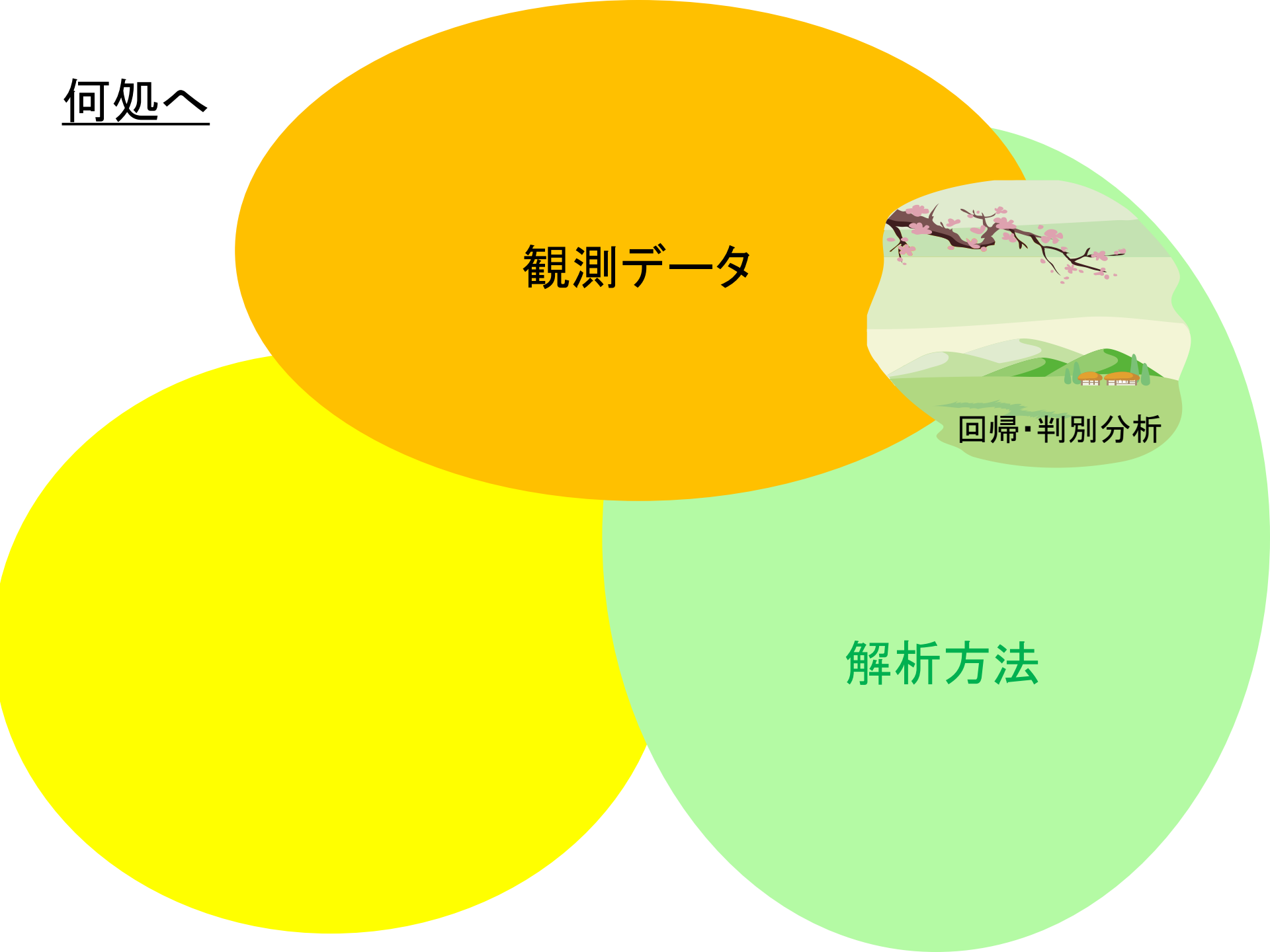
データ解析

第3回: 判別分析

渡辺澄夫

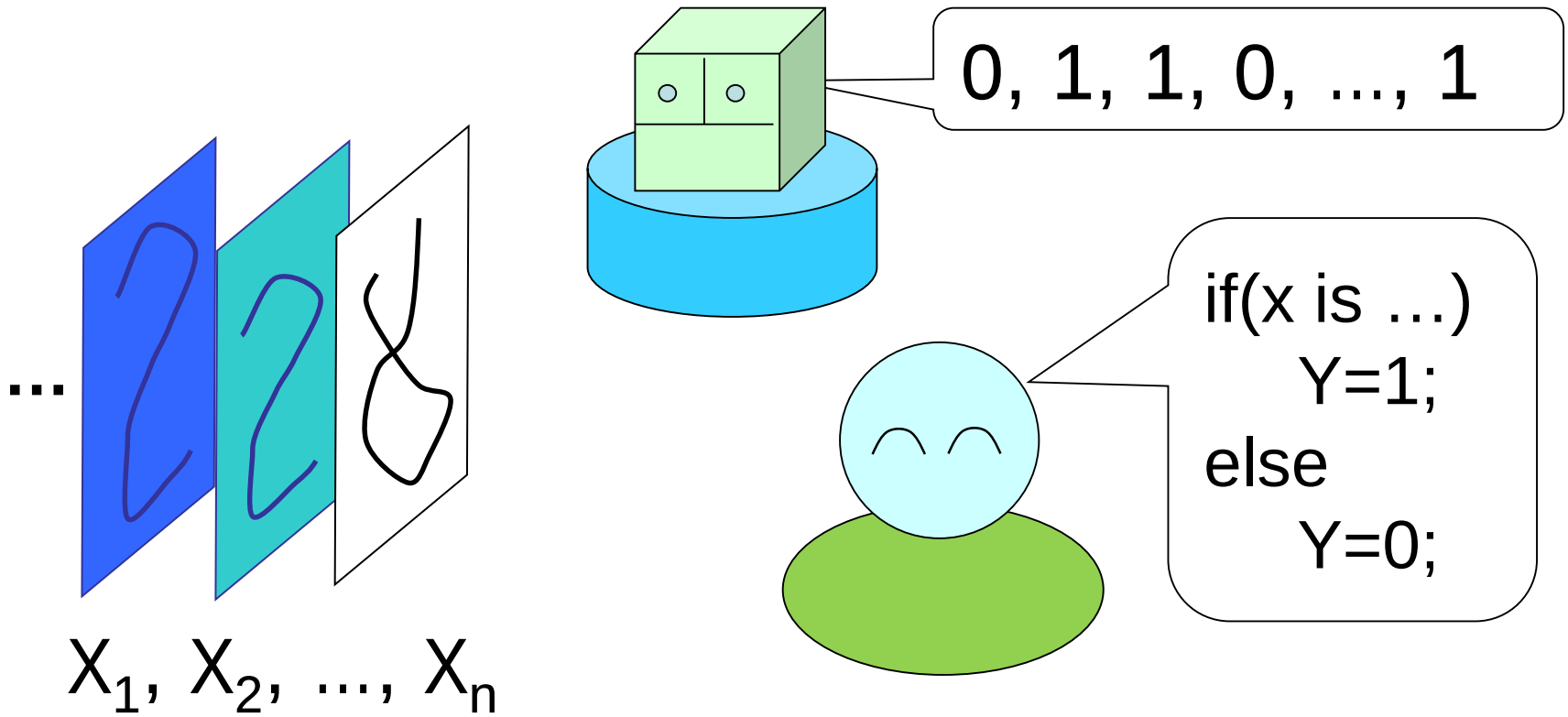
何処へ



最適判別関数

二つの確率変数 (X,Y) の確率密度関数 q(x,y) がわかっているならば
判別確率を最小にする関数を定めることは容易である。

判別分析とは



確率変数 (X,Y) のデータが得られて、X は R^Mに値をとり Y は{0,1}に値をとる確率変数とする。X に対してどのように 0 と 1 を定めるのが最も間違いが少なくなるだろうか。

同時確率密度関数

$R^M \times \{0,1\}$上の同時確率密度関数 $q(x,y)$ を考える。

$$q(x,0) = a\, q_0(x), \qquad q(x,1) = b\, q_1(x),$$

ここで $p_0(x)$ と $p_1(x)$ は確率密度関数で $a+b=1$ とする. 両方合わせて

$$q(x,y) = (a\, q_0(x))^{1-y} (b\, q_1(x))^{y}$$

と書くことができる (y は 0 または 1)。周辺確率密度関数は

$$q(x) = a\, q_0(x) + b\, q_1(x),$$

であるから、条件つき確率は

$$q(y|x) = (a\, q_0(x))^{1-y} (b\, q_1(x))^{y} \;/\; (a\, q_0(x) + b\, q_1(x)),$$

回帰関数 $g(x)=E[Y|x]$ は $q(1|x)$ と一致する。

$$g(x) = \Sigma_{y=0,1}\, y\, q(y|x) = b\, q_1(x) \;/\; (a\, q_0(x) + b\, q_1(x)) = q(1|x).$$

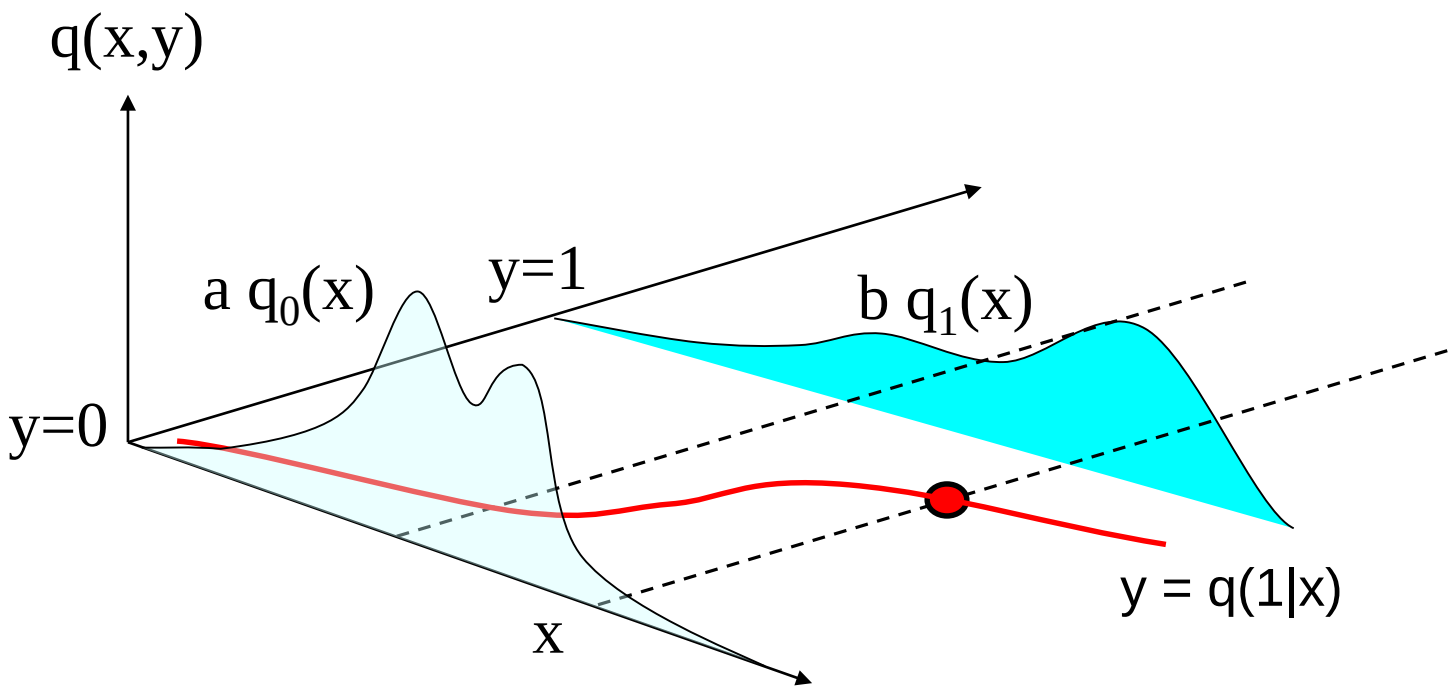
確率的推論

x が与えられたとき y=0,1 となる確率はそれぞれ

$$q(0|x) = a\, q_0(x) \;/\; (a\, q_0(x) + b\, q_1(x))$$

$$q(1|x) = b\, q_1(x) \;/\; (a\, q_0(x) + b\, q_1(x))$$

であるから、x が与えられたときの y は0か1かに対する答えは、
「もし q(1|x)>1/2 ならば y=1, それ以外は y=0 」とするのが合理的である。



判別分析は大小比較

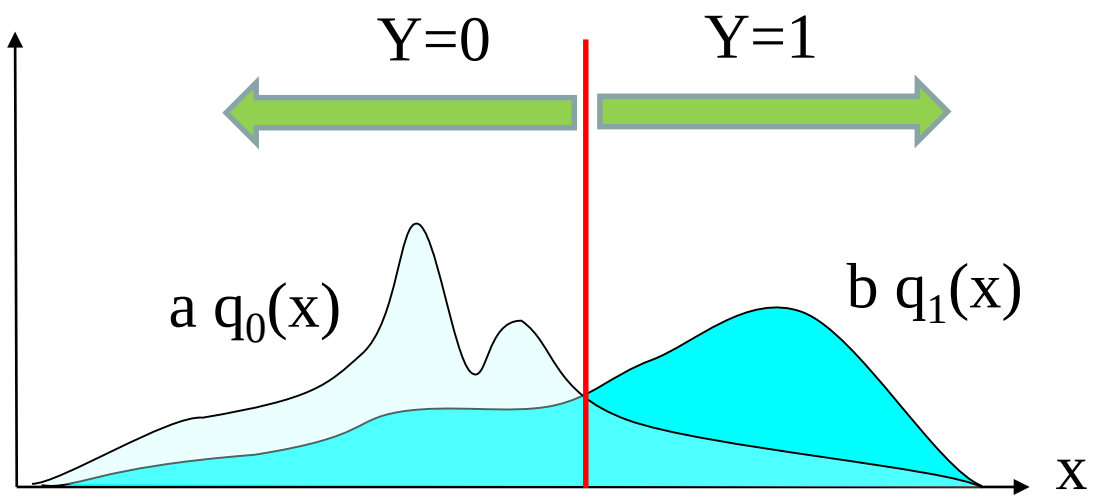
x が与えられたとき y=1 となる確率は

$$q(1|x) = b\, q_1(x) \;\; / \;\; (a\, q_0(x) + b\, q_1(x))$$

$$= 1 / \{ 1 + (a\, q_0(x)) / (b\, q_1(x)) \}$$

であり、　$q(1|x) > 1/2 \quad \Leftrightarrow \quad a\, q_0(x) < b\, q_1(x)$　なので、

要するに $a\, q_0(x)$ と $b\, q_1(x)$ の大小比較で判定するだけです。



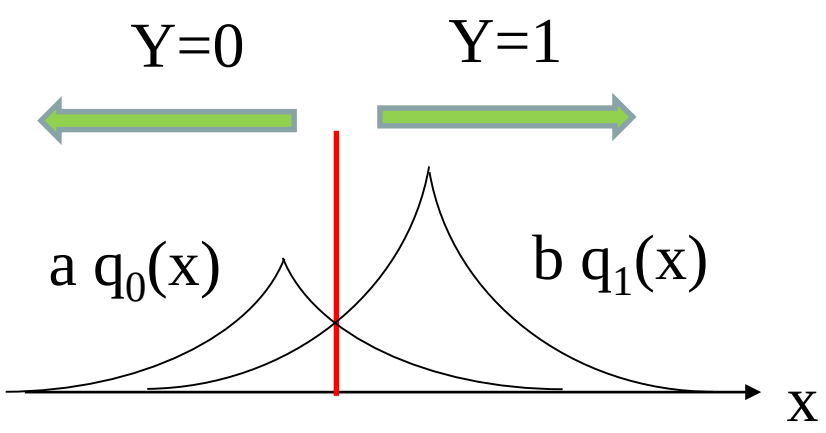
例

簡単な例として M=1, a=1/3, b=2/3,

$q_0(x) = (1/2)\exp(-|x|)$,

$q_1(x) = (1/2)\exp(-|x-1|)$,

の場合を考えてみよう。



$$q(1|x) = (2/3)\exp(-|x-1|) \ / \ \{ (1/3)\exp(-|x|) + (2/3)\exp(-|x-1|) \}$$

$$= 1 \ / \ \{ \ (1/2)\exp(-|x|+|x-1|) + 1 \}$$

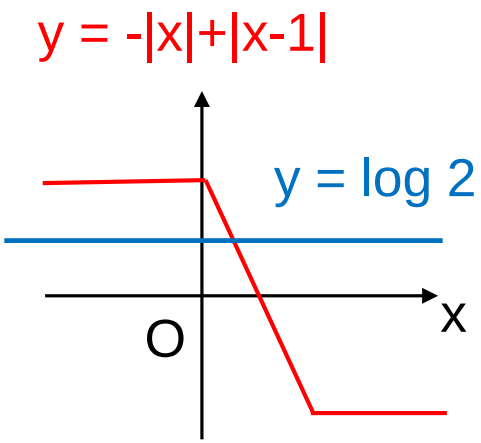
となるので

$$q(1|x) > 1/2 \ \Leftrightarrow \ 1 \ / \ \{ \exp(-|x|+|x-1|)/2 + 1 \} > 1/2$$

$$\Leftrightarrow \ \exp(-|x|+|x-1|)/2 < 1$$

$$\Leftrightarrow \ -|x|+|x-1| - \log 2 < 0$$

$$\Leftrightarrow \ x > (1-\log 2)/2$$



推論は「X>(1-log 2)/2 ならば Y=1, それ以外は Y=0」になる。

最適判別関数

定理. 条件 $q(x)>0$ を仮定する。h を R^Mから$\{0,1\}$への関数とする。関数 h を用いた推論 $y=h(x)$ が間違う確率は

$$\mathrm{ErrProb}(h) = \Sigma_{y=0,1} \int dx \ |y - h(x)| \, q(x,y)$$

である。これは $h(x)$ = 「$q(1|x)>1/2$ ならば 1, それ以外は 0 」のときに限り最小値を取る （ここでは＞と≧は同じものだと考えることにする）。

証明. 仮定から $|y-h(x)|$ は 0 または1だから $|y-h(x)| = |y - h(x)|^2$ である.

回帰関数は $g(x)=q(1|x)$ なので

$$\mathrm{ErrProb}(d) = \Sigma_{y=0,1} \int dx \ |y - h(x)|^2 \, q(x,y)$$

$$= \int dx \ \{ \Sigma_{y=0,1} |y - q(1|x)|^2 \, q(x,y) + |q(1|x)-h(x)|^2 \, q(x)\}.$$

各 x ごとに $|q(1|x)-h(x)|$ を最小にする関数は題意のものだけである.（証明終）

注意. 誤り確率を最小にする判定法は何であるかと問うならばそれは条件つき確率が1/2より大きいかどうかで0と1を分ける方法だということです。

符号長最小化

定理. 条件 $q(x)>0$, $0<q(1|x)<1$ を仮定し $q(1|x)$ は x の連続関数とする。$f(x)$ を R^Mから区間$(0,1)$への連続関数とする。符号長 $K(f)$ を次式で定義する。

$$K(f) = -\Sigma_{y=0,1} \int dx \;\; q(x,y) \{ y \log f(x) + (1-y) \log (1-f(x)) \}$$

とおく。$K(f)$ は $f(x) = q(1|x)$ のときに限り最小値を取る。

証明. $\Sigma y=0,1$ の和を取ると回帰関数 が $q(1|x)$ であることから

$$K(f) = -\int dx \;\; q(x) \{ q(1|x) \log f(x) + (1-q(1|x)) \log (1-f(x)) \}$$

定数 c が $0<c<1$ なら $t \in (0,1)$の関数 $c \log t + (1-c) \log (1-t)$ は $t=c$ でのみ最大となる (t で微分して増減表を書くとよい)。従って各 x ごとに $f(x)=q(1|x)$ のときのみ最小となる。(証明終)

注意. 汎関数 $K(f)$ は $q(1|x)$ に対する $f(x)$ の対数損失と呼ばれるもので、カルバック・ライブラ情報量(＋定数)と一致します。それは $q(1|x)$ を情報源とするときの符号化 $f(x)$ の符号長を意味しています。条件つき確率 $q(1|x)$ は符号長 $K(f)$ を最小とするものでもあるということがわかりました。

何がわかったのか

$R^M \times \{0,1\}$上の同時確率密度関数 $q(x,y)$ がわかっているとき
「x に対する y は1か」という問いに対する答えは $q(1|x)$ であり、それは

誤り確率　$\mathrm{ErrProb}(h) = \Sigma_{y=0,1} \int dx \ |y - h(x)| \, q(x,y)$

符号長　$K(f) = -\Sigma_{y=0,1} \int dx \ q(x,y) \{ y \log f(x) + (1-y) \log (1-f(x)) \}$

をどちらも最小とするものであることが分かりました。

実世界では $q(x,y)$ はわかりません。このときデータだけをもとにして

(1) $q(x,y)$ を推定してから $q(1|x)$ を求める
(2) 誤り確率を最小化する $h(x)$ を求める
(3) 符号長を最小化する $f(x)$ を求める

は同じではない結果を与えます。データから推定したものは真のものから
ずれていますが、ずれかたが異なるのです。

判別関数の推定

実世界では真の分布はわからないのでデータから $q(1|x)$ を推定する。

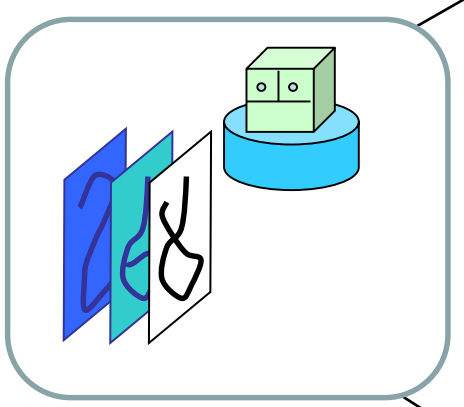
物語はそこから始まる。

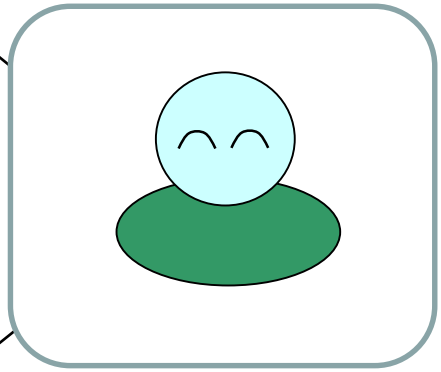
判別分析の枠組み

データ

$X_1, X_2, \ldots, X_n$

$Y_1, Y_2, \ldots, Y_n$





テストデータ

X

Y

情報源

(X,Y) は $R^M \times \{0,1\}$ に
値をとる確率変数

統計モデル

$y=f(x,w)$

判別分析の方法

データ $\{(X_i,Y_i)\,;\,i=1,2,...,n\}$ から条件つき確率 $q(1|x)$ を推定しよう。

判別分析では $q(x,y) = (a\,q_0(x))^{1-y}(b\,q_1(x))^{y}$ を推定し、それらを用いて $q(1|x)$ を推定する。

$$\{(X_i,Y_i)\,;\,i=1,2,...,n\} \;\rightarrow\; q(x,y) \;\rightarrow\; q(1|x)$$

周辺化すると y の分布は $q(x) = (a)^{1-y}(b)^{y}$

y=0 のとき $q(x|0) = q_0(x)$

y=1 のとき $q(x|1) = q_1(x)$

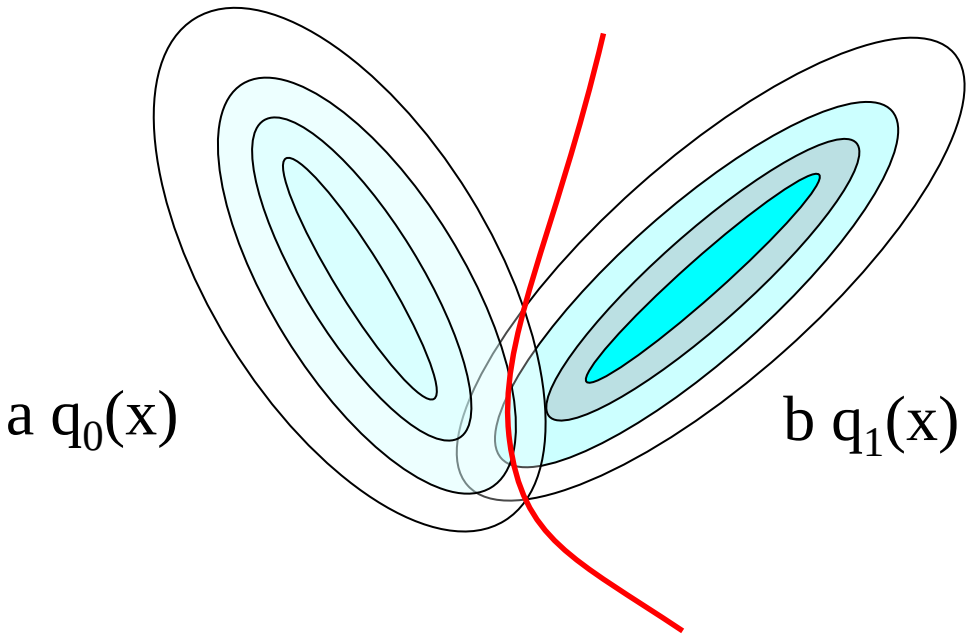
これらの推定を行なう。$q_0(x)$ と $q_0(x)$ は正規分布で推定する。

判別分析の第1の方法

データ $\{(X_i,Y_i)\,;\,i=1,2,\ldots,n\}$ から $q(1|x)$（確率かつ回帰関数）を推定する。

<u>判別分析</u>. $aq_0(x)$, $bq_1(x)$ をデータからモデルを用いて推定し、

$q(1|x) = bq_1(x) / (aq_0(x)+bq_1(x))$ と推定する。



$a\,q_0(x) = b\,q_1(x)$: 識別境界

注意. $q_0(x)$ と $q_1(x)$ の推定には正規分布を使うことにする。

$q_0(x) \leftarrow$ 平均 c_0, 分散 S_0
$q_1(x) \leftarrow$ 平均 c_1, 分散 S_1

$q(1|x) > 1/2$
$\Leftrightarrow a\,q_0(x) < b\,q_1(x)$

最尤推定法

パラメータ w を持つ x の確率密度関数 p(x|w) が与えられたときデータ $\{X_i ; i=1,2,\ldots,n\}$ から w を推測する方法のひとつに最尤推定法がある。尤度関数 L(w) を次式で定義する。

$$L(w) = \prod_{i=1}^{n} p(X_i|w).$$

もしもモデル p(x|w) がデータを発生したとすれば, L(w) はデータの発生確率（の密度関数）である。これを最大化することでパラメータを定めるのが最尤法である。

平均 a, 分散共分散行列 S をパラメータに持つM次元正規分布を次式で定義する。

$$p(x|a,S) = 1/\{(2\pi)^M \det(S)\}^{1/2} \exp(-(1/2)(x-a)^T S^{-1}(x-a))$$

補題. 尤度 $L(a,S) = \prod_{i=1}^{n} p(X_i|a,S)$ を最大にするパラメータ（最尤推定量）は

$$a^* = (1/n)\,\Sigma_i X_i, \qquad S^* = (1/n)\,\Sigma_i (X_i-c_0)(X_i-c_0)^T.$$

最尤推定法（つづき）

（証明）対数関数は単調増加なので、L(a,S) を最大化することは log L(a,S) を最大化することと等価である。$Z(a) = (1/n)\ \Sigma_i\ (X_i-a)\ (X_i-a)^T$ とおく。

$$2 \log L(a,S) = n \ \log \det(S) \ - \Sigma_i\ (X_i-a)^T S^{-1}(X_i-a)$$

$$= n \log \det(S) \ - \Sigma_i\ \mathrm{tr}(\ S^{-1}(X_i-a)\ (X_i-a)^T\ )$$

$$= n \log \det(\ Z(a)^{1/2}\ S^{-1} Z(a)^{1/2}\ ) \ - \ n\ \mathrm{tr}(\ Z(a)^{1/2}\ S^{-1} Z(a)^{1/2}\ ) - n \log \det Z(a)$$

<u>まず a について</u>。S を固定して a について平方完成すると

$$\Sigma_i\ (X_i-a)^T S^{-1}(X_i-a) = \|\ S^{-1/2}(a-a^*)\|2 + (a\text{によらない定数})$$

となることから S によらず $a=a^*$ で log L(a,S) は最大となることがわかる。
<u>次に S について</u>。$Y= Z(a^*)^{1/2}\ S^{-1}\ Z(a^*)^{1/2}$ とおくと

$$2 \log L(a,S) = n\ (\ \log \det Y - \ \mathrm{tr}\ Y\ ) + (\text{パラメータによらない定数})$$

となる。Y は正定値行列であり固有値を $y_1,y_2,..,y_M$ と書く。関数 $f(x) = \log x - x$ と書くと $(\ \log \det Y - \ \mathrm{tr}\ Y\ ) = \Sigma_i\ f(x_i)$ であるが f(x) は x =1 のときだけ最大値をとるので log L(a,S)が最大となるのは $y_1=y_2=\ldots=y_M=1$ のときだけである。このとき Y は対角化に用いる行列に依存せず単位行列である。よって $S= Z(a^*)^{-1}$ で最大になる（証明終）。

判別分析のアルゴリズム

0. データ $\{(X_i,Y_i)\,;\,i=1,2,\ldots,n\}$ $(X_i \in R^M,\ Y_i=0,1)$が与えられたとする。

1. Y が0と1のデータ個数 n_0,n_1 $(n_0+n_1=n)$ から (a,b) を $(n_0/n,n_1/n)$ で推定。

2. $q_k(x)$ $(k=0,1)$ を、平均 c_k, 分散共分散行列 S_k の正規分布で推定。

$$\begin{cases} c_0 = (1/n_0)\,\Sigma_i\,(1-Y_i)\,X_i \\ c_1 = (1/n_1)\,\Sigma_i\,Y_i\,X_i \\ S_0 = (1/n_0)\,\Sigma_i\,(1-Y_i)\,(X_i-c_0)(X_i-c_0)^T \\ S_1 = (1/n_1)\,\Sigma_i\,Y_i\,(X_i-c_1)(X_i-c_1)^T \end{cases}$$

実際のデータが正規分布から
発生することはめったにないが
正規分布を使って推測している

→ $q_k(x)$ の推定結果：$q_k(x) = [(2\pi)^M \det(S_k)]^{-1/2} \exp(-(X-c_k)^T S_k^{-1}(X-c_k)/2)$

3. 条件つき確率を $q(1|x) = bq_1(x) / (aq_0(x)+bq_1(x))$ と推定する。

4. 新しいデータ X を「$q(1|X)>1/2$ ならば $Y=1$, それ以外は$Y=0$」と判定。

判別分析つづき

補題. 前ページの記号を用いる。命題「$q(1|x)>1/2$ 」は次と同値である。

$2\log(n_0/n_1) + \log\det(S_1S_0^{-1}) +(x-c_1)^TS_1^{-1}(x-c_1) -(x-c_0)^TS_0^{-1}(x-c_0)<0.$

新しい x に対する y は「不等式が成立なら1、それ以外は0」と判定する.

（証明）計算するだけですが、計算を書いておきます。

$q(1|x) = bq_1(x) / (aq_0(x)+bq_1(x)) = 1 / \{ 1+(a/b)(q_0(x)/q_1(x)) \}$ だから

$q(1|x)>1/2 \Leftrightarrow (a/b)(q_0(x)/q_1(x)) <1 \Leftrightarrow \log(a/b) + \log(q_0(x)/q_1(x))<0$

ここで

$$\log \frac{q_0(x)}{q_1(x)} = \log \frac{[(2\pi)^M\det(S_0)]^{-1/2} \exp(-(X-c_0)^TS_0^{-1}(X-c_0)/2)}{[(2\pi)^M\det(S_1)]^{-1/2} \exp(-(X-c_1)^TS_1^{-1}(X-c_1)/2)}$$

$= (1/2)\log\det(S_1S_0^{-1}) +(x-c_1)^TS_1^{-1}(x-c_1)/2 -(x-c_0)^TS_0^{-1}(x-c_0)/2$ （証明終）

判別分析のプログラム

データ $\{(X_i,Y_i)\,;\,i=1,2,\ldots,n\}$ が与えられたとき

$$
\begin{cases}
n_0 = \Sigma_i\,(1-Y_i) \\
n_1 = \Sigma_i\,Y_i \\
c_0 = (1/n_0)\,\Sigma_i\,(1-Y_i)\,X_i \\
c_1 = (1/n_1)\,\Sigma_i\,Y_i\,X_i \\
S_0 = (1/n_0)\,\Sigma_i\,(1-Y_i)\,(X_i-c_0)(X_i-c_0)^T + (\varepsilon/n)\,I \\
S_1 = (1/n_1)\,\Sigma_i\,Y_i\,(X_i-c_1)(X_i-c_1)^T + (\varepsilon/n)\,I
\end{cases}
$$

S_0,S_1 が可逆にならないと困るので単位行列 I と小さい値ε>0を加えた．事前分布を仮定して事後確率最大化推定をしたと考えてもよい。

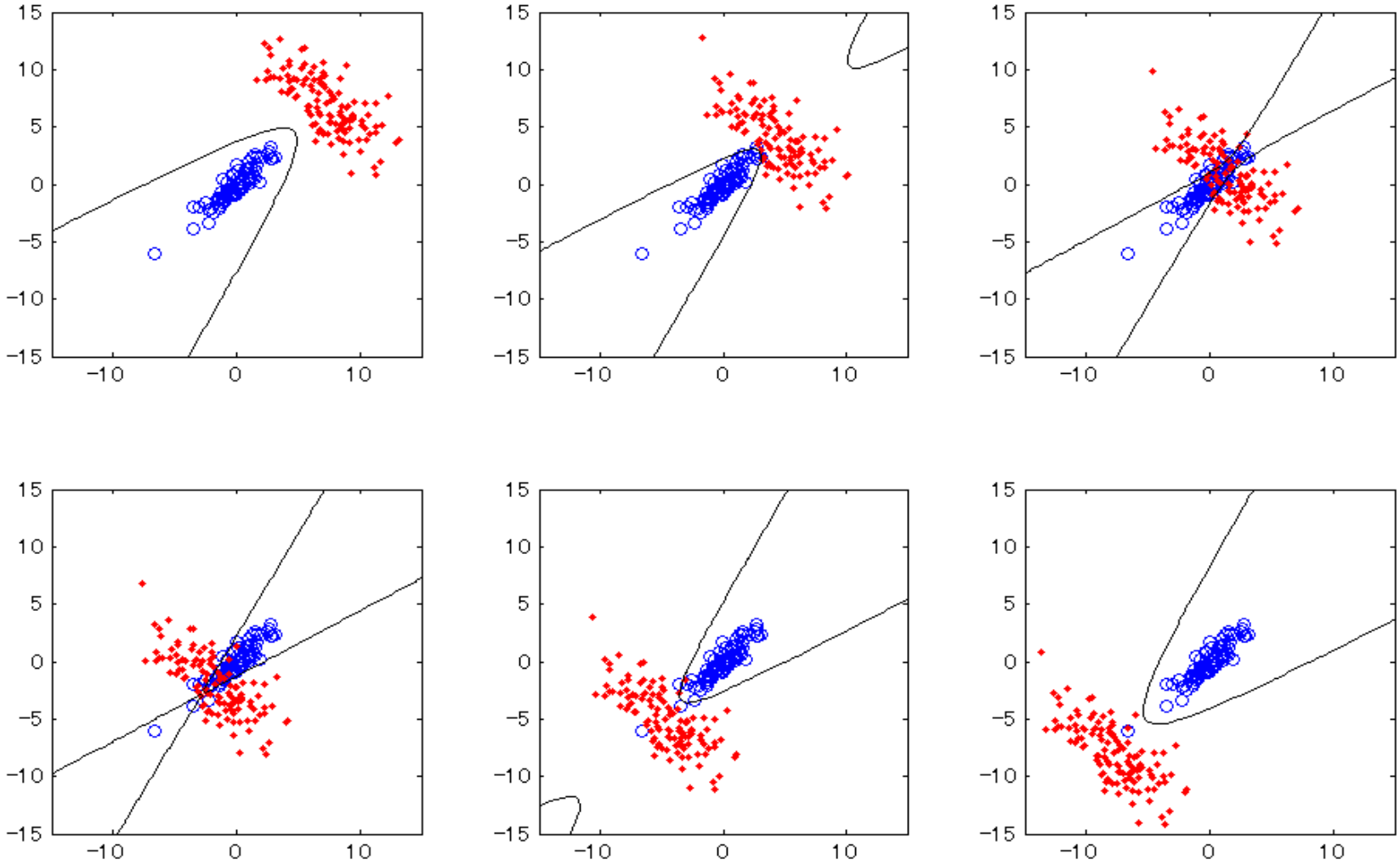
新しい x に対する判定法は

「 $2\log(n_0/n_1) + \log\det(S_1S_0^{-1}) + (x-c_1)^TS_1^{-1}(x-c_1) - (x-c_0)^TS_0^{-1}(x-c_0) < 0$

ならば Y=1、それ以外は Y=0」.

簡単な場合にやってみた

ふたつのデータを正規分布から出して判別分析が作る境界を描いてみた。



判別分析の方法3通り

データ $\{(X_i,Y_i)\,;\,i=1,2,\ldots,n\}$ から $q(1|x)$（確率かつ回帰関数）を推定するには判別分析のほかにも次の方法を用いることもできる。

2. パラメータを持つ関数 $f(x,w)$ を用いて回帰関数 $q(1|x)$ を推定する。
二乗誤差 $\sum_{i=1}^{n}(Y_i-f(X_i,w))^2$ を最小化して w を推定する。

3. パラメータを持つ関数 $f(x,w)$ を用いて確率 $q(y|x)$ を推定する。
尤度関数 $\prod_{i=1}^{n} f(X_i,w)^{Y_i}\,(1-f(X_i,w))^{1-Y_i}$ を最大化して w を推定する。

注意. 関数 $q(1|x)$ は回帰関数でもあり条件つき確率でもある。2は回帰関数を推定する方法であり、3は条件つき確率を推定する方法である。

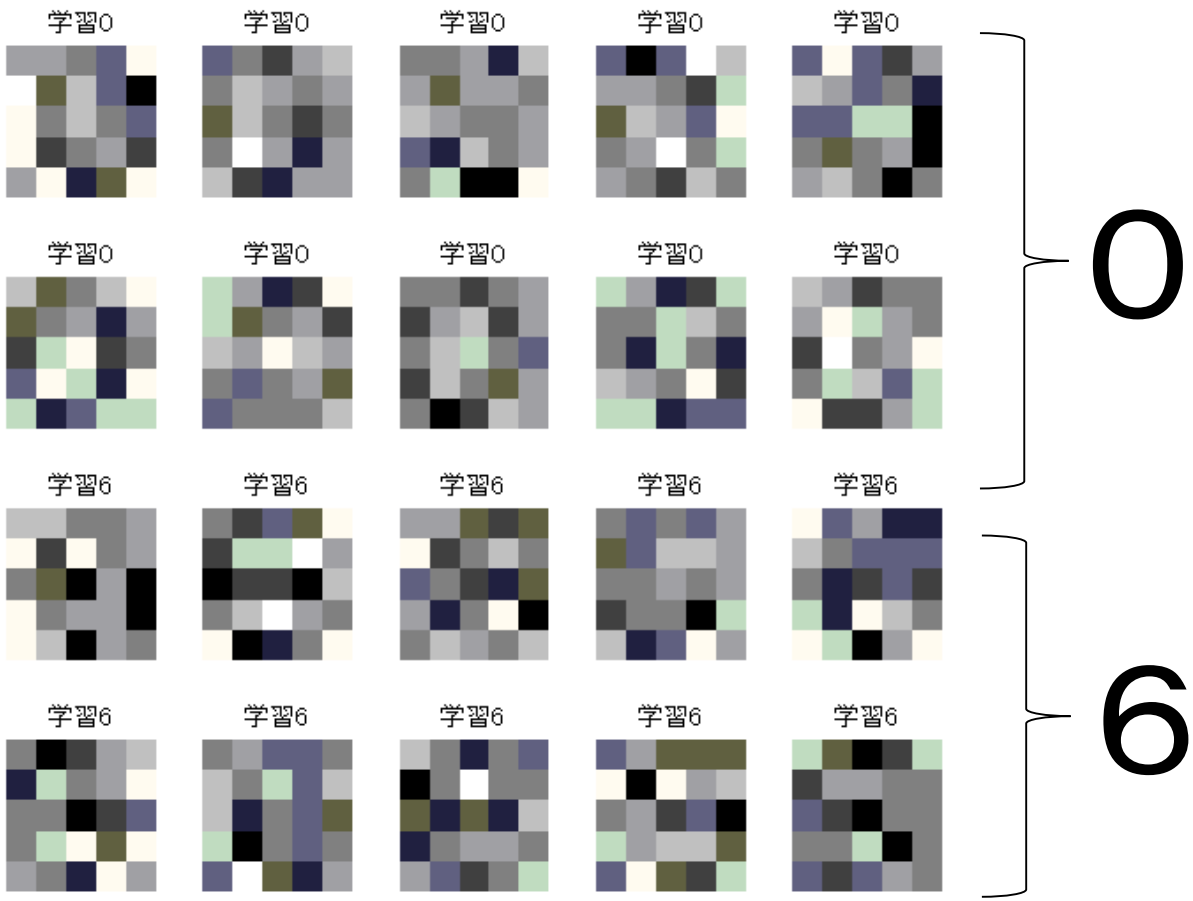
注意.「判別分析」という言葉は方法1を意味することが多いので、講義では1の方法を紹介した。しかしながら、実は1の方法は目的とする判別には不要なものまで推定しているため特別な場合以外は精度がよくないことが知られている。2と3はどちらも精度がよく、広く使われるがどちらが良いかはケースによる。目標が同じでも方法が異なるので推定結果は同じではない。

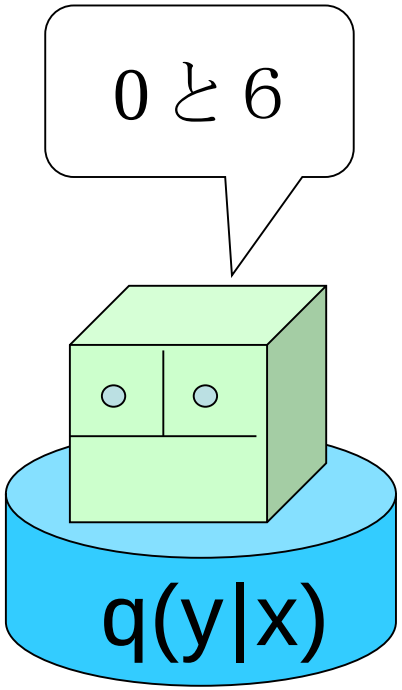
具体的な例

情報 X を 0 と 1 に識別するという操作は多くの情報システムの
性能を決定しているものであり、今日では極めて多くの応用がある。

判別分析で文字を認識してみる

「0」と「6」の文字の例を200個ずつ合計400個を使って判別分析をつくり、
それとは別のテスト用データ400個で認識率を計算してみた。
各文字は25次元の実数を要素とするベクトルなので判別分析できる。

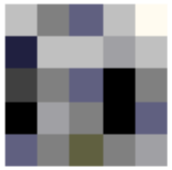




テストデータの判別結果

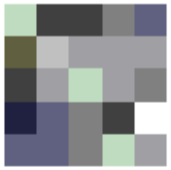
本当は0なのに6と誤判定した(400個中３個）

判定6,実は0



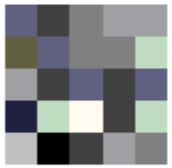
判定6,実は0

判定6,実は0



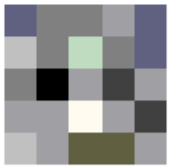
本当は6なのに0と誤判定した（４００個中８個）

判定0,実は6



判定0,実は6

判定0,実は6

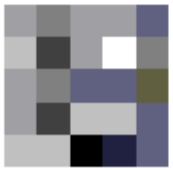


判定0,実は6

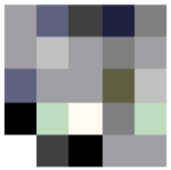
判定0,実は6

判定0,実は6

判定0,実は6



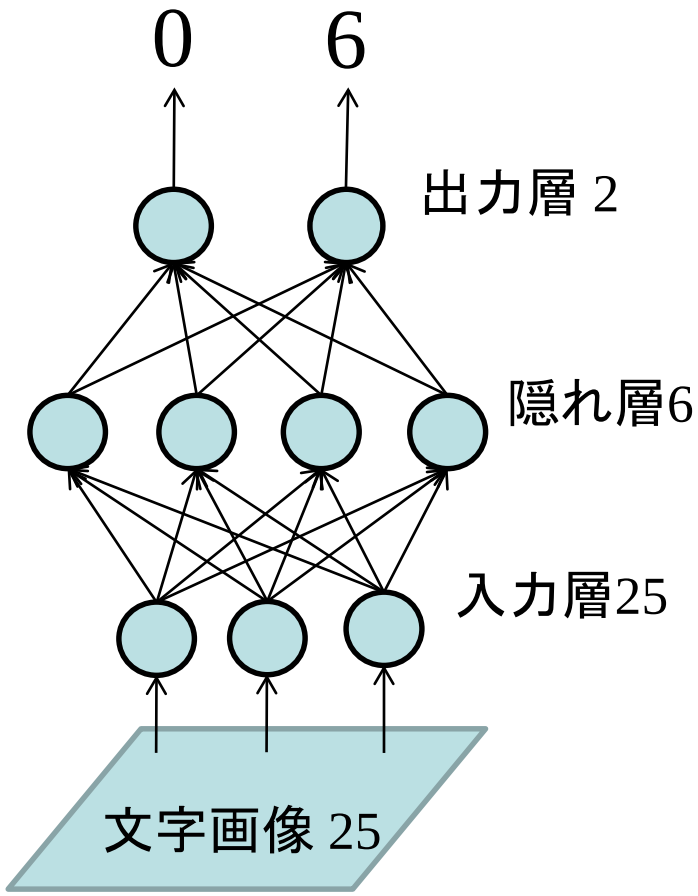
判定0,実は6



学習用データで
誤識別は３個、
テストデータでは
誤識別は11個で、
間違えた文字は
左のようなものだった。

人間でも判定は
難しいケースのように
見えます。

参考：神経回路網での文字認識と比較してみた



パラメータを持つ関数 f(x,w) を用いて

回帰関数 q(1|x) を推定する。

二乗誤差 $\sum_{i=1}^{n}(Y_i-f(X_i,w))^2$ を

最小化して w を推定する。

関数 f(x,w) として神経回路網を用いた。

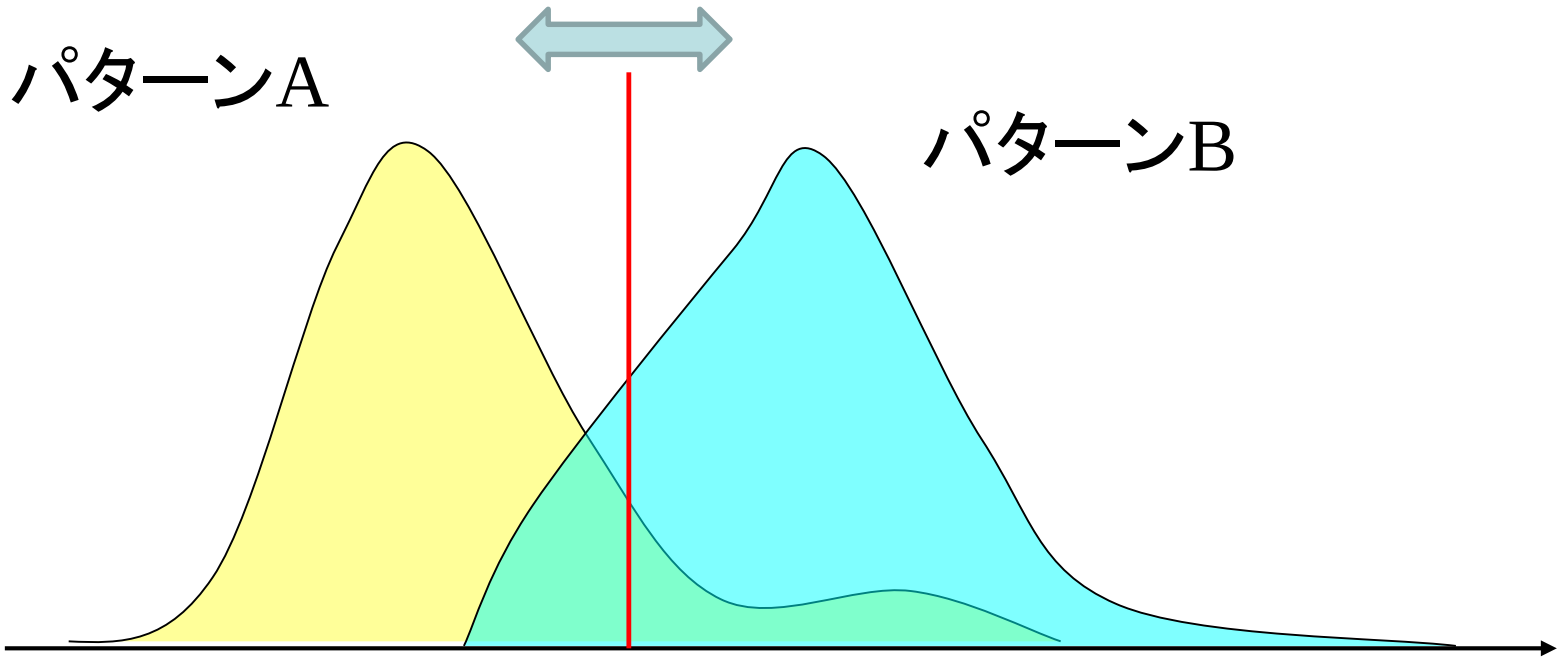
学習用のデータで誤識別は0個
テストデータでは8個だった。

判別分析での誤識別11個の集合と
神経回路網の誤識別8個の集合で
どちらにも含まれている要素は1個だった。
推定のしかたによって異なる識別境界が
できるようである。

この講義では神経回路網は
学習しませんが、比較のための
参考として載せています。

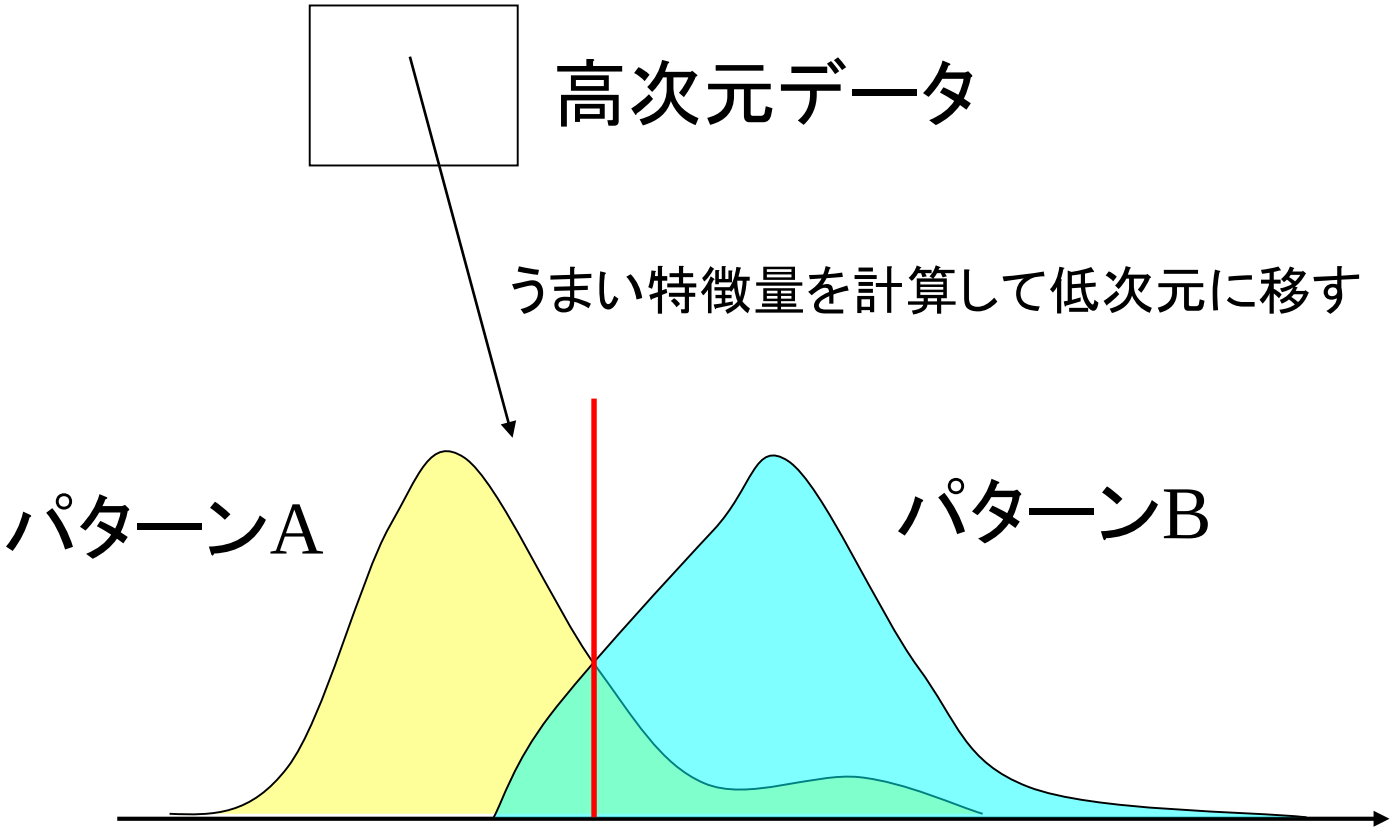
パターン識別:今昔物語（1）

もしも二つのパターンの確率分布がわかっていれば
誤識別を最小にする境界を見つけるのは難しくない。



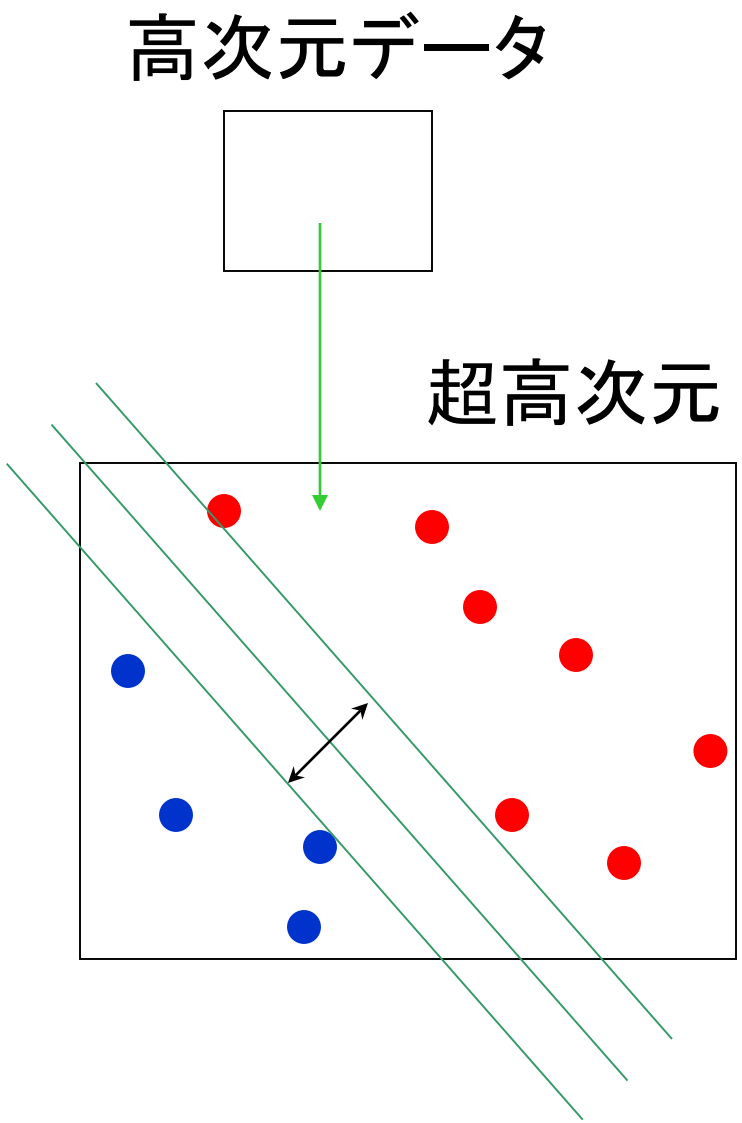
遠い昔、パターンの識別にはA,Bの確率分布を使う方法がよいだろうと
思われていた。深層学習は昔は「邪道」だと考えられていた。

パターン識別：今昔物語（2）



昔、優れた特徴量を探し出す方法で競争があった。
深層学習は自動的に特徴量を見つけるといわれている。

パターン識別：今昔物語（3）



高次元データをさらに
超高次元へ移して
識別する方法（SVM)や

深い階層を持つ神経
回路網を使う方法が
使われるようになっている。

判別分析は精度の点では
SVMや神経回路網には
及ばないことが多いが

判別分析は識別の過程が
明確なので、得られた結論の
理由づけが考えやすいという
特長がある。

判別分析いろいろ

１．入力 X を0か1に判定する操作は、音声文字画像の認識や理解などに広く用いられる基本的なものであり、広く応用されている。今日では人間の能力を超えるものもある。

２．判別分析の長所は、X から Y への推論が理解しやすいところになる。またもしも Y=0,1 に対する X が正確に正規分布であるときには、判別分析も精度は悪くはない。

３．入力 X を0か1に判定する装置が複数個あるとき、それらの判定が異なる場合には、重ね合わせの推論で精度が向上する場合がある。（アンサンブル学習）。

4．ここではY として2個のものの識別を考えたが、3個以上の識別も同様にして行なうことができる。2個のものの識別を組み合わせることもできるが2個のものの組み合わせ方で最適なものはまだわかっていない。