \log_e P(A|x) - \log_e\big(1 - P(A|x)\big)$$

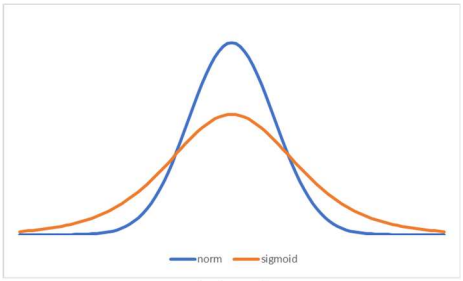
ところで、$1 - P(A|x) = P(not\ A|x)$ですから、

$$x = \log_e \frac{P(A|x)}{P(not\ A|x)}$$

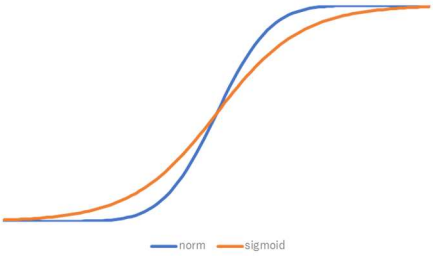
この関数をロジット関数といい、$\log_e \frac{P(A|x)}{P(not\ A|x)}$をロジットと呼びます。また、あることが起こる確率と起こらない確率の比$\frac{P(A|x)}{P(not\ A|x)}$をオッズ比と言います。このような言い方をする場合には、ロジット関数と対比させて、元のシグモイド関数をロジスティック関数と呼ぶのが一般的です。

VII-2-1-3. シグモイド関数と正規分布（ガウス関数）の比較

図 103 に正規分布とシグモイド関数、その積分の累積確率密度関数を示しました。2 つの曲線を比べると、確率密度曲線が左右対称の釣り鐘型の凸曲線であること、累積確率曲線が$\left(0 \quad \frac{1}{2}\right)$に変曲点を持つ S 字曲線であることなどよく似ています。正規分布関数は二項分布のテイラー展開だから、理論的背景を持った確率密度関数としてすぐに思いつきますが、正規分布だけがすべてに適用可能な唯一絶対の正しい唯一の確率分布ではありません。統計解析の実務でも、明らかに正規分布ではない現象については、ノンパラメトリックな統計解析法が用いられます。明らかに正規分布ではないという統計的な判断は可能ですが、正しい正規分布と差がないという、ないことの証明は悪魔の証明ですから、絶対に正しく正規分布だという証明はありません。実際、ほぼ完璧に正規分布する現象の方が稀です。正規分布が統計解析的に扱いやすい関数であり、正規分布を仮定して解析を行い、



確率分布



累積確率分布

図 103. 正規分布（ガウス関数）とシグモイド関数の比較

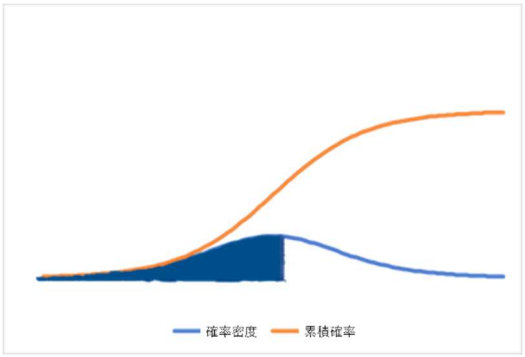


図 104. S 字曲線とその微分

それで問題がない場合が多いから、多くの場合使われるということに過ぎません。シグモイド関数の導関数は正規分布関数よりなだらかな凸関数です。むしろこのように広がりを持って分布する現象の方が多いような気がします。いずれにしても、正規分布にこだわる必要もなく、それで良い結果が得られるのならば、シグモイド関数の導関数のように母集団が分布することを仮定することは無理のあることではありません。シグモイド関数は、様々な分野で判別分析やクラスター分析等の確率として使われています。ある値を境界に「そうであるか、あるいは、そうでないか」という排反事象の確率はその境界まで（あるいは境界以降）の累積確率になるからです。例えば、致死的な負傷した患者が、救急病院に運ばれるまでの時間 t と致死率の関係は、ある時点で致死率が増加しますが、致死率 1 に近づくと、致死率の増加速度が低下する S 字曲線になるでしょう。図 104 に S 字曲線とその微分関数を示しました。青線が赤線の S 字曲線の導関数で確率密度分布関数で、赤線の S 字曲線は青い色で示した部分の面積（累積密度関数）という関係になります。シグモ

イド関数とその導関数もこの図のような関係にあり、シグモイド関数は青の部分の面積がある境界値以下である確率の総和です。

VII-2-1-4. 勾配降下法を使った１次元２クラス判別

シグモイド関数を理解したので、S 字曲線の関数としてシグモイド関数を使うことにして、表 58 に示したデータで体重 1 変数で人種 A、B を判別すると問題に戻ります。これまでの議論からp_nは以下のように入れ子になった複合関数で表現できます。

$$\alpha_n = a_0 + a_1 w_n \qquad \text{i}$$

$$p_n = \varsigma(\alpha_n) \qquad \text{ii}$$

式 i は、$w_{n0} = 1$のダミー変数をつくると考えると、$\boldsymbol{w}_n = (w_{n0} \quad w_{n1})$となるベクトルとして表現できるので係数もベクトル化して、$\boldsymbol{a} = (a_0 \quad a_1)$と表現し、

$$\alpha_n = \sum_{d=0}^{1} a_d\, w_{nd} = \boldsymbol{a} \cdot \boldsymbol{w}_n$$

・は内積を表す

と表す方がすっきりしていてプログラミングなどではわかりやすいかもしれません。ここで最適化するのは$\boldsymbol{a} = (a_0 \quad a_1)$です。最尤法で最適化するとするのであれば、全尤度（確率の総積）を最大化することになります。尤度は

$$L(\boldsymbol{a}) = \prod_{n=1}^{N} P(t|w_n) = \prod_{n=1}^{N} {p_n}^t (1-p_n)^{1-t}$$

ですが、この尤度の値は大変大きな値になって、コンピュータ‐の計算限度を超えてしまいます。対数は大小関係が変わらずに数値を圧縮できるので、一般には対数尤度を最大化します。

$$LL(\boldsymbol{a}) = \sum_{n=1}^{N} \log_e({p_n}^t (1-p_n)^{1-t})$$

$$= \sum_{n=1}^{N} (t \log_e p_n + (1-t)\log_e(1-p_n)) \qquad \text{iii}$$

対数尤度を係数$\boldsymbol{a}$の関数として$\boldsymbol{a}$で微分することを考えます。一旦、Σ を外して中だけを考えます。

$$f_n(\boldsymbol{a}) = t(\log_e p_i) + (1-t)\log_e(1-p_i) \qquad \text{iv}$$

式 I,ii,iv は入れ子になった複合関数だから

$$\frac{\partial f_n(\boldsymbol{a})}{\partial \boldsymbol{a}} = \frac{\partial f_n(\boldsymbol{a})}{\partial p_n}\frac{\partial p_n}{\partial \alpha_n}\frac{\partial \alpha_n}{\partial \boldsymbol{a}}$$

です。

$$\frac{\partial f_n(\boldsymbol{a})}{\partial p_n} = t(\log_e p_n)' + (1-t)\,(\log_e(1-p_n))' = t\frac{1}{p_n} - (1-t)\frac{1}{1-p_n}$$

$$= \frac{t(1-p_n)-(1-t)p_n}{p_n(1-p_n)} = \frac{t-p_n}{p_n(1-p_n)}$$

$$\frac{\partial p_n}{\partial \alpha_n} = \left(\frac{1}{1+e^{-\alpha_n}}\right)' = -\frac{e^{-\alpha_n}\cdot(-1)}{(1+e^{-\alpha_n})^2} = \frac{e^{-\alpha_n}}{(1+e^{-\alpha_n})^2} = p_n(1-p_n)$$

$$\frac{\partial \alpha_n}{\partial a_0} = \frac{\partial(a_0 w_{n0} + a_1 w_{n1})}{\partial a_0} = w_{n0} = 1$$

$$\frac{\partial \alpha_n}{\partial a_1} = \frac{\partial(a_0 w_{n0} + a_1 w_{n1})}{\partial a_1} = w_{n1}$$

$$\frac{\partial f_n(a_0)}{\partial a_0} = \frac{t-p_n}{p_n(1-p_n)}\big(p_n(1-p_n)\big)\times 1 = t-p_n$$

$$\frac{\partial f_n(a_1)}{\partial a_1} = \frac{t-p_n}{p_n(1-p_n)}\big(p_i(1-p_i)\big)\times w_{n1} = (t-p_i)w_{n1}$$

このように式が作れます。これをΣ記号の中に戻して

$$\frac{\partial LL(\boldsymbol{a})}{\partial \boldsymbol{a}} = \sum_{n=1}^{N} \frac{\partial f_n(\boldsymbol{a})}{\partial \boldsymbol{a}}(t-p_i)\boldsymbol{w_n}$$

$$\frac{\partial LL(a_0)}{\partial a_0} = \sum_{n=1}^{N} \frac{\partial f_n(a_0)}{\partial a_0}(t-p_i)$$

$$\frac{\partial LL(a_1)}{\partial a_1} = \sum_{n=1}^{N} \frac{\partial f_n(a_1)}{\partial a_1}(t-p_i)w_{n1}$$

最後にこれらの微分式を0とおいて連立微分方程式を解き、最適なa_0、a_1を求めるというのが解析数学的な解法ですが、この場合は、非線形の超越関数である指数関数が間にあるために、代数的に連立方程式の解を求めることが出来ません。コンピュータの計算力を使えば、暫定的に$\boldsymbol{a}$を与えて、微分値を計算して、微分の傾斜に沿って$\boldsymbol{a}$の値を修正することを繰り返す方法で近似的に解を求めることが出来ます。その方法を勾配法と呼びます。

機械学習では、対数尤度に-1を掛けたものを、交差エントロピー誤差（cross entropy error function）と呼び、これをデータ数nで割ったものを線形の係数ベクトルに対する平均交差エントロピー誤差 （$E(\boldsymbol{a})$）と呼びます。

$$E(\boldsymbol{a}) = \frac{-1}{N}\,LL(x) = \frac{-1}{N}\sum_{n=1}^{N}\log_e P(t|x_n) = \frac{-1}{N}\sum_{n=1}^{N}(t\log_e p_n + (1-t)\log_e(1-p_n))$$

$$= \frac{1}{N}\sum_{n=1}^{N}(-t\log_e p_n + (t-1)\log_e(1-p_n))$$

式の意味は、平均的な対数尤度を計算して、これが上向きに凸の関数だから、正負を入れ替えて、下向きに凸の関数にするということです。

この関数は$\boldsymbol{a}$の関数です。$\boldsymbol{a}$の値が与えられれば、そこから、ベクトル$\boldsymbol{a}$の要素による平均エントロピー誤差の微分値が計算されます。この微分値はその要素の方向の関数の傾きです。微分値が正ならば反対方向の方向、微分値が負ならばその反対方向に、$\boldsymbol{a}$の要素の方

向に微小分だけ、$\boldsymbol{a}$の値を修正すれば、平均交差エントロピー誤差が小さくなります。下向きに凸の関数ですから、これを繰り返せば微分値が近似的に 0 となり、極値を与える$\boldsymbol{a}$が得られるはずです。これが勾配降下法です。これは長い繰り返し計算になります。また、関数がいくつかの極値を待つ場合には、全体最適的に最小値を与える$\boldsymbol{a}$に接近する前に、部分最適的な極値に落ち込んでしまい、最小値にたどり着けない場合もあります。そうしたことをできるだけ避けて効率的に勾配降下法を行う様々な工夫がありますが、ここではもっとも単純な勾配法（勾配降下法）の手順を示します。

ここで考えている 1 次元のデータの場合、最適化する係数は$a_0, a_{1,}$の二つです。$E(\boldsymbol{a})$の$\boldsymbol{a}$による微分をします。

$$E(\boldsymbol{a}, \boldsymbol{x}) = -\frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{a}, \boldsymbol{x}_n)$$

で、すでにやったように

$$\frac{\partial f_n(\boldsymbol{a})}{\partial p_n} = \frac{t - p_n}{p_n(1 - p_n)}$$

$$\frac{\partial p_n}{\partial \alpha_n} = p_n(1 - p_n)$$

$$\frac{\partial \alpha_n}{\partial \boldsymbol{a}} = \boldsymbol{w_n}$$

ですから、

$$\frac{\partial E(\boldsymbol{a})}{\partial a_0} = -\frac{1}{N}\sum_{n=1}^{N} \frac{\partial f_n(\boldsymbol{a})}{\partial p_n}\frac{\partial p_n}{\partial \alpha_n}\frac{\partial \alpha_n}{\partial a_0} = -\frac{1}{N}\sum_{n=1}^{N} \frac{t - p_n}{p_n(1 - p_n)} p_n(1 - p_n)\boldsymbol{w_n}$$

$$= \frac{1}{N}\sum_{n=1}^{N} (p_n - t_n)\,\boldsymbol{w_n}$$

となります。個々の係数別に分けて書くと次のようになります。

$$\frac{\partial E(\boldsymbol{a})}{\partial a_0} = \frac{1}{N}\sum_{n=1}^{N} (p_n - t_n)$$

$$\frac{\partial E(\boldsymbol{a})}{\partial a_1} = \frac{1}{N}\sum_{n=1}^{N} (p_n - t_n) w_{n1}$$

w_{n0}についてそれが人種 A だった場合には、t_n=1 だから、微分値は$p_n - 1$、人種が A でなければp_nになるということです。

勾配降下法では、初期値$a_{0,0}, a_{1,0}$を決めて次のように順次$a_{0,k}, a_{1,k}$を移動させ、微分値が近似的に 0 になる点を探します。

$$a_{0,k} = a_{0,k-1} - \eta\frac{\partial E(\boldsymbol{a})}{\partial a_0}$$

$$a_{1,k} = a_{1,k-1} - \eta\frac{\partial E(\boldsymbol{a})}{\partial a_0}$$

ηは移動する率であり、学習率と呼ばれています。学習率も分析者が適当な値を決めるのですが、式からわかるように、微分値が小さくなるにつれて、移動する幅も小さくなります。学習率が小さいと、計算回数が増えて計算時間が長くなりますが、あまり大きくすると極値を飛び越えて安定しません。図 105 に勾配降下法による係数の推定値の移動の様子をイメージとして描いてみました。ただこれだけでは、実際の計算方がわからないかもしれません。ここでは機械学習の例として判別分析を取り上げています。プログラミングをしてコンピュータで計算するのが前提です。プログラムを示されてもそのプログラム言語を使い慣れていない人には意味が解らないでしょう。そこで、エクセルシートを手で動かして勾配降下法で最適$a_0, a_{1,}$を求めてみました。その計算プロセスと結果を表 59 と図 106 に示しました。図 106 は$a_0 - a_{1,}$平面上での勾配降下法による点の移動の軌跡です。途中で

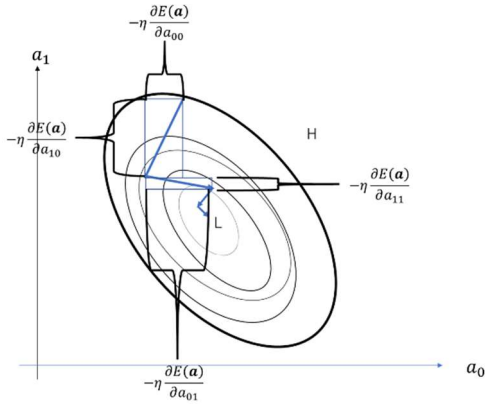


図 105. 勾配降下法による点（$a_0, a_{1,}$）の移動

表 59. 勾配降下法の計算例（体重データ）

step	A0	A1	A0微分値	A1微分値	学習率	境界体重
0	-77.5	1	0.284961	0.003796	-1	-77.5
1	-77.785	0.996204	-0.40697	-0.00508		-78.0813
2	-77.378	1.001282	0.003947	0.000188		-77.2789
3	-77.909	0.994322	-0.74155	-0.00936		-78.3539
4	-77.1674	1.00368	0.939456	0.012232		-76.8845
5	-78.1069	0.991448	-1.2554	-0.01592		-78.7806
6	-76.8515	1.007368	1.489623	0.019382	-0.5	-76.2893
7	-77.5963	0.997677	-0.0384	-0.00036		-77.7769
8	-77.5771	0.997855	0.00094	0.000149		-77.7439
9	-77.5775	0.99778	-0.00654	5.35E-05		-77.7501
10	-77.5743	0.997753	-0.00513	7.16E-05		-77.749
11	-77.5717	0.997718	-0.0054	6.82E-05		-77.7492
12	-77.569	0.997684	-0.00535	6.88E-05		-77.7491
13	-77.5663	0.997649	-0.00536	6.87E-05		-77.7491
14	-77.5637	0.997615	-0.00536	6.87E-05		-77.7491
15	-77.561	0.99758	-0.00536	6.87E-05		-77.7491
16	-77.5583	0.997546	-0.00536	6.87E-05		-77.7491
17	-77.5583	0.997546	-0.00536	6.87E-05		-77.7491

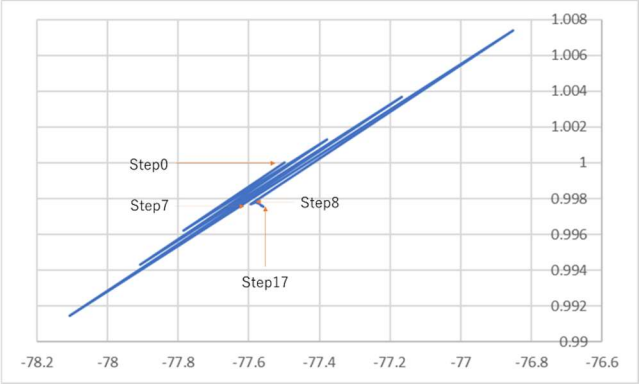


図 106. 勾配降下法による点（a_0, a_1）の移動(体重)

学習率を 1 から 0.5 に変えましたが、Step 7 までは、図の右上と左下の間で、大きく振幅しています。Step8 以後は、左上から右下に向かって、ゆっくりと移動します。このことから、右上と左下が高く Step8 と Step17 を結ぶ線のあたりに谷底の線があり、この谷底線は右下に向けてゆっくりとなだらかに下っているということがわかります。やがて勾配は 0 になります。勾配が小さいので、手計算でやっているとなかなか収束しません。傾き -0.005 を近似的に 0 として良いかどうかは微妙なところですが、著者はここで力尽きました。目的は境界値となる体重を求めることだから、それがわかれば、無理に収束させなくても良いかもしれないと言い訳をして、とりあえず収束したことにしました。境界値は、確率 1/2 なる値だから、

$$\frac{1}{2} = \frac{1}{1 + e^{-(a_0 + a_1 x)}}$$

$$2 = 1 + e^{-(a_0 + a_1 x)}$$

$$e^{-(a_0 + a_1 x)} = 1$$

$$a_0 + a_1 x = 0$$

$$x = \frac{-a_0}{a_1}$$

境界値の体重xの移動も表 62 に示しました。すでに境界値は近似的に収束しています。そこで

$$a_0 = -77.558$$

$$a_1 = 0.9975$$

を最適解としました。

図 107 に、人種 B の割合とロジスティック回帰の結果を重ね合わせて示しました。 2 つの線はほぼ一致しています。このことから、ロジスティック回帰によって、確率 2 分の一になる点を、判別の境界値とすればよいことが確認できます。

いつでもこのように容易に答えが得られるわけではありません。たとえば、体長について、同様のモデルでロジスティック回帰を行うと容易に収束しません。試行錯誤の結果、谷底線上に初期値を置くことに成功して、a_0の微分値が 0.01 以下になったのは、表 60 に

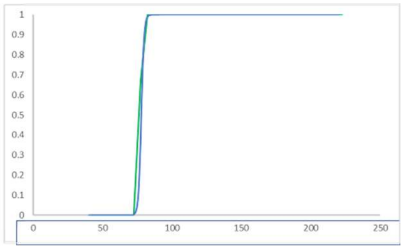


図 107. 人種 B の割合とロジスティック回帰の比較。

表 60. 勾配降下法の計算例（体長データ）

step	a0	a1	a0微分値	a1微分値
0	-35	0.2	-30.4275	-0.18955
1	-31.5347	0.221787	-2.32518	-0.01565
:	:	:	:	:
113	-28.6649	0.241192	-0.00999	-6.8E-05
114	-28.6549	0.24126	-0.00979	-6.6E-05
115	-28.6451	0.241326	-0.0096	-6.5E-05

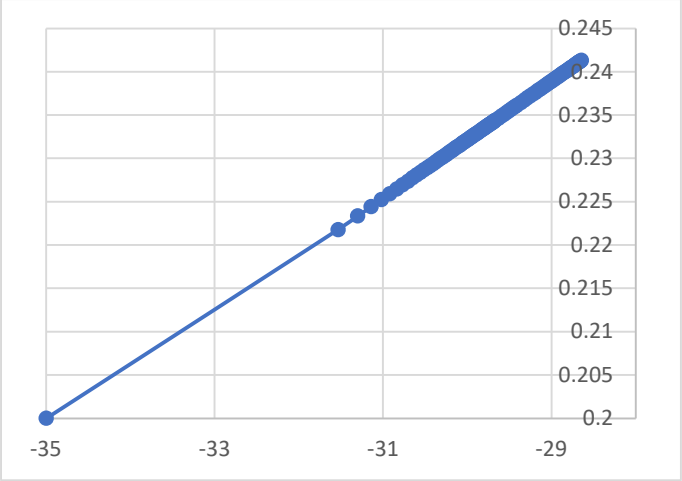


図 108.勾配降下法による点（a_0, a_1）の移動(体長)

示したように 114 ステップ目でした。それでも収束しないので。例によって、力尽きて収束したことにしました。図 108 に移動の様子を示しました。たまたま谷底線に初期値が落ちたので、谷底線を左右を大きく振動するという段階はこの図には見られませんが、試行錯誤の段階では、推定値は大きく振動して移動しました。マクロを作らずにエクセルで作業したら、丸 1 日以上かかりました。こんなに苦労して得た結果を、ロジスティック曲線としてグラフに描くと、図 109 の灰色の線でした。明らかに、この灰線は私たちが期待した曲線ではありません。緑色の線は期待正答率の線です。期待正答率は、1cm ずつ移動しながら 10cm の幅の中でとった移動平均です。私たちはデータからこれに近い図を頭の中に何となく描いて、黄色で示したような S 字曲線のようなもの期待しますが、それは勝手

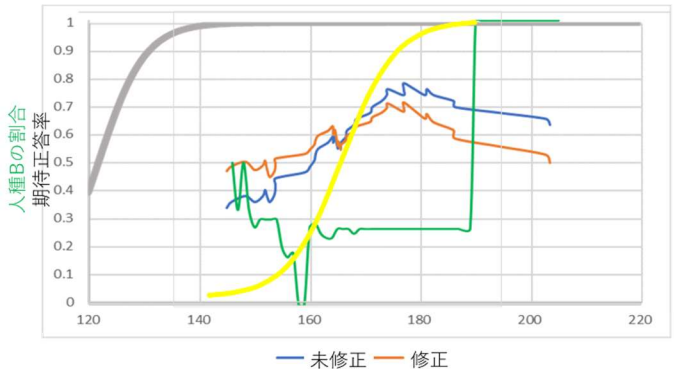


図 109. 体長と人種 B の割合との関係とロジスティック回帰の結果

表 61. 人種 A、B の体長データ

順位i	身長	t	順位i	身長	t	順位i	身長	t	順位i	身長	t
1	144.8	0	13	158.9	1	25	165.4	0	37	176.8	1
2	145.6	1	14	159.9	1	26	165.5	1	38	176.8	1
3	148.2	1	15	160.6	1	27	166.4	1	39	178.8	0
4	149.8	0	16	160.8	1	28	166.5	1	40	180.7	0
5	151.3	1	17	161.3	1	29	167.8	1	41	181.0	1
6	151.7	1	18	163.1	1	30	168.3	1	42	182.2	0
7	152.0	0	19	164.0	1	31	170.7	1	43	185.9	0
8	152.5	0	20	164.0	0	32	171.1	1	44	186.1	0
9	153.3	1	21	164.3	1	33	172.8	1	45	194.3	0
10	153.6	1	22	164.6	0	34	173.7	1	46	202.5	0
11	153.6	1	23	164.9	0	35	173.9	1	47	203.3	0
12	153.6	1	24	165.0	1	36	176.7	0			

な期待に過ぎないのです。表 61 に体長の短い方から順番にデータを示しました。体長が最も小さいデータは、人種 B の 144.8cm であり、このデータでは最低値も最高値も人種 B なのです。移動平均で考えれば、体長< 145.6の範囲では、人種 B の割合は 1 です。私たちは、144.8cm の人種 B は例外的であり、それ以下では人種 B の割合は 0 に近いだろうと予想して、黄色の線を頭に描いてしまいます。コンピュータというか「ロジスティック回帰」は、体長< 145.6では人種 B の期待値は 0 から 1 まであり得ると考えます。その結果、灰色の線が描かれました。このデータセットは、等分散性を前提としないで、判別分析を行った場合にどうなるかを考えるために、著者が作ったデータセットです。人種 A と人種 B の体長の分散は等しくありません。また、データが粗な部分と密な部分を意図的に作りました。人種 A の分散は、人種 B の分散の範囲中に収まっています。アメリカは多人種国家だから、アメリカ人という人種はありません。とても大きな人もとても小さな人もいるでしょう。そんな例だと思えばよいでしょう。このように、極端に不等分散で一方の分布が他方の分布に覆いかぶさっている場合には、ロジスティック回帰でも判別分析はできません。分析の前にデータがどのように分散しているのかを調べておくことが必要です。ここでは、ロジスティック回帰を簡単に理解するために、シグモイド関数を使って一つの境界値で判別をするということを試みました。一つの変数で選別するということはあまりないでしょう。一つの変数では判別できなくても、多次元に説明変数を増やせばその

組み合わせで、分布に違いが出るかもしれません。次の項では次元を増やしてロジティック回帰を多次元の 2 クラス判別に発展させます。

VII-2-1-5.　多次元 2 クラス判別

VII-2-1-5-1. 考え方

前項で揚げた例のように、一つの要素で判別を行うことは多くの場合困難でしょう。しかし、人はしばしば体形で人種を識別しているし、身長は体形の重要な要素の一つにすぎません。手足の長さや胸囲・頭の大きさなど、いくつかの要素を組み合わせて、判別境界線や判別境界面を作れば、より正確な判別が可能になります。表 55 のデータを使って、説明変数を体長と体重の 2 つの変数にして、ロジスティック回帰をします。図 97 にはぼんやりと判別境界線が見えます。 1 次元の時には境界は点でしたから境界平面の傾きなどを考える必要はありませんでした。多次元になると、n 次元の場合は n ― 1 次元平面が境界平面になります。これは傾きを持っています。この境界平面と図 110 の物差しの位置と傾

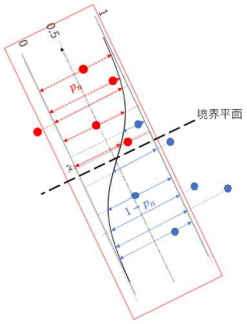


図 110. データ散布からシグモイド関数の計算のイメージ図

むきを調整して S 字曲線が 0.5 となる点を通るp方向の軸を合わせて、さらに、確率(p)、(1-p)の総積が最大になるように、境界平面と直交方向の縮尺を調整するという考え方です。これを 3 次元空間で立体的に表したのが図 111 です。図 111 の Z 軸（垂直軸）はデータ A に属する確率を表しています。実際のデータで A であったものは確率 1 で A なので、Z=1 の平面（四角形 ABCD を含む平面）上に A の属するデータをプロットしました（軸 BA が体長、軸 BC が体重)。実際のデータで B であったものは、Z=0 の平面（四角形 DFGH を含む平面）上にプロットしました（軸 FE が体長、軸 FG が体重)。A であるデータのプロットから Z=0 平面に下した垂線とロジスティック曲面（S）の交点から Z=0 平面までの距離が、その体長・体重の人が人種 A である確率(p)であり、B であるデータのプロットから、Z=1 平面に下した垂線とロジスティック曲面(S)の交点から Z=1 平面までの距離が、その体長・体重の人が B である確率(p)です。すべてのデータについて p を求めれば、その総積が尤度になります。この尤度を最大化するように、ロジスティック曲面を移動変形するというのが、計算の内容です。Z は A である確率(p)だから、p を任意の確率として、Z=p の平面で、最適化されたロジスティック曲面を切断すれば、その切断面に表れる直線が、任意の確率水準で選ばれた判別関数（モデル）判別直線）となります。何の前提もなければ、確率水準は 1/2 が選ばれるでしょう。その場合は、図の四角形 IJKL

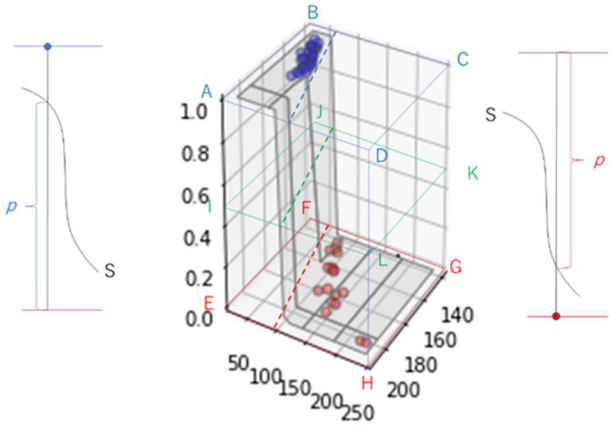


図 111. 2 次元 2 クラス判別分析で推定された確率とデータの分布

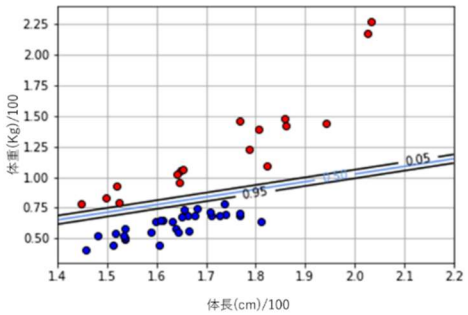


図 112.データと判別境界線(青線：確率 0.5、黒線：確率 0.95,0.05)

の平面（Z=0.5）で切断することになります。このようにして得られた点や直線を、体長―体重平面に投影すると、地図の等高線のような表現になります。図 112 は人種 A、B の判別について、平面に投影して判別境界線を等高線的に描いた例です。計算は大変なので Pyhton でやりました。ただし、体長・体重の値をそのまま使うと、指数関数の値が、コンピューターの計算限界を超えて無限大になってしまい、これを分母とすると 0 が生じ対数計算が止まってしまうので、体長・体重の単位を 100 分の 1 にしました。

上記の図を使った比喩的説明は感覚的には納得しやすいのですが、それだけでは具体的な計算過程がわからないので、数式を使って線形代数学的な解説を加えます。結論は極めたシンプルなのですが、今までの説明とつながるように、線形判別分析のところで説明した点と平面の距離の公式を作るところから始めます。高校で習った点と平面の公式と同じ考え方ですがちょっとだけ変形します。図 113 のように、p次元空間に点$\boldsymbol{X}$があり、その空間内に$p-1$次元平面$\boldsymbol{\phi}_A$があります。

$p-1$次元平面$\boldsymbol{\phi}_A$の式を

$$a_0 + ta_1h_1 + \cdots + ta_ph_p = a_0 + t\sum_{i=1}^{p} a_1h_1 = 0 \qquad \text{i}$$

とします。なお、$(a_1 \quad \cdots \quad a_p)$は平面$\boldsymbol{\phi}_A$の法線の単位ベクトルです。

$$\sum_{i=1}^{p} {a_i}^2 = 1$$

式iから

$$t\sum_{i=1}^{p} a_1 h_1 = -a_0$$

$$\sum_{i=1}^{p} a_1 h_1 = -\frac{a_0}{t} \qquad \text{ii}$$

tはベクトル$\overrightarrow{\boldsymbol{OA}}$の長さです。

$$\overrightarrow{\boldsymbol{OA}} = (ta_1 \quad \cdots \quad ta_p)$$

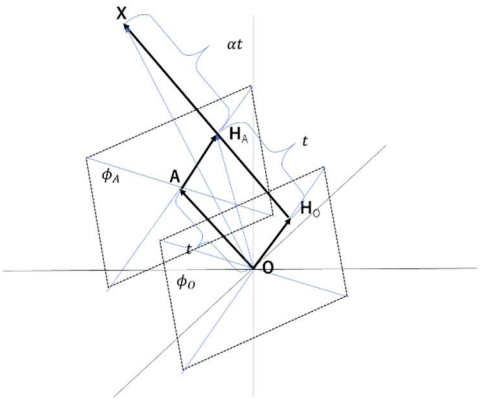


図 113. p次元空間での点と平面の関係

点$\boldsymbol{H_A} = (h_1 \quad \cdots \quad h_p)$を平面$\boldsymbol{\phi_A}$から点の向けた法線の脚とすると、

$$\overrightarrow{\boldsymbol{OH_A}} = (h_1 \quad \cdots \quad h_p)$$

点$\boldsymbol{H_A} = (h_1 \quad \cdots \quad h_p)$は平面$\boldsymbol{\phi_A}$上の点だから

$$= a_0 + t\sum_{i=1}^{0} a_1 h_1 = 0$$

$$ta_0 = -t\sum_{i=1}^{0} a_1 h_1$$

一方

$$\overrightarrow{\boldsymbol{H_AX}} \parallel \overrightarrow{\boldsymbol{OA}}$$

$$\overrightarrow{\boldsymbol{H_AX}} = \alpha\overrightarrow{\boldsymbol{OA}} \qquad \text{iii}$$

$$\left|\overrightarrow{\boldsymbol{H_AX}}\right| = \alpha\left|\overrightarrow{\boldsymbol{OA}}\right| = \alpha t \qquad \text{iv}$$

このように$\overrightarrow{\boldsymbol{H_AX}}$を$\overrightarrow{\boldsymbol{OA}}$のスカラー倍のベクトルとして表すことが出来ます。$\overrightarrow{\boldsymbol{H_AX}}$はベクトル$\overrightarrow{\boldsymbol{OH_A}}$の先端からベクトル$\overrightarrow{\boldsymbol{OX}}$の先端に向けたベクトルだから

$$\overrightarrow{\boldsymbol{H_AX}} = \overrightarrow{\boldsymbol{OX}} - \overrightarrow{\boldsymbol{OH_A}} = \left(x_1 - h_1 \quad \cdots \quad x_p - h_p\right)$$

なので、式 iii は

$$\left(x_1 - h_1 \quad \cdots \quad x_p - h_p\right) = \alpha(ta_1 \quad \cdots \quad ta_p) = \alpha t(a_1 \quad \cdots \quad a_p)$$

と書けます。
ここで、両辺について、法線の単位ベクトル$(a_1 \quad \cdots \quad a_p)$との内積をとると

$$(x_1 - h_1 \quad \cdots \quad x_p - h_p) \cdot (a_1 \quad \cdots \quad a_p) = \alpha t (a_1 \quad \cdots \quad a_p) \cdot (a_1 \quad \cdots \quad a_p)$$

式 ii を使って左辺は

$$(x_1 - h_1 \quad \cdots \quad x_p - h_p) \cdot (a_1 \quad \cdots \quad a_p) = \sum_{i=1}^{p} a_i x_i - \sum_{i=1}^{p} a_i h_i = \sum_{i=1}^{p} a_i x_i + \frac{a_0}{t}$$

$$= \frac{1}{t}\left(a_0 + t \sum_{i=1}^{p} a_i x_i \right)$$

$(a_1 \quad \cdots \quad a_p)$は単位ベクトルだから、$(a_1 \quad \cdots \quad a_p) \cdot (a_1 \quad \cdots \quad a_p) = 1$なので、
式 iii を使って右辺は、

$$\alpha t (a_1 \quad \cdots \quad a_p) \cdot (a_1 \quad \cdots \quad a_p) = \alpha t = \left|\overrightarrow{\boldsymbol{H_A X}}\right|$$

左辺＝右辺となります。

$$\frac{1}{t}\left(a_0 + t \sum_{i=1}^{p} a_i x_i \right) = \alpha t = \left|\overrightarrow{\boldsymbol{H_A X}}\right| \qquad \text{v}$$

tは$\overrightarrow{\boldsymbol{OA}}$の長さで、$\overrightarrow{\boldsymbol{OA}} = (ta_1 \quad \cdots \quad ta_p)$ですから、

$$t = \sqrt{\sum_{i=1}^{p} (ta_i)^2}$$

点と平面の距離の公式として知られている以下の公式

$$\left|\overrightarrow{\boldsymbol{H_A X}}\right| = \frac{a_0 + a_1 x_1 + \cdots + a_p x_p}{\sqrt{\sum_{i=1}^{p} {a_i}^2}}$$

にするには、a_1からa_pについて

$$ta_i \to a_i$$

と書き換えるだけです。
しかし、この項で注目するのは、式 v の形の公式です。これは、

$$\alpha t = \frac{a_0}{t} + \sum_{i=1}^{p} a_i x_i \qquad \text{vi}$$

となりますが、シグマの中のa_iは平面$\boldsymbol{\phi_A}$の単位法線ベクトル$(a_1 \quad \cdots \quad a_p)$の成分です。

$$\frac{a_0}{t} = 0, \qquad \sum_{i=1}^{p} a_i h_i = 0$$

ならば、$\boldsymbol{H_A} = (h_1 \quad \cdots \quad h_p)$は原点を含む$\boldsymbol{\phi_A}$と並行した$p-1$次元平面$\boldsymbol{\phi_O}$上にあります。つまり、単位法線ベクトル$(a_1 \quad \cdots \quad a_p)$は$\boldsymbol{\phi_A}$の傾きのみを表しています。v$\left|\overrightarrow{\boldsymbol{OA}}\right| = \boldsymbol{t}$=1 のならば、式viは

$$\alpha = a_0 + \sum_{i=1}^{p} a_i x_i \qquad \text{vii}$$

となります。

比喩的説明で使った物差し傾きの調整は、単位法線ベクトル$(a_1 \quad \cdots \quad a_p)$の最適化、物差しの位置の調整は$a_0$の最適化、物差しの縮尺の調整は t=1 とすることあたります。

なお、viiの式はベクトル$(a_0 \quad a_1 \quad \cdots \quad a_p)$とベクトル$(1 \quad x_1 \quad \cdots \quad x_p)$の内積になっています。一般的な平面と点の距離の公式は

$$\frac{a_0+a_1x_1+\cdots+a_px_p}{\sqrt{\sum_{i=1}^{p} {a_i}^2}} \qquad \text{式 83}$$

ですが、この場合は点と平面の距離の公式ととして、

$$\alpha = a_0 + a_1x_1 + \cdots + a_px_p = \boldsymbol{a} \cdot \boldsymbol{x}$$

$$\boldsymbol{a} = (a_0 \quad a_1 \quad \cdots \quad a_p)$$

$$\boldsymbol{x} = (x_0 \quad x_1 \quad \cdots \quad x_p)$$

a_0は切片、$(a_1 \quad \cdots \quad a_p)$は単位法線ベクトル

x_0はダミー変数・$x_0 = 1$、$(x_1 \quad \cdots \quad x_p)$は変数ベクトル　　式 84

と覚えてしまった方が良いかもしれません。多くのプログラム言語は行列計算に対応しているので、行列の積にすることで繰り返し計算を避けられるからです。

VII-2-1-5-2. 2 次元 2 クラス判別の計算

判別境界平面からの距離$\alpha = \boldsymbol{a} \cdot \boldsymbol{x}$から確率を計算するには、S 字曲線の形（累積確率密度関数）を決めなくてはなりませんが、シグモイド関数を使うことにします。

$$p = \frac{1}{1+e^{-\alpha}} = \frac{1}{1+e^{-\boldsymbol{a}\cdot\boldsymbol{x}}}$$

最尤法で$\boldsymbol{a}$を最適化するには、すべての$\boldsymbol{x}$について t−pを求めてその総積を全尤度としてこれを最大化します。機械学習的な計算では、同じことを平均交差エントロピー誤差（$E(\boldsymbol{a})$）の最小化という計算で行います。

$$E(\boldsymbol{a}) = -\frac{1}{N}\sum_{n=1}^{N}(t_n \log_e p_n + (1-t_n)\log_e(1-p_n))$$

であり、すでに述べたように$E(\boldsymbol{a})$の$\boldsymbol{a}$の係数による偏微分は入れ子になった複合関数の微分で、個々の関数の微分の積で、次のようになります。

$$\frac{\partial E(a_0)}{\partial a_0} = t_n - p_n$$

$$\frac{\partial E(a_1)}{\partial a_1} = (t_n - p_n)x_1$$

$$\frac{\partial E(a_2)}{\partial a_2} = (t_n - p_n)x_2$$

だから、

$$\frac{\partial E(\boldsymbol{a})}{\partial a_0} = \frac{1}{N}\sum_{n=1}^{N}(p_n - t_n) = 0$$

$$\frac{\partial E(\boldsymbol{a})}{\partial a_1} = \frac{1}{N}\sum_{n=1}^{N}(p_n - t_n)\,x_{1n} = 0$$

$$\frac{\partial E(\boldsymbol{a})}{\partial a_2} = \frac{1}{n}\sum_{i=1}^{n}(p_n - t_n)\,x_{2n} = 0$$

という連立微分方程式を作って解を求めると説明すれば、数学的な説明としてはれば良いのですが、関数の中に指数関数を含むので実際には代数的な計算では解けません。機械学習では勾配法降下法によってこれを計算します。図 111, 図 112 は Python の minimizing の機能を使って計算した結果なのですが、そのコードリストを示しても、計算過程がブラックボックスになってしまいます。計算のプロセスが追えるようにエクセルで計算のフォームを作りました。このブログのカテゴリー「やさしい水産学（自習室)」の「参考資料」にある「ロジスティック回帰計算」がその計算フォームです。この資料の中で、人種 A と人種 B の判別のためのロジスティック回帰の計算は、AB という sheet で行いました。人種 A と人種 B の判別に使うデータセットを表 62 に示しました（ここでは、人種 A を t=0、人種 B を t=1 としました。深い意味はありません。図 111 で大きい方の人種のプロットが上の Z=1 平面に来る方が自然で混乱が少ないかもしれないと思ったからです。Sheet の A1－P51 が実際の演算の sheet です。各列は次のように定義されています。
A:データ番号(i)、B:人種の判別（t=0 or 1)、C:体長(x_1)、D:体重(x_2)、E(a_0)、F(a_1)、G:(a_2)、H:$(a_1x_1);F\times C$、I:$(a_2x_2);G\times D$、J:$(y=a_0+a_1x_1+a_2x_2);E+H+I$、

K:$(-y)$;$-J$、L:(e^{-y});exp(-K)、M:$\left(p_i=\frac{1}{1+e^{-y}}\right);(L+1)^{-1}$、

N: 点 i におけるa_0による対数尤度の微分値（a_0方向の傾き）$(p_i-t);M-B$、
O: 点 i におけるa_1による対数尤度の微分値（a_1方向の傾き）$(p_i-t)x_1;N\times C$、
P: 点 i におけるa_2による対数尤度の微分値（a_2方向の傾き）$(p_i-t)x_2;N\times D$

表 62. 人種 A と人種 B の線形判別のデータセット

i	t	x1	x2	i	t	x1	x2	i	t	x1	x2	i	t	x1	x2
1	0	160.6313	44.82035	13	0	181.0023	64.22131	25	0	148.1581	51.77903	37	1	164.8974	105.3962
2	0	163.0945	63.6093	14	0	159.8722	63.32097	26	0	168.287	74.75967	38	1	149.8485	83.00537
3	0	153.5863	58.28133	15	0	153.6039	49.11976	27	0	164.331	55.04781	39	1	176.7396	145.5228
4	0	151.6996	54.56979	16	0	166.5169	55.91289	28	0	172.8068	69.09698	40	1	152.5407	79.22857
5	0	171.1278	68.17206	17	0	153.5524	50.13258	29	0	176.8415	70.65811	41	1	180.7021	138.8345
6	0	163.9663	58.07964	18	0	153.3062	51.98684	30	0	173.75	78.01168	42	1	164.0308	102.8453
7	0	167.834	69.01207	19	0	170.7289	71.49387	31	1	182.2317	109.4391	43	1	164.6233	95.83666
8	0	161.3457	64.97124	20	0	165.0226	67.25768	32	1	186.091	141.8961	44	1	194.3142	144.0872
9	0	145.6229	40.94841	21	0	158.9344	55.22369	33	1	185.8532	147.6327	45	1	144.7686	77.9867
10	0	173.9173	69.12638	22	0	160.8471	64.77768	34	1	152.0412	92.57853	46	1	165.4054	106.6235
11	0	166.369	68.41507	23	0	151.3166	44.23636	35	1	202.5394	216.8869	47	1	178.7588	122.734
12	0	165.4508	73.77345	24	0	176.8014	68.43142	36	1	203.3108	227.2248				

N49、O49、P49 は、それぞれ、その列の値の平均値であり(AVERAGE (N2:N48)、AVERAGE (O2:O48)、AVERAGE (P2:P48)、それぞれa_0方向、a_1方向、a_2方向の平均的な傾き、つまり、平均交差エントロピー誤差の微分値です。N50、O50、P50 は学習率であり、この例では 0.5 としているが、任意の学習率を選ぶことが出来ます。勾配降下法ででは、傾きの方向に下っていくので、傾きが負であれば、正の方向に、傾きに学習率を変えた値の絶対値分だけ前進し、傾きが正であれば、負の方向に、傾きに学習率かけた値の絶対値分だけ戻るので、N49、O49、P49 の値に－学習率を掛けたものを、元のa_0、a_1、a_2、の加えたものが、次のa_0、a_1、a_2となります。この sheet では、元のa_0、a_1、a_2は、それぞれ A52、B52、C52 に与え、次のステップのa_0、a_1、a_2は N51; =$A52-49\times N50$、O51; =$B52-O49\times O50$、P51; $C52-P49\times P50$に計算結果として出てきます。マクロのボタンを押すと、N51、O51、P51 の値が、A52、B52、C52 にコピーされ、D52、E52、F52 に平均エントロピー誤差の微分値である N49、O49、P49 に書き出されます。同時に、A52、B52、C52、D52、E52、F52 の値が、I52、J52、K52、L52、M52、N52 に書き出される。ここまでが一つのステップです。I54 にあるボタンを押すと、ステップが一つ進み、A52、B52、C52、D52、E52、F52、I52、J52、K52、L52、M52、N52 の値が更新されます。ボタンの左側にあるグラフの赤線が$(0=a_0+a_1x_1+a_2x_2)$の直線です。青線は筆者が散布図からおよそこの辺りで収束するだろうとデータ分布から視覚的に予想した直線です。予測された直線に近づいていくことが実感できると退屈な作業も少し頑張れるでしょう。また、このデータでは 40 番目のデータが、最も境界に近いと予想されるので、その値を監視するために、I53 に 40 番目のデータが B と判定される確率を書きだしました。これも収束の一つの目安です。

R2-U48 のところで、元の値を 100 分の 1 に小さくしています。これは、体長(cm)・体重(Kg)の値をそのまま使うと、$(e^{-\alpha})$の値が無限大になり、次の計算で 0 の対数が出てきて不能になるからです。筆者は、すべての傾きが、絶対値で 0.002 以下になったあたりで、力尽きて収束したと判断して、計算を終了しました。

収束値は

$$a_0=0.377555$$
$$a_1=-12.3099$$
$$a_2=24.28287$$

したがって、αは以下のようになります。

$$\alpha=0.377555-12.3099x_1+24.28287x_2$$

A,B どちらかに判定した場合にその確率が 1/2 となるところを境界線とすると

$$p=\frac{1}{2}=\frac{1}{1+e^{-\alpha}}$$
$$1+e^{-\alpha}=2$$
$$e^{-\alpha}=1$$

$$-\alpha = 0$$
$$0 = 0.377555 - 12.3099x_1 + 24.28287x_2$$

この式をx_2が被説明変数となる式に書き換えると、

$$x_2 = -0.015548 + 0.506938x_1$$

となりますが、これを元の単位である、体長 cm、体重 kg の単位に書き換えると、

$$体重(kg) = -1.5548 + 0.506938\ 体長(cm)$$

となります。これが、図 114 に太い灰色の実線で示した直線の式です。ちなみに、Python の minimize　という機能を使って計算した結果は、

$$a_0 = -1.00095251$$
$$a_1 = 0.95034987$$
$$a_2 = -1.9542132$$

これを体長と体重の関係式にすると

$$体重(kg) = -0.512202292 + 0.486308168\ 体長(cm)$$

でした。エクセルシートでの計算は収束のかなり手前で力尽きています。Python による計算ではほぼ瞬時に計算が終了しました。Excel で計算には大変な時間がかかり実際には無理です。これを使って計算することは薦めません。しかし、このシートを動かしてみると、計算のプロセスが具体的にわかります。エクセルでの計算で筆者が適当に与えた初期値は、

$$y = -0.085 - 0.42448x_1 + x_2$$

でした。この値をエクセルシートの A71、B71、C71 に記しました。この値を、A52、B52、C52 にコピーすると、確かに、赤い線と青い線が重なります。ボタンを押して次のステップへ進むと、赤い線は青い線から大きく離れます。しかし、これを繰り返していると、ゆっくりと次第に赤い線は青い線に近づいていきます。この過程で赤い線の傾きも変化しているのですが、それはあまり目立ちません。中間値１ぐらいまでは、比較的少ないステップで、青線近づきますが、それ以後、動きは遅くなります。傾斜が比較的小さくなって、一回に進む距離が短くなるためです。中間値２ぐらいからは、傾きがゆっくりと変わっていき、時間をかければ収束値に至ると期待できます。同じ方法で、人種 A と人種 C の判別境界線を求めました。計算例は、ロジスティック回帰計算の AC の sheet に示しました。
その結果、

$$a_0 = 23.46295$$
$$a_1 = -20.9539$$
$$a_2 = 23.90215$$
$$\alpha = 23.46295 - 20.9539x_1 + 23.90215x_2$$

$\alpha = 0$としてx_2を目的変数に書き換えて、単位を１０００倍にして

$$体重(kg) = -98.163 + 0.877\ 体長(cm)$$

が、境界となる直線式です。図 114 にはこの線を灰色の破線で示しました。

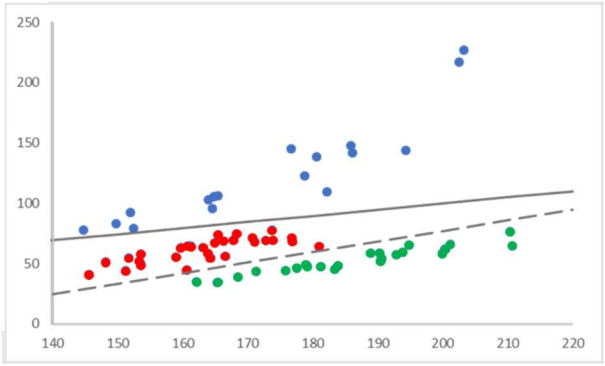


図 114.人種ごとの体長―体重分布と判別直線

VII-2-1-5-3. Python を使った多次元２クラス判別

図 112 は Python を使ったロジスティック回帰のプログラムで作りました。プログラムのコードリストを、このブログのカテゴリー「やさしい水産学（自習室）」-「参考資料」-「python」に「VII-2-1.ロジスティック回帰」としてあげておきました。その中で使ったデータは「Sample data」の中にあります。コードリストは Jupyter Notebook をエディターに使って書きました。リストは、VII-2-1-i.準備とデータの読み込み、VII-2-1-ii.判別するクラスを指定、VII-2-1-iii.関数の定義と実行、VII-2-1-iv.結果の図示、VII-2-1-v.等高線図と 3D グラフ、VII-2-1-vi.3D グラフ(color)、VII-2-1-vii.結果の保存、という構成になっています。

「VII-2-1-i.準備とデータの読み込み」では、#縮尺倍率を決めるというコードがあり、データを 100 分の 1 に小さくしています。これは、数値が大きいとシグモイド関数の分母の指数がコンピュータの計算限界を超えて大きくなって、シグモイド関数が 0 になってしまうために対数が取れずに計算が止まってしまうからです。実際のデータを分析する場合にはそういう問題が起こるので、この部分で数値の大きさを調整してください。「VII-2-1-iv.結果の図示」では、確率 0.05、0.5、0.95 の等高線を書きました。確率水準は#等高線を書く作業の定義のところの、levels=(0.05,0.5,0.95)というところで確率水準を指定しています。「VII-2-1-vii.結果の保存」は結果の出力と保存ですが、Jupyter Notebook のホルダーに、parameter1.csv という名前で推定された係数が保存されます。また、作った図も、ABcontour.png、AB3d_data.png、ABp.png などの名前で保存されます。