Fisherの判別分析を超えて

新村秀一
成蹊大学 経済学部

Beyond Fisher's Linear Discriminant Function

Shuichi Shinmura
Dept. of Economics, Seikei Univ.

要旨

- 単純な判別規則： f(x)>0→class1， f(x)<0 → class2
- 判別分析の深い闇を隠してきた？
  - f(x)=0のケース？
  - 誤分類数0（線形分離可能）のデータを，LDFとQDFは認識不能.
  - QDFは，特定の理数系科目の合否判定で一方の群全てを誤判別.
  - 判別関数と誤分類数の関係が不明.
  - 判別分析は推測統計学と無縁.
  - Fisherの仮説を満たすデータは少ない.
    - そのため，QDFが開発された.
    - ロジスティック回帰の判別成績がLDFやQDFに比べて良いのは，データに対する仮説がないから.
- 以上の問題を，MNM基準による最適線形判別関数（OLDF）で解決！

キーワード: LDFとQDFの問題，MMN基準，SVM，IP，LP，QP，
MMN (Minimum Number of Misclassifications)の単調減少性

1．はじめに

- Fisher（1936）は，**分散比最大化基準**でFisherの線形判別関数（LDF）を定式化
  - 判別分析の世界を切り開いた[1]
- その後，多くの基準で判別関数が提案．
  - **Fisherの仮説**(2群が多次元正規分布し分散共分散が等しい)から，分散比最大化基準と同じLDFが定式化
    - 統計家の混乱
  - 分散共分散が等しくない場合に，2次判別関数（**QDF**）
    - QDFの提案自体，実データがFisherの仮説を満たさないことを示す．
  - **マハラノビスの汎距離**を用いた多群判別
  - 田口[2]のMT理論
  - ロジスティック回帰はデータに特定の分布を仮定しないため，判別成績がLDFやQDFに比べて良い．
    - 医療や金融の分野で利用
    - 地球モデル，MT理論と関連

数理計画法によるアプローチ

- **数理計画法は，関数の最大/最小を求める学問[3－4]**
- 回帰分析や判別分析も，容易に定式化．
  - L_1ノルムの和の最小化基準：MIN= $\Sigma\varepsilon_i$
  - Lpノルムの和の最小化基準：MIN= $\Sigma|\varepsilon_i|^p$
- 1978年にStam[5]は，総括論文で「なぜ統計ユーザーはLpノルム判別分析を利用しないのか?」と指摘し，この分野の研究は終焉．
- Vapnik[6]は，**SVM**を提案して今日まで研究は継続
  - H-SVM（ハードマージン最大化SVM)は，**線形分離可能**なデータから出発
    - 我々統計家は，**線形分離可能**なデータの判別は容易と勘違いし無視
    - H-SVMは，2群を分けるマージンを判別超平面の両側に等距離で設定し，この距離を最大化すれば**「汎化能力」**が良いとする点は評価できる
  - S-SVM（ソフトマージン最大化SVM）：幾つかのケースはマージンの反対側にきてその距離の和を最小化する第2基準を加えた2目的最適化
    - H-SVMの提案した汎化能力という新知見は無関係になる
    - 「**MIN=1/マージン＋c*($\Sigma\varepsilon_i$)**」で2目的を単目的化しているため，ペナルティcと呼ばれる重みを導入 （チューニングというトリック）．
    - 重みを決める**「最適化基準がない」**
      - cを任意に動かせば，マージン距離が最小の状態から最大の状態の任意の判別関数を試行錯誤で手当たり次第に求めている[7]
  - Kernel－SVM：カーブ・フィッティングで次数を上げれば当てはめが良くなる？

2. MNM最小化基準による最適線形判別関数

MNM(Minimum Number of Misclassifications)基準による判別は，教師データで判別成績が良くても，評価データで過評価しないのか?

2.1 背景
2.2 2次元における3個のケースの判別例
2.3 各種OLDFの定式化[11]
(1) IP-OLDFとLP-OLDF
(2) 改定IP-OLDFと改定LP-OLDF
(3) 改定IPLP-OLDF
2.4 MMNの有用性
2.5 SVMによる判別モデル

2.1 背景

- 大学卒業後，心電図診断論理の研究に従事したが，枝分かれ論理に惨敗.
  - 実データは，Fisherの仮説をみたすデータは少ない.
  - 正常と異常の2群判別を考えた場合，異常群は正常からある計測値が連続的に大きくなることで異常群が形成．すなわち正常を地球と考え，異常群は地球から噴き出た山と考える(地球モデル[8]).
  - 異常群の典型症例は，異常群の平均値でなく山頂
  - 判別超平面上に多くの異常ケースがくる(合否判定もおなじ).
  - 医師は異常所見として判別境界上の不確かなデータでなく，典型例を集める.
- 計測値が連続的に大きく(あるいは小さく)なることで，異常がより確かになることをBayseの定理を用いた「スペクトル診断[9]」を発表.
  - ロジスティック回帰が医学診断で用いられているが，Bayseの定理を用いるよりも操作性が良い.
  - ロジスティック回帰の考え方は，地球モデルの根拠を与えていると考える.
  - マハラノビス田口理論(MT理論)
- 複数の異常所見との2群判別を考えた場合，異なった複数個の異常所見を判別する説明変数は異なる.
  - これはパターン認識の大家の渡辺先生も「異なったパターンの識別は，識別空間が異なる」と指摘[10]

判別分析の深い闇

判別分析の**単純な判別規則**(**f(x)>0→クラス1，f(x)＜0→クラス2**)に隠れて
大きな問題が未解決のまま放置.

- f(x)＝0のケースをどちらに判別するかは未解決.
  - 統計ソフトでSASだけがこの問題を意識して，f(x)＝0の扱いをユーザーが指定可
  - JMPでは，ロジスティック回帰で2＊2の分割表に判別結果をまとめる際，どのクラスを陽性と指定するかを問い，判別境界上のケースを全て陽性に判別
- LDF，QDF，S-SVMは，**線形分離可能なデータを認識できない**.
  - 変数選択法は，MNM=0の最小次元を基準にして考えると一定の傾向を示さない.
  - ロジスティック回帰は，回帰係数の推定が不安定になりNM=0になれば，線形分離可能なデータを認識したことが改定IP-OLDFから分かる[11].
  - しかし，最小次元のMNM=0の空間を求める保証はない.
  - H-SVMは，線形分離可能なデータを識別できるが，それ以外のデータに適用できない.
- 判別係数と誤分類数の関係が不明.
  - **判別係数は，定数項が正と0と負の3つの異なった構造**をもち，
- MNMが最少な**最適凸体**の内点を判別係数とするOLDFを考えれば，判別分析の抱える問題が解決できる.
- LDFやQDFは，試験の合否判定という自明な線形分離可能なデータを認識できず，LDFでは誤分類確率が0.34，QDFでは0.9以上になる例もある.

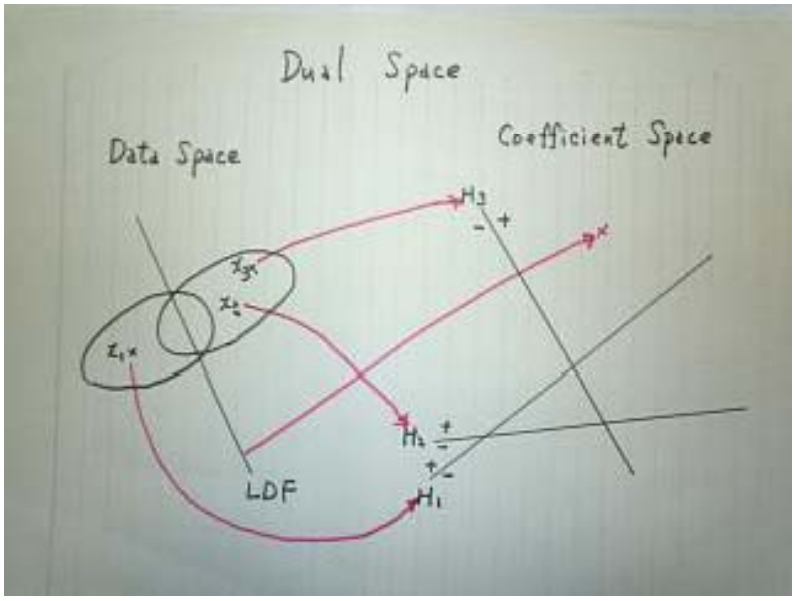
2.2 2次元における3個のケースの判別例

- データ空間
  - CLASS 1： $x_1 = (-1/18, -1/12)$,
  - CLASS 2： $x_2 = (-1, 1/2)$, $x_3 = (1/9, -1/3)$

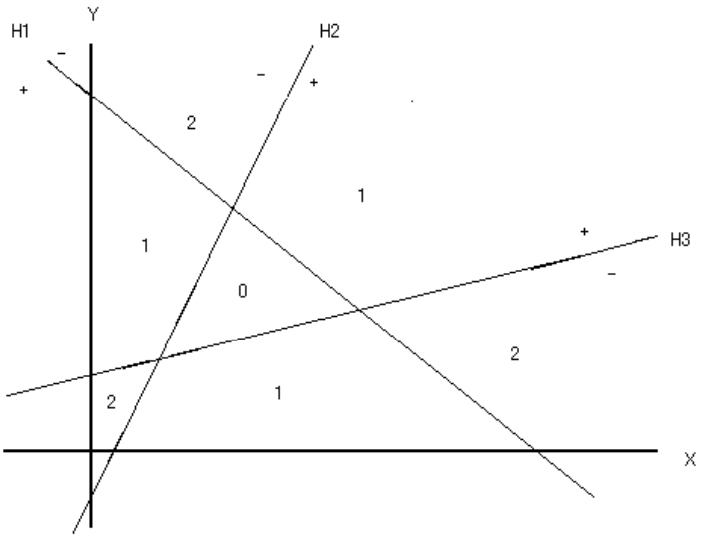
- IP-OLDFとLP-OLDF の定式化 （判別関数：$f(x)= b_1X_1+b_2X_2+1$)
  MIN=Σe_i；
  $-(1/18) \times b1 - (1/12) \times b2+1 \geqq -e_i$；
  $-1(-b1 + (1/2) \times b2 + 1) \geqq -e_i$；
  $-1((1/9) \times b1 - (1/3) \times b2 + 1) \geqq -e_i$；
  - e_iがO/1の整数変数，e_iが実数

- *判別係数上の線形超平面*
  - $H1 = -(1/18) \times b1 - (1/12) \times b2+1=0$,
  - $H2 = -b1 + (1/2) \times b2 + 1=0$,
  - $H3 = (1/9) \times b1 - (1/3) \times b2 + 1 =0$

データ空間と判別係数の空間の関係



判別係数上の線形超平面と最適凸体
3つの判別係数の構造



2.3 最適線形判別関数の定式化[11]

(1) IP-OLDFとLP-OLDF

- IP-OLDF

  $MIN = \Sigma e_i$;

  $y_i \times (\boldsymbol{x}_i'\boldsymbol{b}+1) >= -M*e_i$; (2.1) i=1,…,n

  $\boldsymbol{Y}_i$: クラス1 → 1, クラス2 → -1

  $\boldsymbol{x}_i$: 説明変数のデータ, $\boldsymbol{b}$: 判別係数,

  e_i: 判別されるケースは0, 1/0の整数値.
  誤分類されるケースは1

  誤分類されるケースの制約式は, $y_i \times (\boldsymbol{x}_i'\boldsymbol{b}+1) >= 0$ から$y_i \times (\boldsymbol{x}_i'\boldsymbol{b}+1) >= -M$に変わることで制約式が満たされる.

  目的関数は, この1になる誤分類数の和を最小化する.

- 配置行列が一般位置にある場合は最適凸体の頂点を求めるが, ない場合は正しい最適凸体の頂点を求めないこともある.
- e_iを非負の実数にしたものをLP-OLDFという.
  - L_1ノルム判別分析の一種である.
  - 誤分類されるケースの判別超平面上からの距離の和を最小化しても, 現実的に意味がない.

(2) 改定IP-OLDFと改定LP-OLDF

- 改定IP-OLDFは, パターン認識のマージン概念を取り入れて定式化(2.2)

  $MIN = \Sigma e_i$;

  $y_i \times (\boldsymbol{x}_i'\boldsymbol{b}+b_0) >= 1-10000 \times e_i$; (2.2)

  - 判別 : $y_i \times (\boldsymbol{x}_i'\boldsymbol{b}+1) >= 1$ から
  - 誤判別 : $y_i \times (\boldsymbol{x}_i'\boldsymbol{b}+1) >= -9999$ に変わる.
- 判別超平面上にケースはなくなり, 正しい最適凸体の内点が得られる.
- 改定LP-OLDFは, (2.2)のe_iを0/1の整数変数から非負の決定変数に変える.

(3) 改定IPLP-OLDF

- 改定IP-OLDFは，計算時間がかかる.
- 改定IPLP-OLDFは，MNMの近似解を求める
  - 第１段階：改定LP-OLDFを適用し，e_i =0になるSVで正しく判別されたケースを選ぶ
  - 第2段階：第１段階の選ばれたケースを0に固定.
    - 第1段階で1と誤判別されたものだけに改定IP-OLDFを適用し，MNMの近似値を高速で求める.
      - ４種の実データで2万件のリサンプリングデータを作成
    - これで**100重交差検証法の適用**が可能.
      - ４種の実データで100組のリサンプリングデータを作成
      - 100重交差検証法
- 実際の分析は，改定IP-OLDFを用いる

学生 データ (31 モデル, 20s vs. 40s)

Ind.Var.	IP	IPEC	LP	LPEC	SV	IPLP	IPLPEC	%
x1,x2,x3,x4,x5	3	2004	4	2391	8	3	2004	−3
x1,x2,x3,x5	3	2004	4	2350	11	3	2004	−3
x1,x2,x3,x4	3	2004	5	2737	11	3	2004	−3
x1,x3,x4,x5	3	2004	6	3464	12	3	2004	−3
x1,x2,x4,x5	4	2099	6	3170	13	4	2099	0
x2,x3,x4,x5	3	2004	6	3464	12	3	2004	−3
x1,x2,x3	3	2004	4	2350	11	3	2004	−3
x1,x3,x5	3	2004	4	2350	11	3	2004	−3
x1,x3,x4	5	2486	7	3803	13	5	2486	0
x1,x2,x4	5	2486	7	3803	13	5	2486	0
x1,x2,x5	3	2004	6	3464	12	3	2004	−3
x2,x3,x4	4	2637	7	4399	10	4	2637	−3
x2,x3,x5	3	2004	4	2350	11	3	2004	−3
x3,x4,x5	3	2004	4	2350	8	3	2004	−3
x1,x4,x5	6	3720	8	4527	15	6	3720	−4
x2,x4,x5	5	2808	7	3333	12	5	2808	−2
x3	8	4527	8	4527	15	8	4527	−3
x1	7	3587	6	2903	9	7	3587	0
x2	7	4641	3	2004	11	7	3628	−1
x4	13	6290	13	6290	13	13	6290	1
x5	15	10000	15	10000	15	15	10000	−13

Iris データ (15 モデル,446s vs.30s: 15倍)

Ind. Var.	IP	IPEC	LP	LPEC	SV	IPLP	IPLPEC	%
X1, X2, X3, X4	1	204	2	411	3	1	204	0
X2, X3, X4	2	411	2	411	5	2	411	-0.1
X1, X3, X4	2	414	2	414	6	2	414	-0.1
X1, X2, X4	4	799	7	1413	11	4	799	0
X1, X2, X3	2	402	3	616	10	2	402	0
X2, X4	5	1020	6	1232	11	5	1024	-0.1
X3, X4	3	622	6	1252	6	3	622	-0.1
X1, X3	4	817	5	1031	10	4	823	-0.1
X1, X4	5	1024	6	1232	10	5	1024	-0.1
X2, X3	6	1209	6	1209	14	6	1213	-0.1
X1, X2	25	4924	27	5391	61	25	4975	0.1
X4	6	1232	6	1232	10	6	1232	-0.2
X3	7	1413	7	1408	12	7	1408	0
X1	27	5362	25	4954	57	27	5362	0.2
X2	37	7351	34	6841	78	37	7351	0.2

銀行データ(63 モデル,133399s vs. 2688s: 50倍)

Var.	p	IP	IPEC	LP	LPEC	SV	IPLP	IPLPEC	%
X1, X2, X3, X4, X5, X6	6	0	0	0	0	0	0	0	0
X2, X3, X4, X5, X6	5	0	0	0	0	0	0	0	0
X1, X3, X4, X5, X6	5	0	95	0	0	0	0	0	0
X1, X2, X4, X5, X6	5	0	799	0	0	0	0	0	0
X1, X2, X3, X4, X6	5	0	807	0	531	0	0	531	-2.7
X1, X2, X3, X4, X5	5	1	371	2	496	3	1	389	-1.4
X1, X2, X3, X5, X6	5	1	115	1	115	1	1	115	-0.1
X3, X4, X5, X6	4	0	0	0	0	0	0	0	0
X2, X4, X5, X6	4	0	0	0	0	0	0	0	0
X1, X4, X5, X6	4	0	95	0	0	0	0	0	0
X2, X3, X4, X6	4	0	0	0	0	0	0	0	0
X1, X3, X4, X6	4	0	1303	0	531	0	0	531	-2.7
X1, X2, X4, X6	4	0	1303	0	531	0	0	531	-2.7
X4, X5, X6	3	0	0	0	0	0	0	0	0
X3, X4, X6	3	0	0	0	0	0	0	0	0
X1, X4, X6	3	*0*	*1303*	0	531	0	0	531	-2.7
X2, X4, X6	3	0	0	0	0	0	0	0	0
X3, X4, X5	3	2	198	3	282	5	2	198	0
X2, X4, X5	3	2	198	2	198	7	2	198	0
X4, X6	2	0	0	0	0	0	0	0	0

CPD(40 モデルs, 38170s vs. 380s : 100倍)

P	Type	IP	IPEC	%	LP	LPEC	%	SV	IPLP	IPLPEC	%
1	FBfb	20	2142	-2.4	20	2142	-2.4	50	20	2142	-2.4
2	FBfb	13	1815	-3.7	17	1931	-2.6	38	13	1815	-3.7
3	FBfb	12	1647	-3.2	18	1991	-2.5	37	12	1524	-2.6
4	Ffb	10	1285	-2.3	13	1378	-1.5	32	10	1285	-2.3
4	B	11	1468	-2.8	19	2159	-2.9	36	11	1468	-2.8
5	Ff	11	1468	-2.8	19	2159	-2.9	35	11	1468	-2.8
5	b	7	1043	-2.3	13	1477	-2.0	26	7	1043	-2.3
5	B	11	1468	-2.8	18	2094	-3.0	33	11	1468	-2.8
6	B	9	1136	-1.9	13	1469	-1.9	30	9	1136	-1.9
6	b	7	1043	-2.3	14	1626	-2.3	24	7	1043	-2.3
6	Ff	7	1043	-2.3	14	1523	-1.8	24	7	1043	-2.3
6	DOC1	12	1533	-2.7	18	2097	-3.0	35	12	1533	-2.7
6	DOC2	11	1361	-2.2	17	1927	-2.6	36	11	1361	-2.2
7	B	9	1136	-1.9	13	1469	-1.9	29	9	1136	-1.9
7	Ffb	6	887	-1.9	14	1523	-1.8	24	6	887	-1.9
8	F	6	887	-1.9	12	1289	-1.4	23	6	887	-1.9
12	fb	3	370	-0.6	8	855	-0.9	13	3	370	-0.6
13	FB	3	240	0.1	8	946	-1.4	14	3	370	-0.6
13	fb	3	390	-0.7	9	1020	-1.4	12	3	390	-0.7
14	FB	3	370	-0.6	7	777	-1.0	15	3	370	-0.6
14	fb	2	214	-0.2	7	807	-1.1	14	2	214	-0.2
15	FB	3	370	-0.6	8	855	-0.9	13	3	370	-0.6
15	fb	2	202	-0.2	5	591	-0.9	10	2	202	-0.2
16	FB	2	202	-0.2	5	482	-0.3	14	2	202	-0.2
16	fb	2	214	-0.2	5	481	-0.3	8	2	202	-0.2
17	FB	2	334	-0.8	8	744	-0.4	8	2	214	-0.2
18	FB	2	334	-0.8	5	591	-0.9	7	2	214	-0.2
19	FB	2	221	-0.3	6	542	-0.2	7	2	102	0.3

2.4 MMNの有用性

幾つかの新知見

（1） **MMNの単調減少性**

- *p*変数モデルの*MMN*をMMN_pとし，1変数追加した*(p+1)*変数モデルの$MMN_{(p+1)}$は，必ず減少する（$MMN_p \geqq MMN_{(p+1)}$）.
- 証明： 追加した説明変数の判別係数を0とすれば，その*(p+1)* 変数モデルの*誤分類数*はMMN_pになる．よって，*(p+1)*変数モデルの$MMN_{(p+1)}$はMMN_pより小さいか等しい．これで，変数増加法で選ばれるモデルの*MMN*も単調に減少する．

(2) 正規性からの乖離

- データがFisherの仮説を満たせば，得られた誤分類数はMMNに等しくなる.

- LDFの誤分類数がMMNから乖離するほど，データが仮説から乖離.
  - この点に関しては，実データと分散共分散行列が等しい2万件の正規乱数データを作成し，評価データとして分析した結果が裏付けた.
  - 実データで得られたLDFを評価データに適用したところ，評価データの誤分類数が教師データより少ない結果が多く現れた[12].
  - これは，実データから計算された分散共分散行列そのものが母集団であり，2万件の正規乱数はそこから忠実にサンプリングされた標本である．元データは単に分散共分散行列が等しい正規分布から偏った標本に過ぎない.
  - LDFはこの分散共分散行列から計算され，正規乱数データのほうが正規性からの乖離が少ないので，元データより誤分類数が少なくなるのは当然.

(3) MMNによる他の判別手法の評価

- 他の判別手法の誤分類数をMNMで単回帰分析し，比較評価.
- 誤分類数とMNMをプロットし評価.

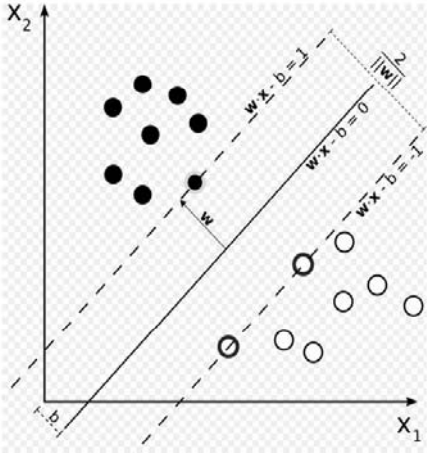
- 4種の実データを用いた評価法の方針
  - 原則全てのモデルで検証
    - 判別成績の良いモデルだけでは種々の困難な現実の問題点が明らかにならない.
    - 3種の実データではこれを行う
  - CPDデータは，約52万個のモデルがある[13-14].
    - 逐次変数選択法で代表的な40モデルを選んで評価に用いた.
    - 多重共線性があり，QDFは多重共線性の影響を強く受ける.
- 2変数の115組の正規乱数データでは，QDFは教師データで判別成績が良く，評価データで悪い[15].
- 実データと2万件の正規乱数データの比較で，Kernel SVMとの比較を行い，評価データでKernel SVMの汎化能力が悪い例 [12].
- 100重交差検証法
- 応用研究：合否判定

(4) 線形分離可能な最小次元のデータ空間の発見

- 改定IP-OLDFは，線形分離可能な最小次元のデータ空間を発見できる．
- 線形分離可能な最小次元の空間が分かれば，その変数を含む判別モデルは，MNMの単調減少性からすべて線形分離可能．
- スイス銀行紙幣データ[16]は，逐次F検定やAICでは5変数モデルを選ぶが，2変数で線形分離可能[17]
- 統計学では，この点に関して明快な説明ができない．
  - しかし，このデータは2次元で線形分離可能なためである．
  - 逐次F検定は追加した変数による偏差平方和の増分を検定して5変数モデルを選ぶ．
  - もし，2変数で線形分離可能であれば，それ以上のモデルを選択することは間違い．
  - 逐次変数選択法は，線形分離可能なデータに適用してはいけない．

2.5 SVMによる判別モデル

- H-SVM
  $MIN= // w // ^2/2=(a_1^2+a_2^2+\ldots+a_p^2)/2$;
  $y_i * g(\mathbf{x}_i)> 1$; $i=1,\ldots,n$
- S-SVM.
  $MIN= (a_1^2+\ldots+a_p^2)/2+c*\Sigma e_i$; (2.3)
  $y_i * g(\mathbf{x}_i)> 1-e_i$; $i=1,\ldots,n$
- 2目的最適化：L_1ノルムにペナルティCと呼ぶ重みを導入し，2目的を単目的化．
  - 単位の異なる多目的基準を加重和で単目的化することは問題
  - ポートフォーリオ分析のように，2次式で表される分散（リスク）を目的関数として最小化し，一次式で表される利益を制約式に取り込み，この利益水準を何段階かで変えて効率的フロンティアを描くべき
- 式(2.3)でCを何段階かで変えても，多くの結果が変わらない[12]．
  - より高速で効率的なアルゴリズムを開発したが，理論的に改定IP-OLDFより劣るので意味のないモデル．

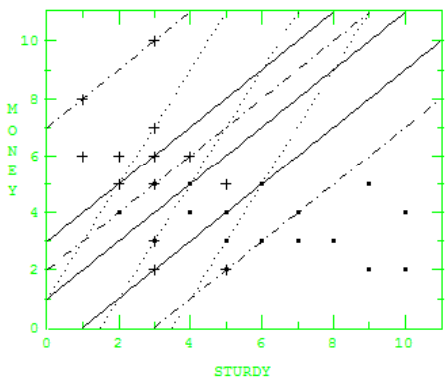


SVMと改定LP-OLDFの関係

				判別関数, 合格と不合格のSV	SV上のケース		
Cの値	L0	L1	SVM	支出=b*勉強+c	合格群	不合格群	NM
①$10^{6}$	0.25	14.5	1.4E7	Y=x+1, x-1, x+3	(6,5), (5,4)	(2,5), (3,6)	6+1
②$10^{-1}$	0.16	15	1.66	Y=2x-3, 2x-7, 2x+1	(6,5), (5,3)	(3,7), (2,5)	5+4
③$10^{-2}$	0.04	19	0.23	Y=x+2, x-3, x+7	(6,3) (5,2), (7,4)	(3,10), (1,8)	4+4
改定LP-OLDF	0.25	14.50		Y=x+1, x-1, x+3	(6,5), (5,4)	(2,5), (3,6)	6+1
改定IP-OLDF	0			Y=6E-6X+4.9999	(6,5)	(5,5)	5

- S-SVMは，改定LP-OLDFのBigM定数をいろいろ変えて，良い結果をつまみ食いする手法.

S-SVMの秘密



- Penalty cを10^{6}から10^{-6}まで13段階で変更
  - 実線（マージン最小）→破線->一点鎖線（マージン最大）まで変わっていく

3. データと評価方法

3.1 検証のためのデータ

- **4種の実データ(教師データ)**
  - Fisherのアイリスデータ:セトサ，バーシクル，バージニカ各50例，4個の計測値.
    - セトサは，他の2群と完全に判別できるので，これを省いた15個の2群判別 [1].
  - CPD(児頭骨盤不均衡)データ:自然分娩群(180例)と帝王切開(60例)の2群判別.
    - 17個の計測値と，2組の計測値の差の19個の説明変数を:3個の多重共線性.
    - 52万個の判別モデルがあり変数選択法で選んだ40個のモデルで検討[13-14].
  - スイス銀行紙幣データ[16]:1000スイスフラン紙幣データの各100枚の真札と偽札
    - 6個の計測値による2群判別.
    - 逐次F検定やAICで5変数モデルが選ばれるが，2変数で線形分離可能[11].
  - 学生の成績データ: 40人の学生の成績とそれに関する5変数のデータ.
    - 便宜上70点以上を合格，未満を不合格とした2群判別に [18].

- 評価データ
  - 実データと同じ平均と分散共分散をもつ正規乱数
  - 実データから，2万件のリサンプリング標本
  - 実データから100組の標本をリサンプリングし，100重交差検証法に用いた[11].

3.2 スイス銀行紙幣データによる教師データの評価

Var.	p	LDF	MNM
X1—X6	6	1	0
X2—X6	5	1	0
X1, *X3*—X6	5	1	0
X1, *X2*, X4—X6	5	1	0
X1—X4, X6	5	1	0
X1—X5	5	7	2
X1—X3, X5, X6	5	2	1
X4—X6	3	1	0
X3, X4, X6	3	1	0
X1, X4, X6	3	1	0
X2, X4, X6	3	1	0
X4, X6	2	3	0

- (X4,X6)でMNM=0で，これを含む全モデルでMNM=0
  - 変数選択法は，5または6変数を選ぶ
  - この結果は普遍的でないのでは？という意見.
- 誰もが簡単に確認できる方法の開発
  - 誤分類数が0でないデータの2群間の距離を拡大すれば，MNM=0のデータを作成可.
  - AIC, Cpなどは，同じモデルを選ぶ

CPDデータ(距離2倍)

実データと距離2倍データは、Cpは6変数、AICは5変数と同じモデルを選ぶ。
実データのMNMは6変数が最小で、距離2倍データは1変数でMNM=0と、異なったモデルを選ぶ

実データ							距離2倍データ				
変数	p	Cp	AIC上昇	AIC下降	MNM	LDF	Cp	AIC上昇	AIC下降	MNM	LDF
X2, X9, X12, X15, X17, X18	6	7		-589.7	8	16	7	-828.6	-828.6	0	0
X2, X9, X12, X15, X18	5	6	-590.5	-590.5	10	17	7	-828.9	-828.9	0	0
X9, X12, X15, X18	4	6	-590.9		10	17	7	-828.5		0	0
X9, X12, X18	3	6	-590.7		12	19	8	-827.2		0	1
X9, X12	2	8	-588.8		13	17	12	-823.4		0	1
X12	1	24	-573.1		19	23	40	-798.3		0	2

アイリスデータ(距離2倍)

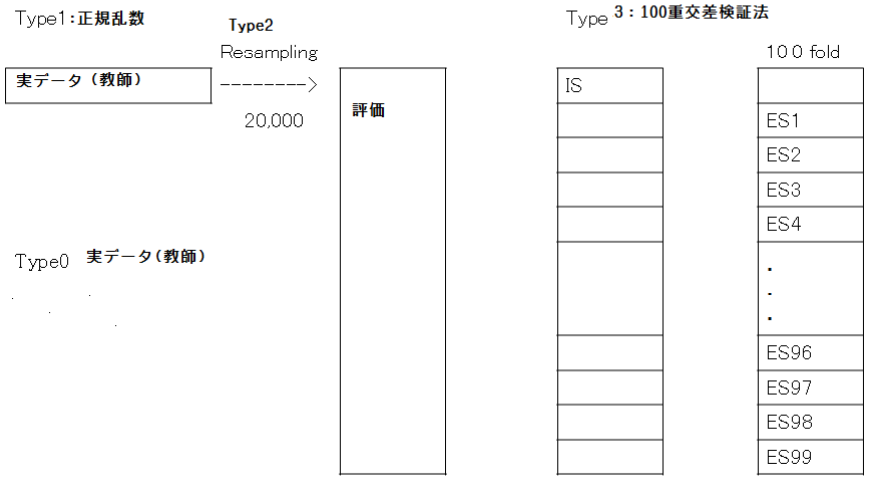
実データ						距離2倍データ				
変数	p	AIC上昇	MNM	LDF	差	Cp	AIC上昇	MNM	LDF	差
x1-x4	4	-281.8	1	3	2	5.0	-402.8	0	0	0
x2, x3, x4	3	-276.4	2	4	2	11.9	-395.8	0	0	0
x1, x3, x4	3		2	3	1	15.9		0	0	0
x1, x2, x4	3		3	5	2	33.3		0	0	0
x1, x2, x3	3		2	7	5	50.9		0	0	0
x2, x4	2	-261.3	3	5	2	34.3	-376.7	0	0	0
x3, x4	2		3	6	3	36.7		0	0	0
x1, x3	2		3	6	3	50.3		0	0	0
x1, x4	2		5	6	1	57.4		0	0	0
x2, x3	2		5	7	2	93.0		0	0	0
x1, x2	2		24	25	1	545.1		11	14	3
x4	1	-259.4	5	6	1	55.4	-362.2	0	0	0
x3	1		5	8	3	100.9		0	0	0
x1	1		24	27	3	546.4		12	15	3
x2	1		29	42	13	942.0		19	17	-2

銀行データ

		距離1.25倍				元データ			
Model	p	Cp	AIC(上昇)	MMN	LDF	Cp	AIC(上昇)	MMN	LDF
X1-X6	6	7.0	(-863.5)	0	0	7.0	(-779.4)	0	1
X2-X6	5	5.3	-864.8	0	0	5.3	-781.00	0	1
X3-X6	4	10.5	-895.5	0	0	10.3	-776.00	0	1
X4- X6	3	10.9	-859.1	0	0	10.7	-775.60	0	1
X4, X6	2	111.8	-779.1	0	0	107.0	-698.50	0	3
X6	1	3.53.9	-678.8	0	1	292.0	-603.90	2	2

		距離0.75倍				距離0.5倍			
Model	p	Cp	AIC(上昇)	MMN	LDF	Cp	AIC(上昇)	MMN	LDF
X1-X6	6	7.0	(-675.8)	1	2	7.0	(-543.5)	5	12
X2-X6	5	5.3	-677.50	1	2	5.3	-544.80	6	12
X3-X6	4	9.8	-672.80	1	1	8.9	-541.10	7	13
X4- X6	3	10.1	-672.60	1	2	8.8	-541.10	8	14
X4, X6	2	97.9	-601.00	4	6	78.7	-481.90	16	19
X6	1	253.8	-516.60	6	8	184.4	-417.40	53	56

4. 100重交差検証法



4. 100重交差検証法

	LDF-IPLP			
	教師		評価	
	最小値	最大値	最小値	最大値
アイリス(15)	0.55	5.23	-0.60(2)	2.36
銀行(63個中35個)	0.00	3.63	-0.01(1)	4.35
学生(31)	1.46	8.61	-1.29(3)	7.11
CPD(26)	3.05	7.28	2.21	6.15
	Logi-IPLP			
	教師		評価	
	最小値	最大値	最小値	最大値
アイリス(15)	0.59	5.31	-0.84(2)	1.85
銀行(63個中35個)	-0.28(1)	3.47	-0.04(1)	4.43
学生(31)	-2.12(3)	6.48	-2.89(7)	5.59
CPD(26)	0.13	3.43	0.29	1.74

5. 2010年と2011年の統計入門の試験の合否判定[19]

- 10択100問の試験を実施.
  - 正解と不正解が1/0の値をもつ説明変数.
- 合格水準を得点の10%点，50%点，90%点で検討.
- 大問と小問100問で検討

5.1 大問の分類

大問	中間試験			期末試験		
	内容	得点	小問番号	内容	得点	小問番号
T1	基礎統計量	29	1-8, 21-41	計算	26	1-26
T2	計算	12	9-20	相関回帰	30	27-56
T3	正規分布	19	42-60	分割表	21	57-77
T4	JMPの解釈	40	61-100	同左	23	78-100

表5 2010年と2011年の中間大問の合否判定

P	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi
1	T4	6	11	11	9	T4	16	16	16	16	T3	10	24	24	27
2	T2	2	11	9	6	T3	9	12	12	10	T4	5	20	11	10
3	T1	1	8	5	3	T1	2	5	6	2	T1	0	20	10	0
4	T3	0	9	2	0	T2	0	3	6	0	T2	0	20	11	0

P	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi
1	T2	9	15	15	17	T4	9	9	9	9	T3	6	14	14	7
2	T4	4	11	9	9	T1	4	5	7	4	T4	1	14	6	1
3	T1	0	9	10	0	T3	1	3	3	2	T1	0	13	5	0
4	T3	0	9	11	0	T2	0	3	3	0	T2	0	14	9	0

表6 2010年度の中間試験の小問の判別結果

4問が全員正解で，フルモデルは96問
QDFは，フルモデルで一方の群を他方に誤判別
ロジスティック回帰は，90%で最小次元を発見できない

P	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi
1	36	11	11	11	11	93	28	28	28	28	57	13	19	19	18
6	61	0	2	1	0	99	9	9	12	9	59	5	8	7	5
7	38					84	6	7	7	6	58	4	9	7	4
12	65					22	0	2	4	0	3	1	6	13	1
13	72					67					63	0	4	13	1
14	99					34					8		5	13	0
96	46		0	109	0	44		0	61	0	30		0	13	0

表7 2010年度の期末試験の判別結果

1問が全員正解で，フルモデルは99問
QDFは，フルモデルで一方の群を他方に誤判別
ロジスティック回帰は，50%と90%で最小次元を発見できない

P	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi
1	18	12	15	15	15	22	12	26	26	26	68	12	29	29	29
・															
10	4	1	5	2	1	63	6	6	10	4	93	6	5	13	2
11	1	1	5	111	1	90	1	6	9	1	34	0	6	13	1
12	2	0	5	111	0	19	0	4	4	1	70		4	13	0
13	44					14		3	4	1	63				
14	54					41		3	3	1	42				
31	75					10		1	62	1	14				
32	41					62		1	62	0	12				
99	73		0	111	0	71		0	62	0	58		0	13	0

表8 2011年度の中間試験の判別結果

2問が全員正解で，フルモデルは98問
QDFは，フルモデルで一方の群を他方に誤判別

P	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi
1	13	9	9	9	9	84	19	19	19	19	54	9	28	9	28
・															
8	89	2	5	107	9	58	2	7	3	2	98	1	6	9	5
9	22	2	2	107	7	83	2	5	5	2	94	0	6	9	0
10	17	2	3	107	2	23	2	5	5	2	56				
11	90	1	4	107	2	82	2	5	5	3	52				
12	14	0	2	107	0	63	1	5	5	1	82				
13	18					26	1	5	6	1	40				
14	78					52	1	4	5	1	55				
15	69					98	0	3	6	0	78				
98	87		0	107	0	91		0	61	0	51		0	9	0

表9 2011年度の期末試験の判別結果

3問が全員正解で，フルモデルは97問
QDFは，フルモデルで一方の群を他方に誤判別

P	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi	Var	MNM	LDF	QD	Logi
1	14	10	10	10	10	32	30	30	30	30	100	12	23	23	23
・															
7	81	1	4	3	2	79	8	12	12	10	49	1	3	12	1
8	31	0	4	4	0	83	6	9	8	6	95	0	2	12	0
9	25					34	4	9	8	7	92				
10	41					89	4	9	5	6	39				
11	58					12	3	8	7	5	19				
12	12					46	2	6	7	4	70				
13	91					30	0	6	7	0	62				
97	73		0	110	0	11		0	62	0	99		0	12	0

QDFが誤判別する現象面的な理由

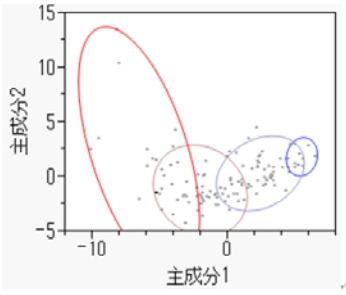


図2 2011年度の中間試験と期末試験の大間のスコアプロット

- QDFが，合否の一方の群を他方に誤判別するのは，数学や統計に限定．
- JMPが正則化法に切り替えても起きる場合がある
- スコアプロット（95%正規分布）
  - 左から右は，3水準で分かれる4群
  - 90%以上の成績優秀群は小さくなる
  - 10%の合否判定の場合，合格群は不合格群と直角になることが多い

6. 終わりに

- 本研究では，1998年から12年間行ってきたOLDFの理論的背景と，100重交差検証法でMNM基準が頑強なことを示した．
- 応用研究の一つとして，LDFとQDFは線形分離可能なデータを認識できないばかりか，誤分類数が大きいことを合否判定データで検討した．
  - 大問では4問全てを用いても合否判定ができない．
  - 小問では，*MNM=0*になる最小の説明変数のモデルで合否判定ができなかった．フルモデルでは，LDFの誤分類数は0になったが，QDFは一方の群が全て誤判別された．
- 名義ロジスティック回帰は，教師データに合わせて導出．
  - このため，多くの事例で判別結果が良いため，LDFとQDFに代わって医療や金融で近年多用．
  - 回帰係数の推定値が全て不安定になるまで変数を追加すると，多くの判別結果でOLDFが見つけた最小次元の特徴空間を発見．
    - 2010年度の中間の90%，期末の50%と90%で，最小次元の特徴空間を見つることができなかった．

本研究の重大な影響

- 今回の結果は次のような重大な懸念が考えられる．
- 過去に行われた医学診断やパターン認識などで，誤分類確率が0でなくても，線形分離可能であることを見過ごしてきた可能性．
- 判別分析の応用としてゲノム判別が研究．
  - 少ないケースで多くの説明変数から分散共分散を推定する研究が行われている
  - 分散共分散行列に基づくLDFとQDFに問題があるので，適用には注意する必要がある．

SAS社への提言

- 筆者の最適線形判別関数は，
  - SAS社，IBM社（SPSS+ILOG)，NTTデータ（数理システムの子会社化）の3社が開発できる
  - 統計ソフトのトップランナーとして，SAS社が提供すべき
- 筆者は，最大限協力したい

文献

[1] Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7,179–188.
[2]田口玄一(1999). タグチメソッドわが発想法. 経済界.
[3]新村秀一(2007). ExcelとLINGOで学ぶ数理計画法. 丸善.
[4]新村秀一(2011). 数理計画法による問題解決法. 日科技連出版社.
[5]Stam, A. (1997). Nontraditional approaches to statistical classification : Some per-spectives on Lp-norm methods. Annals of Operations Research, 74, 1-36.
[6] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, 1995.
[7] 新村秀一（2006). 改定IP-OLDFによるSVMのアルゴリズム研究. オペレーションズ・リサーチ, 51/11, 702-707.
[8] 新村秀一(1984). 医療データ解析，モデル主義，そしてOR. オペレーションズ・リサーチ. 29-7, 415-421.
[9]新村秀一，北川護，高木義人，野村裕(1973). 二段階重みづけによるスペクタル診断. 第12回日本ME学会大会論文集，107-108.
[10]渡辺慧（1978). 認識とパターン. 岩波書店.
[11] 新村秀一(2010). 最適線形判別関数. 日科技連出版社.
[12]新村秀一，ユンイエブン（2007). OLDFとSVMの比較研究(4)－種々のデータによるSVMとの比較-. 成蹊大学経済学部論集，37-2，89－119.
[13]新村秀一，三宅章彦(1983). 重回帰分析と判別解析のモデル決定（1）－19変数をもつC.P.Dデータの多重共線性の解消 －. 医療情報学，3-3.507-124.
[14]新村秀一(1996). 重回帰分析と判別分析のモデル決定(2)－19変数を持つCPDデータのモデル決定－. 成蹊大学経済学部論集，第27巻第1号，180-203.
[15]新村秀一，垂水共之(1999). 2変数正規乱数データによるIP-OLDFの評価. 計算機統計学12-2，107-123.
[16]Flury,B. & Rieduyl, H. (1988). Multivariate Statistics: A Practical Approach. Cambridge University Press.
[17]後藤昌司(2002). 統計学科学における事例の解剖.. 計算機統計学, 15(2), 185-217.
[18] 新村秀一（2004). JMP活用 統計学とっておき勉強法. 講談社.
[19] 新村秀一（2011). 合否判定データにおける判別分析の問題点. 応用統計学，3，157-173.
[20] Firth,D.(1936).Bias reduction of maximum likelihood estimates. Biometrika,80,27-38.