流通科学大学論集―経済・情報・政策編―第 23 巻第 2 号，47-73（2015）

データマイニング・ケーススタディ
― ある予測問題の観点から ―

Data Mining ― A case study on some prediction problem ―

磯貝　恭史*
Takafumi Isogai

データマイニングの分類問題で，よく利用されている手法群に対し，ある品質管理問題で調査されたデータを用いて，それらの予測精度の比較を行った結果を報告する．取り上げた手法は，線形判別，MT システム，ロジスティック判別，ニューラルネット，サポート・ベクター・マシーン，k 近傍法，バギング，ブースティングの 8 手法である．
キーワード：多変量解析，機械学習，学習データ，検証データ，計算機統計学

Ⅰ．はじめに

データマイニングという言葉は，昨今，大量のデータ（ビッグデータ）から有用な情報を引き出すデータ解析手法，あるいは，その集合体という意味で用いられている．そして，大量のデータとは，通常は観測単位の個数の多さを意味する大標本データであり，その取り扱いの妥当性については数理統計学の大標本理論などが中心的役割を果たしてきた．しかしながら，計算機の発展とデータベース構築の技術の進歩により，観測単位 1 個あたりの観測特性の記録項目が飛躍的に増加し，何百次元という観測特性を持つ高次元のデータが随所に現れるようになった．この高次元データも，観測特性の多さという意味での情報を大量に持つビッグデータだと考えられている．高次元データで観測単位の個数が次元数より少ないものは，通常の統計解析手法では一括して取り扱うことが難しい．

このようなデータを自由自在に処理するために，計算機科学の中の機械学習の手法（Vapnik（1998）[11] , Cristianini & Shawe-Taylor（2000）[3] , Bishop（2006）[1]），計算機統計学の手法（Giudici & Figini（2009）[6]）が導入されてきた．ビッグデータを取り扱うデータマイニングは，従来の多変量解析法（小西（2013）[8]）を含め，これら機械学習の手法および計算機統計学の手法から構成されている．今では，これらのデータマイニングの様々な手法は，「統計的学習理論」として，統一的な視点から解説されるようになってきている（Hastie et al.（2009）[7]）．

*流通科学大学商学部、〒651-2188　神戸市西区学園西町 3-1

（2014 年 9 月 2 日受理）

データマイニングは多くの分野で利用されているが，その情報を解説書から得ようとすると，手法についての説明の煩雑なものが多い．また，適用事例についても各分野特有の知識が必要とされ簡単には理解しがたいことがある．

データマイニングの考え方を簡単に理解するにはどうしたらよいか？一つの道としては，分かり易いデータを用いて，ある共通のトピックについて多くのデータマイニングの手法を適用してみることである．というわけで，この小論では，実験で計画的に採取されたデータを用いて，予測の観点から，各種の手法をそのデータに適用した結果を報告する．

論文構成としては，第 2 章で我々が用いるデータの概要を述べる．第 3 章ではデータマイニングの各手法の簡単な解説とデータに適用した結果を述べる．採りあげる手法は，線形判別，MT システム，ロジスティック判別，ニューラルネット，サポート・ベクター・マシーン，k 近傍法，バギング，ブースティングの 8 手法である．

Ⅱ．データの概要

このケーススタディで用いるデータの簡単な説明を行う．データはアルミボトルを用いた飲料の内圧測定に関する調査データである．アルミボトル飲料の内圧（缶内圧）が低すぎると不良品と判定される．通常は，缶内圧の高低の検査のために，アルミボトルの上部に急激な振動による衝撃を与えて，その打検音を観測して良不良の判定を行っている．打検音はそのアナログデータが直接利用されるわけではなく，スペクトルデータに変換されて用いられている．各スペクトルの観測値は 512 個の周波数（横軸）に対する振幅（縦軸）という形で与えられる 512 次元データである．スペクトルデータの詳細については，夏木・打田・磯貝（2012）[9] 参照．

我々の用いるデータベースは，実験により様々な内圧の下での打検音のスペクトルデータが集められた調査データである．各アルミボトルから得られるデータの構成は

（スペクトルデータ，内圧の実測値，良不良の判定，実験条件）

となっている．調査データの全観測データ数は 900 である．

このケーススタディでは，アルミボトルの打検音のスペクトルデータから内圧の良不良（高低）の判定がどの程度予測可能かを，各種手法を用いて調べる．

予測精度の評価は次のように行う．まず，900 個のデータを無作為に二つに分けて，学習用データ（観測数 650）と検証用データ（観測数 250）とする．学習用データを用いて，各種手法の予測式を求め，その予測式を用いて検証用データのスペクトルデータから良不良の判定の予測を行う．予測が間違うとき，すなわち誤判別を行う場合には二つのケース（良品を不良品と判定するか，あるいは，不良品を良品と判定する）がある．その誤判別の件数を合計し，その合計を検証用データの観測数 250 で割って予測精度の比較を行う．

スペクトルデータを取り扱うとき，手法の中に 512 次元データを取り扱えないものも存在する

ため，このケーススタディでは 512 次元データから部分集合を取り出して，320 次元のスペクトルデータを用いて手法の予測精度の評価を行う．

スペクトルデータの記号としては

一般的には $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$，データとしては $\mathbf{x}_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)^T\ (i=1,2,\cdots,n)$

を用いる．$\mathbf{x}$ の中の $x_i\ (i=1,\cdots,p)$ は取り上げた周波数に対する振幅を表している．$\mathbf{x}$ は，通常，共変量などと呼ばれている．また，良不良の判定に関する記号としては，y を用いて

$$y=\begin{cases}1\Leftrightarrow\ \text{良品}\\ -1\Leftrightarrow\text{不良品}\end{cases}\quad\text{または}\quad y=\begin{cases}1\Leftrightarrow\ \text{良品}\\ 0\Leftrightarrow\text{不良品}\end{cases}$$

などで表す．y をラベル変数と呼ぶ．

Ⅲ．データマイニング

１．線形判別

まず，フィッシャーの線形判別法の簡単な紹介を行う．データが二つの群 G_1 と G_2 から取り出されているとして，群 G_1 と G_2 から取り出された p 次元データを

$$\mathbf{x}_i^{(1)}=\left(x_{i1}^{(1)},x_{i2}^{(1)},\cdots,x_{ip}^{(1)}\right)^T\in G_1\qquad(\text{i}=1,2,\cdots,n_1),$$

$$\mathbf{x}_i^{(2)}=\left(x_{i1}^{(2)},x_{i2}^{(2)},\cdots,x_{ip}^{(2)}\right)^T\in G_2\qquad(\text{i}=1,2,\cdots,n_2)$$

とする．新しく p 次元データ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときに，この $\mathbf{x}$ が群 G_1 か群 G_2 のどちらに属するのかを判定するのが我々の課題である．

一つの解決策が，群 G_1 と G_2 を描写する確率分布を導入することで得られる．確率分布として p 次元正規分布を採りあげると，p 次元正規分布の確率密度関数は

$$f(\mathbf{x}|\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

で与えられる．ここで平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 Σ は

$$\boldsymbol{\mu}=\left(\mu_1,\mu_2,\cdots,\mu_p\right)^T=\left(E[x_1],E[x_2],\cdots,E[x_p]\right)^T,$$
$$\Sigma=\left(\sigma_{ij}\right)=\left(E\left[(x_i-\mu_i)(x_j-\mu_j)\right]\right)=E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right]\quad(i=1,2,\cdots,p;j=1,2,\cdots,p)$$

で定義される．記号 $E[\bullet]$ は密度 $f(\mathbf{x}|\boldsymbol{\mu},\Sigma)$ に関する期待値を示す．

群 G_1 と G_2 の確率密度関数 $f(\mathbf{x}|G_1)$ と $f(\mathbf{x}|G_2)$ を

$$f\left(\mathbf{x}\middle|G_1\right)=f(\mathbf{x}\middle|\boldsymbol{\mu}_1,\Sigma_1)=\frac{1}{(2\pi)^{p/2}\left|\Sigma_1\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)^T\Sigma_1^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)\right\},$$
$$f\left(\mathbf{x}\middle|G_2\right)=f(\mathbf{x}\middle|\boldsymbol{\mu}_2,\Sigma_2)=\frac{1}{(2\pi)^{p/2}\left|\Sigma_2\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)^T\Sigma_2^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)\right\}$$

で与える．さて，新しく p 次元データ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときに，確率密度関数 $f\left(\mathbf{x}\middle|G_1\right)$ と $f\left(\mathbf{x}\middle|G_2\right)$ を用いて，

$$f\left(\mathbf{x}\middle|G_1\right)\geq f\left(\mathbf{x}\middle|G_2\right)\Rightarrow\mathbf{x}\in G_1$$
$$f\left(\mathbf{x}\middle|G_1\right)<f\left(\mathbf{x}\middle|G_2\right)\Rightarrow\mathbf{x}\in G_2$$

と判定することにする．すなわち，$\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ の確率密度の値が大きい群に $\mathbf{x}$ を割り当てるのである．この判定方式を書き直せば

$$\log\frac{f\left(\mathbf{x}\middle|G_1\right)}{f\left(\mathbf{x}\middle|G_2\right)}\begin{cases}\geq 0\Rightarrow\mathbf{x}\in G_1\\ <0\Rightarrow\mathbf{x}\in G_2\end{cases}$$

となる．ここで，群 G_1 と G_2 の分散共分散行列 Σ_1 と Σ_2 が等しい（$\Sigma_1=\Sigma_2=\Sigma$）と仮定すれば，判定方式は

$$\log\frac{f\left(\mathbf{x}\middle|G_1\right)}{f\left(\mathbf{x}\middle|G_2\right)}=\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^T\Sigma^{-1}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right)\begin{cases}\geq 0\Rightarrow\mathbf{x}\in G_1\\ <0\Rightarrow\mathbf{x}\in G_2\end{cases}$$

と簡単になって $\mathbf{x}$ に関する線形式が得られる．この線形式

$$L(\mathbf{x})=\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^T\Sigma^{-1}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right)$$

をフィッシャーの線形判別関数と呼ぶ．

通常，群 G_1 と G_2 の $\boldsymbol{\mu}_1,\Sigma_1,\boldsymbol{\mu}_2,\Sigma_2$ は未知なので，観測データから推定する．各推定値は

$$\mathbf{x}_i^{(1)}\in G_1\ (i=1,\cdots,n_1)\Rightarrow\hat{\boldsymbol{\mu}}_1=\bar{\mathbf{x}}^{(1)}=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_i^{(1)},\ \hat{\Sigma}_1=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\mathbf{x}_i^{(1)}-\bar{\mathbf{x}}^{(1)}\right)\left(\mathbf{x}_i^{(1)}-\bar{\mathbf{x}}^{(1)}\right)^T$$
$$\mathbf{x}_i^{(2)}\in G_2\ (i=1,\cdots,n_2)\Rightarrow\hat{\boldsymbol{\mu}}_2=\bar{\mathbf{x}}^{(2)}=\frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{x}_i^{(2)},\ \hat{\Sigma}_2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(\mathbf{x}_i^{(2)}-\bar{\mathbf{x}}^{(2)}\right)\left(\mathbf{x}_i^{(2)}-\bar{\mathbf{x}}^{(2)}\right)^T$$

となる．群 G_1 と G_2 の分散共分散行列 Σ_1 と Σ_2 が等しい（$\Sigma_1=\Sigma_2=\Sigma$）場合には，Σ の推定値は

$$\hat{\Sigma}=\frac{1}{n_1+n_2-2}\left\{\left(n_1-1\right)\hat{\Sigma}_1+\left(n_2-1\right)\hat{\Sigma}_2\right\}$$
$$=\frac{1}{n_1+n_2-2}\left\{\sum_{i=1}^{n_1}\left(\mathbf{x}_i^{(1)}-\bar{\mathbf{x}}^{(1)}\right)\left(\mathbf{x}_i^{(1)}-\bar{\mathbf{x}}^{(1)}\right)^T+\sum_{i=1}^{n_2}\left(\mathbf{x}_i^{(2)}-\bar{\mathbf{x}}^{(2)}\right)\left(\mathbf{x}_i^{(2)}-\bar{\mathbf{x}}^{(2)}\right)^T\right\}$$

で与えられる．$\hat{\Sigma}_1,\hat{\Sigma}_2,\hat{\Sigma}$ は不偏分散共分散行列である．

このとき，推定値を用いたフィッシャーの線形判別関数による判定方式は

$$L(\mathbf{x})=\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right)^T\hat{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right)^T\hat{\Sigma}^{-1}\left(\bar{\mathbf{x}}^{(1)}+\bar{\mathbf{x}}^{(2)}\right)\begin{cases}\geq 0\Rightarrow \mathbf{x}\in G_1\\ <0\Rightarrow \mathbf{x}\in G_2\end{cases}$$

で与えられる．以下の例 1 では，この線形判別関数を用いる．

例１．フィッシャーの線形判別関数

まず,我々の調査データから取り出した学習データを用いて,線形判別関数の推定式を求める．推定式の学習程度を見るために学習データのスペクトルデータを用いて良不良の判定の予測を行い，誤判別率を求める．これを学習精度と呼ぶ．つぎに，推定式を検証用のスペクトルデータに適用して，アルミボトル缶の良不良の判定を予測して誤判別率（予測精度）を求めた結果を表 1 に示す．

表１．フィッシャーの線形判別関数の予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	293	13	306
良品	8	336	344
誤判別率＝(13+8)/650=0.0323			

予測精度

実測	予測		
	不良品	良品	合計
不良品	80	34	114
良品	32	104	136
誤判別率＝(34+32)/250=0.264			

表 1 の結果を見れば，学習精度は誤判別率が 3.2%であるので，推定式はかなりの判別力があると思われるが，予測精度は誤判別率が 26.4%であるので，4 回に 1 回は間違うという程度である．

２．MT システム

群 G のデータ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が p 次元正規分布に従うとき，その確率密度関数は

$$f(\mathbf{x}|\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right\}$$

で与えられる．このときの確率密度関数の等高線は

$$f(\mathbf{x}|\boldsymbol{\mu},\Sigma)=\text{定数}(>0)$$

で定義され，中心が $\boldsymbol{\mu}$ の楕円体の形状をしている．これを書き直せば

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)=\text{定数}(>0)$$

となり，正値行列 Σ^{-1} を用いた $\mathbf{x}-\boldsymbol{\mu}$ の 2 次形式で等高線が与えられることが判る．この 2 次形式を $\mathbf{x}$ と平均 $\boldsymbol{\mu}$ との平方距離と見なしたものがマハラノビスの平方距離

$$D^2=\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)$$

である．

群 G のデータ $\mathbf{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)^T \ (i = 1,2,\cdots,n)$ が与えられたとき，次の推定値 $\hat{\boldsymbol{\mu}}$ と $\hat{\Sigma}$

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i,\ \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$$

を用いて群 G のデータのマハラノビスの平方距離

$$D_i^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T \hat{\Sigma}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \quad \left(i = 1,2,\cdots,n\right)$$

が求められる．

MT システムとは群 G として良品の群などの興味ある群を取り上げ，群 G のデータについて，まず，マハラノビスの平方距離 $D_i^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T \hat{\Sigma}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \ \left(i = 1,2,\cdots,n\right)$ を計算し，D_i^2 の最大値 $\max_i D_i^2$ を求める．つぎに，新しいデータ $\mathbf{x} = \left(x_1, x_2, \cdots, x_p\right)^T$ が与えられたときに，$\mathbf{x}$ が群 G に属するかどうかをマハラノビスの平方距離 $\left(\mathbf{x} - \bar{\mathbf{x}}\right)^T \hat{\Sigma}^{-1}\left(\mathbf{x} - \bar{\mathbf{x}}\right)$ を用いて

$$\max_i D_i^2 - \left(\mathbf{x} - \bar{\mathbf{x}}\right)^T \hat{\Sigma}^{-1}\left(\mathbf{x} - \bar{\mathbf{x}}\right)\begin{cases}\geq 0 \Rightarrow \mathbf{x} \in G \\ < 0 \Rightarrow \mathbf{x} \notin G\end{cases}$$

と判定する方式である．MT システムの詳細については，立林ら（2008）[10] を参照．

例２．MT システム

我々の調査データから取り出した学習データを用いて MT システムを構成する．学習データの中に良品と不良品の群があるので，良品の群を G_1，不良品の群を G_2 とする．良品の群 G_1 を用いて，$\boldsymbol{\mu}_1$ と Σ_1 の推定値 $\hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{x}}^{(1)}$ と $\hat{\Sigma}_1$ を求め，マハラノビス平方距離

$$\left\{D_i^{(1)}\right\}^2 = \left(\mathbf{x}_i^{(1)} - \bar{\mathbf{x}}^{(1)}\right)^T \hat{\Sigma}_1^{-1}\left(\mathbf{x}_i^{(1)} - \bar{\mathbf{x}}^{(1)}\right) \quad \left(i = 1,2,\cdots,n_1\right)$$

を計算する．不良品の群 G_2 のデータ $\mathbf{x}_i^{(2)}\left(i = 1,\cdots,n_2\right)$ に対しても，マハラノビス平方距離

$$\left\{D_i^{(2)}\right\}^2 = \left(\mathbf{x}_i^{(2)} - \bar{\mathbf{x}}^{(1)}\right)^T \hat{\Sigma}_1^{-1}\left(\mathbf{x}_i^{(2)} - \bar{\mathbf{x}}^{(1)}\right) \ (i = 1,\cdots,n_2)$$

を求める．$\left\{D_i^{(1)}\right\}^2 \left(i = 1,2,\cdots,n_1\right)$ と $\left\{D_i^{(2)}\right\}^2 \ (i = 1,\cdots,n_2)$ のボックス・プロットを図１に与える．

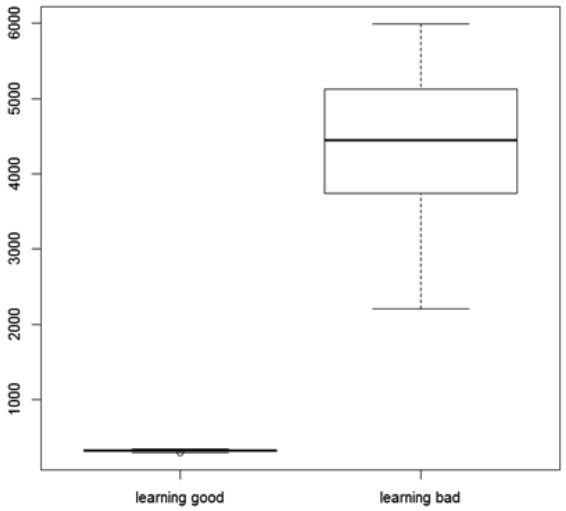

図１．マハラノビス平方距離のボックス・プロット

図１を見るとMTシステムが完璧に学習データの群G_1とG_2を分離することが判る．

検証データ$\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$で予測をするに当たって，学習データの不良品群$G_2$の分析から，群$G_2$のマハラノビス平方距離$\left\{D_i^{(2)}\right\}^2$ $(i=1,\cdots,n_2)$の最小値$\min_i\left\{D_i^{(2)}\right\}^2$ より小さいマハラノビス平方距離$\left(\mathbf{x}-\overline{\mathbf{x}}^{(1)}\right)^T\hat{\Sigma}_1^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}^{(1)}\right)$のものは，学習データの良品群$G_1$に属させればよいと考えられる．すなわち，

$$\min_i\left\{D_i^{(2)}\right\}^2-\left(\mathbf{x}-\overline{\mathbf{x}}^{(1)}\right)^T\hat{\Sigma}_1^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}^{(1)}\right)\begin{cases}>0\Rightarrow\mathbf{x}\in G_1\\ \le 0\Rightarrow\mathbf{x}\notin G_1\end{cases}$$

を判別方式として採用する．表２に判別結果をまとめる．

表２．MTシステムの予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	306	0	306
良品	0	344	344
誤判別率＝0/650=0			

予測精度

実測	予測		
	不良品	良品	合計
不良品	113	1	114
良品	132	4	136
誤判別率＝(1+132)/250=0.532			

表２の結果から，MTシステムは学習データに対する判別効率は良好であるが，予測精度については，特に，良品を不良とはねてしまうので少し問題があるのかもしれない．

３．ロジスティック判別

ここでは，我々の調査のスペクトルデータ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ と良不良の判定を表す y 変数

$$y=\begin{cases}1\Leftrightarrow & \text{良品}\\ 0\Leftrightarrow & \text{不良品}\end{cases}$$

を用いて，次の確率

$\Pr\left(y=1|\mathbf{x}\right)$ ： $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときに良品となる確率

$\Pr\left(y=0|\mathbf{x}\right)$ ： $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときに不良品となる確率

$$\Pr\left(y=1|\mathbf{x}\right)+\Pr\left(y=0|\mathbf{x}\right)=1$$

を求めることを考える．

ここで，

$$\Pr\left(y=1|\mathbf{x}\right)=\pi\left(\mathbf{x}\right),\Pr\left(y=0|\mathbf{x}\right)=1-\pi\left(\mathbf{x}\right)$$

と置く．

$\pi\left(\mathbf{x}\right)$ が判れば，データ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときの良不良の判定は

$$\frac{\Pr\left(y=1|\mathbf{x}\right)}{\Pr\left(y=0|\mathbf{x}\right)}=\frac{\pi\left(\mathbf{x}\right)}{1-\pi\left(\mathbf{x}\right)}\begin{cases}\geq 1\Rightarrow & \text{良品}\\ <1\Rightarrow & \text{不良品}\end{cases}$$

で行える．ここで，次の仮定

$$\log\left\{\frac{\pi\left(\mathbf{x}\right)}{1-\pi\left(\mathbf{x}\right)}\right\}=c+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$

を置くと，$\pi\left(\mathbf{x}\right)$ は

$$\begin{aligned}\pi\left(\mathbf{x}\right)&=\frac{\exp(c+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}{1+\exp(c+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}\\&=\frac{\exp\left(c+\boldsymbol{\beta}^T\mathbf{x}\right)}{1+\exp\left(c+\boldsymbol{\beta}^T\mathbf{x}\right)}\end{aligned}$$

と表せる．この関数形を持つ $\pi\left(\mathbf{x}\right)$ をロジスティック・モデルと呼ぶ．

ロジスティック・モデルを利用すれば，データ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときの良不良の判定は

$$\log\left\{\frac{\Pr\left(y=1|\mathbf{x}\right)}{\Pr\left(y=0|\mathbf{x}\right)}\right\}=\log\left\{\frac{\pi\left(\mathbf{x}\right)}{1-\pi\left(\mathbf{x}\right)}\right\}=c+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p\begin{cases}\geq 0\Rightarrow & \text{良品}\\ <0\Rightarrow & \text{不良品}\end{cases}$$

で行える．この左辺にある「確率の比」の対数をとった量は，対数オッズと呼ばれる．

$\pi(\mathbf{x})$の$\boldsymbol{\beta}$とcの推定は以下のように行う．データ$\left\{\mathbf{x}_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)^T,y_i\right\}(i=1,2,\cdots,n)$が与えられたとき，データの独立性を仮定すると，$y_i\ (i=1,\cdots,n)$の同時確率密度は

$$L(\boldsymbol{\beta},c)=\prod_{i=1}^{n}\pi(\mathbf{x}_i)^{y_i}\left(1-\pi(\mathbf{x}_i)\right)^{1-y_i}$$

となるので，$L(\boldsymbol{\beta},c)$を$\boldsymbol{\beta}$とcについて最大化する．そこで，$L(\boldsymbol{\beta},c)$の対数を求めて

$$\begin{aligned}\log L(\boldsymbol{\beta},c)&=\sum_{i=1}^{n}\left[y_i\log\pi(\mathbf{x}_i)+(1-y_i)\log\left\{1-\pi(\mathbf{x}_i)\right\}\right]\\&=\sum_{i=1}^{n}\left[y_ic+y_i\boldsymbol{\beta}^T\mathbf{x}_i-\log\left\{1+\exp\left(c+\boldsymbol{\beta}^T\mathbf{x}_i\right)\right\}\right]\end{aligned}$$

を$\boldsymbol{\beta}$とcについて最大化する．$\log L(\boldsymbol{\beta},c)$の最大値を与える$\boldsymbol{\beta}$と$c$を$\hat{\boldsymbol{\beta}}$および$\hat{c}$とする．

新しくデータ$\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$が与えられたとき，良品$(y=1)$が現れるか，不良品$(y=0)$が現れるかの判定は，推定した$\hat{\boldsymbol{\beta}}$と$\hat{c}$を用いて，

$$\log\left\{\frac{\hat{\pi}(\mathbf{x})}{1-\hat{\pi}(\mathbf{x})}\right\}=\hat{c}+\hat{\boldsymbol{\beta}}^T\mathbf{x}=\hat{c}+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p\begin{cases}\geq 0\Rightarrow y=1 & (\text{良品})\\<0\Rightarrow y=0 & (\text{不良品})\end{cases}$$

または，同じことであるが，確率$\pi(\mathbf{x})$の推定値を用いて

$$\hat{\pi}(\mathbf{x})=\frac{\exp\left(\hat{c}+\hat{\boldsymbol{\beta}}^T\mathbf{x}\right)}{1+\exp\left(\hat{c}+\hat{\boldsymbol{\beta}}^T\mathbf{x}\right)}=\frac{\exp(\hat{c}+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p)}{1+\exp(\hat{c}+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p)}\begin{cases}\geq\dfrac{1}{2}\Rightarrow y=1 & (\text{良品})\\<\dfrac{1}{2}\Rightarrow y=0 & (\text{不良品})\end{cases}$$

という判別方式で行う．この判別法をロジスティック判別と呼ぶ．

例３．ロジスティック判別

我々の調査の学習データを用いてロジスティック・モデルの$\boldsymbol{\beta}$とcを推定する．得られた推定値$\hat{\boldsymbol{\beta}}$と$\hat{c}$を利用して，学習データの学習精度（誤判別率）と，検証データの予測精度（誤判別率）を調べた結果が表３である．

表３．ロジスティック判別の予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	306	0	306
良品	0	344	344
誤判別率＝0/650=0			

予測精度

実測	予測		
	不良品	良品	合計
不良品	83	31	114
良品	44	92	136
誤判別率＝(31+44)/250=0.30			

表３からロジスティック判別の学習精度は良好である．予測精度については，表１の線形判別とあまり差がないように見える．

４．ニューラルネット

ここでは，３節でのロジスティック・モデルを拡張して，階層化されたロジスティック・モデルを考える．

ロジスティック関数（あるいは，ロジスティック変換）を

$$\text{logistic(a)}=\frac{\exp(a)}{1+\exp(a)}$$

と表せば，３節のロジスティック・モデル $\pi(\mathbf{x})$ は

$$\pi(\mathbf{x})=\text{logistic}(a),\ a=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$

となっている．特徴的なのは，任意の値をとる線形関数 $a=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$ に対して 0 から 1 の間の値をとるように変換が行われていることである．

ここで，任意の線形関数

$$b=\theta_0+\theta_1z_1+\theta_2z_2+\cdots+\theta_Mz_M$$

に対するロジスティック変換 $\text{logistic}(b)$ を考え，ロジスティック変換の階層化を次のように定義する．

$$\pi(\mathbf{x})=\text{logistic}(b),\ \ b=\theta_0+\theta_1z_1+\theta_2z_2+\cdots+\theta_Mz_M$$
$$\begin{cases} z_1=\text{logistic}(a_1),\ a_1=\beta_{10}+\beta_{11}x_1+\beta_{12}x_2+\cdots+\beta_{1p}x_p \\ z_2=\text{logistic}(a_2),\ a_2=\beta_{20}+\beta_{21}x_1+\beta_{22}x_2+\cdots+\beta_{2p}x_p \\ \qquad\qquad\qquad\qquad\qquad\vdots \\ z_M=\text{logistic}(a_M),\ a_M=\beta_{M0}+\beta_{M1}x_1+\beta_{M2}x_2+\cdots+\beta_{Mp}x_p \end{cases}$$

ここでの未知母数は $\{\theta_i\}\ (i=0,1,\cdots,M)$ および $\{\beta_{ij}\}\ (i=1,2,\cdots,M;\ j=0,1,2,\cdots,p)$ である．

この階層化されたロジスティック・モデルは，階層的ニューラルネット・モデルの１例である．線形関数 $b=\theta_0+\theta_1z_1+\theta_2z_2+\cdots+\theta_Mz_M$ 中の集合 $\{z_i\}\ (i=1,2,\cdots,M)$ は隠れ層（あるいは中間層）と呼ばれ，M は隠れ層のユニット数と呼ばれる．従って，この階層的ニューラルネット・モデルは，入力データ $\mathbf{x}=(x_1,x_2,\cdots,x_p)^T$ に関する p 個のユニット $\{x_i\}\ (i=1,2,\cdots,p)$ からなる入力層，M 個のユニット $\{z_i\}\ (i=1,2,\cdots,M)$ からなる一つの隠れ層，一つのユニット $\pi(\mathbf{x})$ を持つ出力層からできている．

未知母数 $\{\theta_i\}$ ，$\{\beta_{ij}\}$ の推定は，データ $\left\{\mathbf{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^T,y_i\right\}\ (i=1,2,\cdots,n)$ の同時確率密度

$$L\left(\{\theta_i\},\{\beta_{ij}\}\right)=\prod_{i=1}^{n}\pi\left(\mathbf{x}_i\right)^{y_i}\left(1-\pi\left(\mathbf{x}_i\right)\right)^{1-y_i}$$

の対数を用いて

$$\log L\left(\{\theta_i\},\{\beta_{ij}\}\right)=\sum_{i=1}^{n}\left[y_i\log\pi\left(\mathbf{x}_i\right)+\left(1-y_i\right)\log\left(1-\pi\left(\mathbf{x}_i\right)\right)\right]$$

を，未知母数 $\{\theta_i\}$ ， $\{\beta_{ij}\}$ に関して最大化し，推定値 $\{\hat{\theta}_i\}$ ， $\{\hat{\beta}_{ij}\}$ を得る．

新しい入力データ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたとき，良品 $\left(y=1\right)$ が現れるか，不良品 $\left(y=0\right)$ が現れるかの判定は，推定した $\{\hat{\theta}_i\}$ と $\{\hat{\beta}_{ij}\}$ を用いて，

$$\hat{\pi}\left(\mathbf{x}\right)=\text{logistic}\left(\hat{b}\right),\hat{b}=\hat{\theta}_0+\hat{\theta}_1\hat{z}_1+\hat{\theta}_2\hat{z}_2+\cdots+\hat{\theta}_M\hat{z}_M$$

$$\begin{cases}\hat{z}_1=\text{logistic}\left(\hat{a}_1\right),\ \hat{a}_1=\hat{\beta}_{10}+\hat{\beta}_{11}x_1+\hat{\beta}_{12}x_2+\cdots+\hat{\beta}_{1p}x_p\\ \hat{z}_2=\text{logistic}\left(\hat{a}_2\right),\ \hat{a}_2=\hat{\beta}_{20}+\hat{\beta}_{21}x_1+\hat{\beta}_{22}x_2+\cdots+\hat{\beta}_{2p}x_p\\ \qquad\qquad\vdots\\ \hat{z}_M=\text{logistic}\left(\hat{a}_M\right),\ \hat{a}_M=\hat{\beta}_{M0}+\hat{\beta}_{M1}x_1+\hat{\beta}_{M2}x_2+\cdots+\hat{\beta}_{Mp}x_p\end{cases}$$

$$\hat{\pi}\left(\mathbf{x}\right)=\text{logistic}\left(\hat{b}\right)\begin{cases}\geq\dfrac{1}{2}\Rightarrow y=1\quad\left(\text{良品}\right)\\ <\dfrac{1}{2}\Rightarrow y=0\left(\text{不良品}\right)\end{cases}$$

という判別方式で行う．

例４．階層的ニューラルネット

ここで用いる階層的ニューラルネット・モデルは，隠れ層（中間層）のユニット数を $M=3$ としている．調査の学習データを用いて階層的ニューラルネット・モデルの未知母数 $\{\theta_i\}$ ， $\{\beta_{ij}\}$ を推定する．未知母数の個数は 900 個以上で，我々の全データ数よりも多い．未知母 $\{\theta_i\}$ ， $\{\beta_{ij}\}$ の推定のための初期値は，乱数を用いて与える．得られた推定値 $\{\hat{\theta}_i\}$ と $\{\hat{\beta}_{ij}\}$ を利用して，学習データの学習精度（誤判別率）と，検証データの予測精度（誤判別率）を調べた結果が表 4 である．表 4 では二つの局所解を与えている．

表４．階層的ニューラルネットの予測精度

(a)局所解1

学習精度

実測	予測		
	不良品	良品	合計
不良品	306	0	306
良品	0	344	344
誤判別率＝0			

予測精度

実測	予測		
	不良品	良品	合計
不良品	86	28	114
良品	19	117	136
誤判別率＝0.188			

(b)局所解2

学習精度

実測	予測		
	不良品	良品	合計
不良品	303	3	306
良品	15	329	344
誤判別率＝18/650=0.027692			

予測精度

実測	予測		
	不良品	良品	合計
不良品	91	23	114
良品	18	118	136
誤判別率＝41/250=0.164			

表４を見ると，二つの局所解の学習精度は良好である．予測精度については，学習精度の少し劣る局所解２の方がわずかながらも良い．さらに興味深いのは，表３のロジスティック判別の結果や，表１の線形判別の結果よりも，予測精度が良くなっていることである．

５．サポート・ベクター・マシーン

ａ．ハードマージン最適化

図２に，群G_1とG_2のデータ（G_1：正方形，G_2：丸）を打点して示している．我々の関心は，新しいデータ$\mathbf{x}=(x_1,x_2)^T$が与えられたとき，$\mathbf{x}$が群G_1に属するか，群G_2に属するかを判定する直線

$$L(\boldsymbol{\psi},b|\mathbf{x})=\psi_1x_1+\psi_2x_2+b=(\boldsymbol{\psi}*\mathbf{x})+b\begin{cases}>0\Rightarrow\mathbf{x}\in G_1\\<0\Rightarrow\mathbf{x}\in G_2\end{cases}$$

をいかに合理的に決定すれば良いかということである．ここで，二つのｐ次元ベクトル

$\boldsymbol{\psi}=(\psi_1,\psi_2,\cdots,\psi_p)^T$，　$\mathbf{x}=(x_1,x_2,\cdots,x_p)^T$に対する内積に関する記号

$$(\boldsymbol{\psi}*\mathbf{x})=\psi_1x_1+\psi_2x_2+\cdots+\psi_px_p$$

を，$p=2$の場合に用いている．

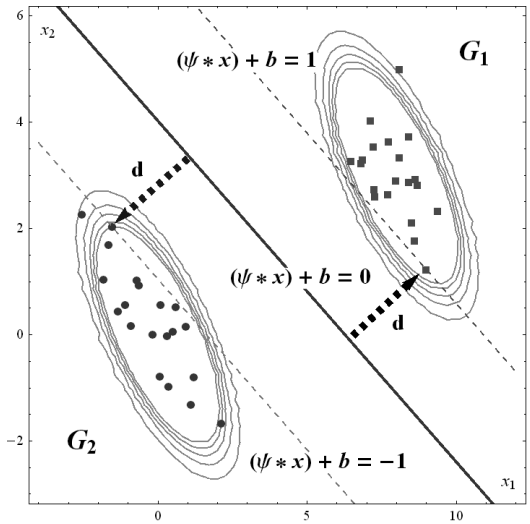

図２．線形分離可能なデータの散布状況

図２では判別のための直線 $L(\boldsymbol{\psi}, b|\mathbf{x}) = (\boldsymbol{\psi} * \mathbf{x}) + b = 0$ を以下のような考え方で導いている．

図２の群 G_1 と G_2 のようなデータの散布状況の時には，いつでも直線を引いて二つの群 G_1 と G_2 を分けることが出来る．このような状況を「線形分離可能」な状況という．そこで，群 G_1 と G_2 を分けるように，平行な直線

$$\begin{cases} \psi_1 x_1 + \psi_2 x_2 + b = 1 & \left((\boldsymbol{\psi} * \mathbf{x}) + b = 1\right) \\ \psi_1 x_1 + \psi_2 x_2 + b = -1 & \left((\boldsymbol{\psi} * \mathbf{x}) + b = -1\right) \end{cases}$$

を描く．二つの直線の方向を決める $\boldsymbol{\psi}$ は図の群 G_1 と G_2 の散布状況から適当に決められるが，b については少し考察が必要である．

群 G_1 の点 $\mathbf{x}_i^{(1)} = \left(x_{i1}^{(1)}, x_{i2}^{(1)}\right)^T \ (i = 1, 2, \cdots, n_1)$ と群 G_2 の点 $\mathbf{x}_j^{(2)} = \left(x_{j1}^{(2)}, x_{j2}^{(2)}\right)^T \ (j = 1, 2, \cdots, n_2)$ は，二つの直線 $(\boldsymbol{\psi} * \mathbf{x}) + b = +1$，$(\boldsymbol{\psi} * \mathbf{x}) + b = -1$ に対して

$$\psi_1 x_{i1}^{(1)} + \psi_2 x_{i2}^{(1)} + b = \left(\boldsymbol{\psi} * \mathbf{x}_i^{(1)}\right) + b \geq +1 \ \ (i = 1, 2, \cdots, n_1)$$
$$\psi_1 x_{j1}^{(2)} + \psi_2 x_{j2}^{(2)} + b = \left(\boldsymbol{\psi} * \mathbf{x}_j^{(2)}\right) + b \leq -1 \ \ (j = 1, 2, \cdots, n_2)$$

を満たしている．二つの直線 $(\boldsymbol{\psi} * \mathbf{x}) + b = \pm 1$ がそれぞれ群 G_1 と群 G_2 のデータ点を通過することから，$\left(\boldsymbol{\psi} * \mathbf{x}_i^{(1)}\right)$ の最小値 $\min_i \left(\boldsymbol{\psi} * \mathbf{x}_i^{(1)}\right)$ と $\left(\boldsymbol{\psi} * \mathbf{x}_j^{(2)}\right)$ の最大値 $\max_j \left(\boldsymbol{\psi} * \mathbf{x}_j^{(2)}\right)$ を用いると

$$\min_i \left(\boldsymbol{\psi} * \mathbf{x}_i^{(1)}\right) + b = 1$$
$$\max_j \left(\boldsymbol{\psi} * \mathbf{x}_j^{(2)}\right) + b = -1$$

が成立している．この二つの式から，

$$b=-\frac{1}{2}\left\{\min_i\left(\boldsymbol{\psi}*\mathbf{x}_i^{(1)}\right)+\max_j\left(\boldsymbol{\psi}*\mathbf{x}_j^{(2)}\right)\right\}$$

であることが判る．この b と方向 $\boldsymbol{\psi}$ を用いて，判別のための直線 $L\left(\boldsymbol{\psi},b\middle|\mathbf{x}\right)=\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=0$ が求められる．

次に，最適な判別のための直線 $L\left(\boldsymbol{\psi},b\middle|\mathbf{x}\right)=\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=0$ を求めるにはどうすればよいか．一つの解決策が Vapnik（1998）[11] により与えられている．それは，二つの直線 $\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=\pm1$ の間の距離が最大になるように方向 $\boldsymbol{\psi}$ と b を決めることである．言い換えれば，直線 $L\left(\boldsymbol{\psi},b\middle|\mathbf{x}\right)=\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=0$ から二つの直線 $\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=\pm1$ までの距離 d （これをマージンと呼ぶ）が最大になるように方向 $\boldsymbol{\psi}$ と b を決めるのである．マージン d は簡単に求められ

$$\begin{aligned}d&=\frac{\min_i\left(\psi_1x_{i1}^{(1)}+\psi_2x_{i2}^{(1)}\right)-\max_j\left(\psi_1x_{j1}^{(2)}+\psi_2x_{j2}^{(2)}\right)}{2\sqrt{\psi_1^{\,2}+\psi_2^{\,2}}}\\&=\frac{\min_i\left(\boldsymbol{\psi}*\mathbf{x}_i^{(1)}\right)-\max_j\left(\boldsymbol{\psi}*\mathbf{x}_j^{(2)}\right)}{2\sqrt{\left(\boldsymbol{\psi}*\boldsymbol{\psi}\right)}}\end{aligned}$$

で与えられる．マージン d は方向 $\boldsymbol{\psi}$ のみの関数であるが，書き換えるとさらに簡単になり，

$$d=\frac{\min_i\left(\boldsymbol{\psi}*\mathbf{x}_i^{(1)}\right)+b-\left\{\max_j\left(\boldsymbol{\psi}*\mathbf{x}_j^{(2)}\right)+b\right\}}{2\sqrt{\left(\boldsymbol{\psi}*\boldsymbol{\psi}\right)}}=\frac{(+1)-(-1)}{2\sqrt{\left(\boldsymbol{\psi}*\boldsymbol{\psi}\right)}}=\frac{1}{\sqrt{\left(\boldsymbol{\psi}*\boldsymbol{\psi}\right)}}$$

となる．すなわち，マージン d の最大化は方向ベクトル $\boldsymbol{\psi}$ の大きさの最小化を意味する．

ここで，データ点 $\mathbf{x}=\left(x_1,x_2\right)^T$ が群 G_1 あるいは群 G_2 に属するという条件は

$$\begin{cases}\psi_1x_1+\psi_2x_2+b\geq+1\Leftrightarrow\mathbf{x}\in G_1\\\psi_1x_1+\psi_2x_2+b\leq-1\Leftrightarrow\mathbf{x}\in G_2\end{cases}$$

となっているので，ラベルを与える変数 y

$$y=\begin{cases}+1\Leftrightarrow\mathbf{x}\in G_1\\-1\Leftrightarrow\mathbf{x}\in G_2\end{cases}$$

を与える．するとデータ $\mathbf{x}=\left(x_1,x_2\right)^T$ および y が与えられたとき，満たすべき条件は一つの式

$$y\left(\psi_1x_1+\psi_2x_2+b\right)\geq1$$

のみで与えられる．

以上のことから，群 G_1 の点 $\mathbf{x}_i^{(1)}=\left(x_{i1}^{(1)},x_{i2}^{(1)}\right)^T\ \left(i=1,2,\cdots,n_1\right)$ と群 G_2 の点 $\mathbf{x}_j^{(2)}=\left(x_{j1}^{(2)},x_{j2}^{(2)}\right)^T$ $\left(j=1,2,\cdots,n_2\right)$ をまとめて，$\mathbf{x}_i\ \left(i=1,2,\cdots,n\right)\left(n=n_1+n_2\right)$，$\mathbf{x}_i$ のラベルを y_i とする．我々のマージン最大化問題は次のようになる．

（ハードマージン最大化）

条件 $y_i\{(\boldsymbol{\psi}*\mathbf{x}_i)+b\}\geq 1\ (i=1,2,\cdots,n)$ の下で，マージン d の最大化を行う．

$\Downarrow$

（ハードマージン最大化の主問題）

条件 $y_i\{(\boldsymbol{\psi}*\mathbf{x}_i)+b\}\geq 1\ (i=1,2,\cdots,n)$ の下で，$\frac{1}{2}(\boldsymbol{\psi}*\boldsymbol{\psi})$ の最小化を行う．

主問題を効率的に解くために，ラグランジュの未定乗数法により，主問題を双対問題に書き換える．ラグランジュ関数は

$$Q(\boldsymbol{\psi},b|\boldsymbol{\alpha})=\frac{1}{2}(\boldsymbol{\psi}*\boldsymbol{\psi})-\sum_{i=1}^{n}\alpha_i\left[y_i\{(\boldsymbol{\psi}*\mathbf{x}_i)+b\}-1\right]\quad(\alpha_i\geq 0,\ i=1,2,\cdots,n)$$

となり，最小値が満たすべき条件は

$$\begin{cases}\dfrac{\partial Q(\boldsymbol{\psi},b,\boldsymbol{\alpha})}{\partial\boldsymbol{\psi}}=\boldsymbol{\psi}-\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i=\mathbf{0}\\ \dfrac{\partial Q(\boldsymbol{\psi},b,\boldsymbol{\alpha})}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0\end{cases}\Rightarrow\begin{cases}\boldsymbol{\psi}=\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i\\ \sum_{i=1}^{n}\alpha_i y_i=0\end{cases}$$

で与えられる．この条件を利用して，ラグランジュ関数 $Q(\boldsymbol{\psi},b|\boldsymbol{\alpha})$ を書き換えると

新しい関数

$$W(\boldsymbol{\alpha})=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i*\mathbf{x}_j)\quad(\alpha_i\geq 0,\ i=1,2,\cdots,n)$$

を得る．双対問題は次のようになる．

（ハードマージン最大化の双対問題）

条件 $\alpha_i\geq 0\ (i=1,2,\cdots,n)$ および $\sum_{i=1}^{n}\alpha_i y_i=0$ の下で，$W(\boldsymbol{\alpha})$ の最大化を行う．

双対問題を解いて得られた解を $\hat{\boldsymbol{\alpha}}$ とする．$\boldsymbol{\psi}$ の推定値はラグランジュ関数の条件から

$$\hat{\boldsymbol{\psi}}=\sum_{i=1}^{n}\hat{\alpha}_i y_i\mathbf{x}_i$$

で与えられる．b の推定値については，二つの直線 $(\hat{\boldsymbol{\psi}}*\mathbf{x})+b=\pm 1$ に載っているデータ点を

$$\begin{cases}(\hat{\boldsymbol{\psi}}*\mathbf{x}^{+})+b=+1\\ (\hat{\boldsymbol{\psi}}*\mathbf{x}^{-})+b=-1\end{cases}$$

とすれば，

$$\hat{b}=-\frac{1}{2}\left\{\left(\hat{\boldsymbol{\psi}}*\mathbf{x}^{+}\right)+\left(\hat{\boldsymbol{\psi}}*\mathbf{x}^{-}\right)\right\}$$

で与えられる．推定値 $\hat{\boldsymbol{\psi}}$ と $\hat{b}$ を用いると，新しいデータ $\mathbf{x}$ に対する判別方式は

$$L\left(\hat{\boldsymbol{\psi}},\hat{b}|\mathbf{x}\right)=\left(\hat{\boldsymbol{\psi}}*\mathbf{x}\right)+\hat{b}=\sum_{i=1}^{n}\hat{\alpha}_i y_i\left(\mathbf{x}_i*\mathbf{x}\right)+\hat{b}\begin{cases}\geq 0\Rightarrow \mathbf{x}\in G_1\\ <0\Rightarrow \mathbf{x}\in G_2\end{cases}$$

で与えられる．

サポート・ベクトルとは，二つの直線 $\left(\hat{\boldsymbol{\psi}}*\mathbf{x}\right)+\hat{b}=\pm 1$ に載っているデータ点 $\mathbf{x}^{+}$，$\mathbf{x}^{-}$ らのことを言う．サポート・ベクトルの個数は，データ数 n に比べて著しく少数であることが判る．すなわち，$\boldsymbol{\psi}$ と b の推定にはごく少数のサポート・ベクトルしか必要としない．この点がサポート・ベクター・マシーンの大きな利点であり特徴である．

b．ソフトマージン最適化

さて，図３に線形分離可能でないデータの状況を与えている．

図３では，群 G_1 の点で，直線 $\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=+1$ より下にあるものが２個 $\left\{\mathbf{x}_i,\mathbf{x}_j\right\}$ あり，群 G_2 の点で，直線 $\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=-1$ より上にあるものが１個 $\left\{\mathbf{x}_k\right\}$ ある．３点 $\left\{\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k\right\}$ は逸脱度を示す変数（スラック変数と呼ぶ）ξ_i，ξ_j，ξ_k (≥ 0) を使って，直線

$$\left(\boldsymbol{\psi}*\mathbf{x}_i\right)+b=1-\xi_i,\ \left(\boldsymbol{\psi}*\mathbf{x}_j\right)+b=1-\xi_j,\left(\boldsymbol{\psi}*\mathbf{x}_k\right)+b=-1+\xi_k$$

の上に載っているものとして取り扱われている．ここで，直線 $\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=0$ によって正しく判別されるのが点 $\mathbf{x}_i$ $(0<\xi_i<1)$ であり，誤判別されるのが点 $\left\{\mathbf{x}_j,\mathbf{x}_k\right\}$ $(\xi_j,\xi_k>1)$ である．二つの直線 $\left(\boldsymbol{\psi}*\mathbf{x}\right)+b=\pm 1$ によって，正しく分離されているデータ点に対しては，逸脱度 $\xi=0$ と考え

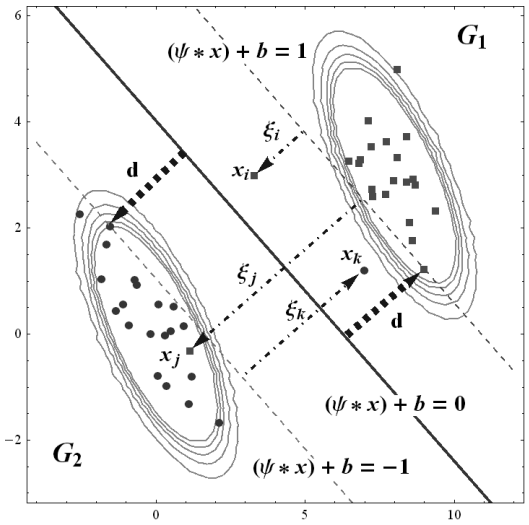

図３．線形分離でないデータの散布状況

る．というわけで，全てのデータ点 $\mathbf{x}_i$ について逸脱度 $\xi_i\ (\geq 0)$ を考えて，逸脱度の和 $\sum_{i=1}^{n}\xi_i$ が出来るだけ小さくなるようにマージン d の最大化を考える．このときのマージンをソフトマージンと呼ぶ．ソフトマージンの最適化は次のように言い表せる．

（ソフトマージンの最適化の主問題）

関数

$$\Phi(\boldsymbol{\psi},\boldsymbol{\xi}) = \frac{1}{2}(\boldsymbol{\psi}*\boldsymbol{\psi}) + C\left(\sum_{i=1}^{n}\xi_i\right)$$

を条件 $y_i\left\{(\boldsymbol{\psi}*\mathbf{x}_i)+b\right\} \geq 1-\xi_i,\ \xi_i \geq 0\ \ (i=1,2,\cdots,n)$ の下で最小化する．このときの $C\ (>0)$ は適当に与えた定数で，ペナルティを調節する役目を持つ．

この主問題に対する双対問題を求めるために，ラグランジュ関数

$$Q(\boldsymbol{\psi},b,\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{\psi}*\boldsymbol{\psi}) + C\left(\sum_{i=1}^{n}\xi_i\right) - \sum_{i=1}^{n}\alpha_i\left[y_i\left\{(\boldsymbol{\psi}*\mathbf{x}_i)+b\right\}-1+\xi_i\right] - \sum_{i=1}^{n}\beta_i\xi_i$$
$$(\alpha_i \geq 0, \beta_i \geq 0,\ \ i=1,2,\cdots,n)$$

を考える．最小値が満たす条件は

$$\begin{cases} \dfrac{\partial Q(\boldsymbol{\psi},b,\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\psi}} = \boldsymbol{\psi} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i = \mathbf{0} \\ \dfrac{\partial Q(\boldsymbol{\psi},b,\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \\ \dfrac{\partial Q(\boldsymbol{\psi},b,\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \end{cases} \Rightarrow \begin{cases} \boldsymbol{\psi} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \alpha_i + \beta_i = C\ \ (i=1,\cdots,n) \end{cases}$$

となる．この条件を用いて，ラグランジュ関数 $Q(\boldsymbol{\psi},b,\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta})$ を書き直せば，再び

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (\mathbf{x}_i * \mathbf{x}_j)\ \ (C \geq \alpha_i \geq 0,\ i=1,2,\cdots,n)$$

を得る．

（ソフトマージン最適化の双対問題）

条件 $C \geq \alpha_i \geq 0\ \ (i=1,2,\cdots,n)$ および $\sum_{i=1}^{n}\alpha_i y_i = 0$ の下で，$W(\boldsymbol{\alpha})$ の最大化を行う．

双対問題を解いて得られた解を $\hat{\boldsymbol{\alpha}}$ とする．$\boldsymbol{\psi}$ の推定値はラグランジュ関数の条件から

$$\hat{\boldsymbol{\psi}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i$$

で与えられる．b の推定値については，二つの直線 $\left(\hat{\boldsymbol{\psi}} * \mathbf{x}\right) + b = \pm 1$ に載っているデータ点を

$$\begin{cases} \left(\hat{\boldsymbol{\psi}} * \mathbf{x}^{+}\right) + b = +1 \\ \left(\hat{\boldsymbol{\psi}} * \mathbf{x}^{-}\right) + b = -1 \end{cases}$$

とすれば，

$$\hat{b} = -\frac{1}{2}\left\{\left(\hat{\boldsymbol{\psi}} * \mathbf{x}^{+}\right) + \left(\hat{\boldsymbol{\psi}} * \mathbf{x}^{-}\right)\right\}$$

で与えられる．推定値 $\hat{\boldsymbol{\psi}}$ と $\hat{b}$ を用いると，新しいデータ $\mathbf{x}$ に対する判別方式は

$$L\left(\hat{\boldsymbol{\psi}}, \hat{b} \middle| \mathbf{x}\right) = \left(\hat{\boldsymbol{\psi}} * \mathbf{x}\right) + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \left(\mathbf{x}_i * \mathbf{x}\right) + \hat{b} \begin{cases} \geq 0 \Rightarrow \mathbf{x} \in G_1 \\ < 0 \Rightarrow \mathbf{x} \in G_2 \end{cases}$$

で与えられる．

ソフトマージン最適化の場合には，サポート・ベクトルは，二つの直線 $\left(\hat{\boldsymbol{\psi}} * \mathbf{x}\right) + \hat{b} = \pm 1$ に載っているデータ点 $\mathbf{x}^{+}$，$\mathbf{x}^{-}$ らに加えて，逸脱度 $\hat{\xi}_i$ が正になるデータ点 $\mathbf{x}_i$ を追加した集合になる．サポート・ベクター・マシーンの詳細については，Vapnik（1998）[11] 参照．

例５．サポート・ベクター・マシーン

我々の調査の学習データにサポート・ベクター・マシーンを適用して，ソフトマージン最適化により，判別直線 $L\left(\boldsymbol{\psi}, b \middle| \mathbf{x}\right) = \left(\boldsymbol{\psi} * \mathbf{x}\right) + b = 0$ の $\boldsymbol{\psi}$ と b を推定する．得られた推定値 $\hat{\boldsymbol{\psi}}$，$\hat{b}$ を利用して，学習データの学習精度と，検証データの予測精度を調べた結果が表５である．

表５．サポート・ベクター・マシーンの予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	295	11	306
良品	9	335	344
誤判別率＝(11+9)/650=0.0307			

予測精度

実測	予測		
	不良品	良品	合計
不良品	89	25	114
良品	21	115	136
誤判別率＝(25+21)/250=0.184			

表５の結果を見ると，サポート・ベクター・マシーンの予測精度は，表４のニューラルネットの予測精度と大差ないことが判る．興味深く感じるのは，サポート・ベクター・マシーンの性質が「線形判別」であるのに対し，ニューラルネットは「非線形判別」であることである．サポート・ベクター・マシーンの持つ本質的な線形性により，現在では，カーネル関数などが導入され，

その適用可能性が大きく拡大されている．

６．ｋ近傍法

パターン認識での密度推定の方法であるｋ近傍法を，我々の分類の問題に適用する．群 G_1（たとえば，良品群）と G_2（不良品群）のデータが与えられているとき，新しいデータ $\mathbf{x}$ がどちらの群に属するかを，以下のような考え方で判定する．

３節のロジスティック判別の時と同様に，新しいデータ $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ と分類の判定を表す y 変数

$$y=\begin{cases}1 \Leftrightarrow \mathbf{x}\in G_1\\ 0 \Leftrightarrow \mathbf{x}\in G_2\end{cases}$$

を用いて，次の確率

$\Pr\left(y=1|\mathbf{x}\right)$： $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときに $\mathbf{x}\in G_1$ となる確率

$\Pr\left(y=0|\mathbf{x}\right)$： $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ が与えられたときに $\mathbf{x}\in G_2$ となる確率

$$\Pr\left(y=1|\mathbf{x}\right)+\Pr\left(y=0|\mathbf{x}\right)=1$$

を求めることを考える．ここで，$\pi\left(\mathbf{x}\right)=\Pr\left(y=1|\mathbf{x}\right)$ と置くと，$\Pr\left(y=0|\mathbf{x}\right)=1-\pi\left(\mathbf{x}\right)$ である．

まず，予め決めてある整数値 k (>0, 奇数とする)の個数だけ群 G_1 と G_2 のデータを含むように，$\mathbf{x}$ を中心に円または球体を描く．k 個のデータのうち，群 G_1 に属しているデータの個数を $k_1\left(\geq 0\right)$，群 G_2 に属しているデータの個数を $k_2\left(\geq 0\right)$ とする．このとき，$\pi\left(\mathbf{x}\right)$ の推定値が

$$\widehat{\pi\left(\mathbf{x}\right)}=\frac{k_1}{k} \quad \text{および} \quad \widehat{1-\pi\left(\mathbf{x}\right)}=\frac{k_2}{k} \quad \left(k_1+k_2=k\right)$$

として求められる．したがって，新しいデータ $\mathbf{x}$ がどちらの群に属するかの判定は

$$\frac{\widehat{\Pr\left(y=1|\mathbf{x}\right)}}{\widehat{\Pr\left(y=0|\mathbf{x}\right)}}=\frac{\widehat{\pi\left(\mathbf{x}\right)}}{\widehat{1-\pi\left(\mathbf{x}\right)}}=\frac{k_1}{k_2}\begin{cases}>1\Rightarrow \mathbf{x}\in G_1\\ <1\Rightarrow \mathbf{x}\in G_2\end{cases} \quad \text{あるいは，簡単に} \quad k_1-k_2\begin{cases}>0\Rightarrow \mathbf{x}\in G_1\\ <0\Rightarrow \mathbf{x}\in G_2\end{cases}$$

で行う．

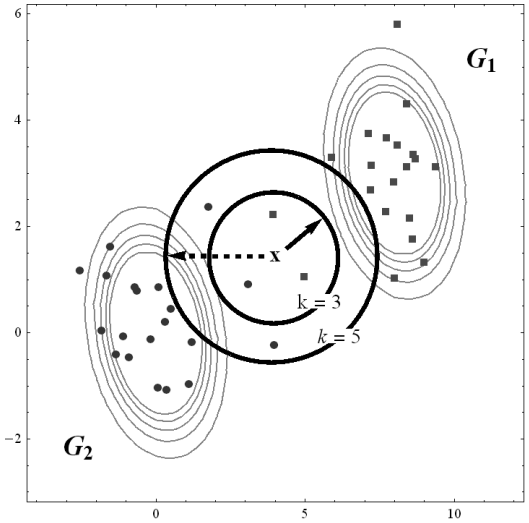

図４．２次元データでのｋ近傍法

図４に２次元の場合の例を示す．群 G_1 のデータが正方形，群 G_2 のデータが丸で与えられている．新しいデータ **x** に対して，$k=3$ のデータ点を含むように，**x** を中心にして円を描くと，$k_1=2$，$k_2=1$ となり，**x** は群 G_1 に属すると判定される．$k=5$ のデータ点を含むように円を描くと，$k_1=2$，$k_2=3$ となり，**x** は群 G_2 に属すると判定される．

ｋ近傍法は２点 $\mathbf{a}=\left(a_1,a_2,\cdots,a_p\right)^T$ と $\mathbf{b}=\left(b_1,b_2,\cdots,b_p\right)^T$ 間の距離としてユークリッド距離

$$d\left(\mathbf{a},\mathbf{b}\right)=\sqrt{\left(\mathbf{a}-\mathbf{b}\right)*\left(\mathbf{a}-\mathbf{b}\right)}=\sqrt{\left(a_1-b_1\right)^2+\left(a_2-b_2\right)^2+\cdots+\left(a_p-b_p\right)^2}$$

を用いることが多い．図４での半径を描いた距離はユークリッド距離で p=2 の場合である．

例６．ｋ近傍法

我々の調査データにｋ近傍法を適用した．学習データの良品群を群 G_1，不良品群を群 G_2 として，検証データを用いて予測精度（誤判別率）を調べてみた．ここでは $k=3$ としている．表６に学習精度と予測精度の結果を示す．

表６．ｋ近傍法の予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	272	34	306
良品	30	314	344
誤判別率＝(34+30)/650=0.0984			

予測精度

実測	予測		
	不良品	良品	合計
不良品	86	28	114
良品	25	111	136
誤判別率＝(28+25)/250=0.212			

表６の結果は，ニューラルネットの表４の結果，およびサポート・ベクター・マシーンの表５の結果とあまり大差がないことが判る．また，表６の結果は線形判別とロジスティック判別の表１と表２の結果よりも良好である．意外にも，予測精度はそれほど悪くない．

ｋ近傍法は本質的に非線形手法で，もし質のよい学習データが十分に集まるならば，分類問題に対して強力な道具になり得る．ただし，そのためには大記憶容量を持ち，超高速で計算処理が行える計算機が必要になる．

７．バギング

二つの群 G_1 と G_2 が存在するとき，与えられたデータ $\mathbf{x}$ に対して，$\mathbf{x}$ がどちらの群に属するかを判定する問題に対して，今までの考え方の手順は次のようになる．まず，$\mathbf{x}$ に対する判定のための予測式あるいは判定方式 $\varphi(\mathbf{x},\boldsymbol{\theta})$ を構成する．つぎに群 G_1 と G_2 の学習データ D を用いて，$\boldsymbol{\theta}$ の推定値 $\hat{\boldsymbol{\theta}}$ を求める．最後に，検証データで $\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}\right)$ の予測精度を評価する．

ここで，群 G_1 と G_2 の M 個の独立な学習データ $D_{(i)}$ $\left(i=1,2,\cdots,M\right)$ が利用出来ると仮定しよう．M 個の学習データ $\left\{D_{(i)}\right\}$ から，M 個の予測式 $\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}\right)$ $\left(i=1,2,\cdots,M\right)$ が得られる．すると，新しい予測式として

$$\overline{\varphi}(\mathbf{x})=\frac{1}{M}\sum_{i=1}^{M}\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}\right)$$

または

$$\varphi_{\max}(\mathbf{x})=\ \{\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}\right)\ \left(i=1,2,\cdots,M\right)\text{により}\mathbf{x}\text{が属すると判定された個数の最も多い群}\}$$

の利用が考えられる．

通常は，群 G_1 と G_2 の M 個の独立な学習データ $D_{(i)}$ $\left(i=1,2,\cdots,M\right)$ は利用できないため，学習データ D から，ブーツストラップ（bootstrap）と呼ばれるリサンプリング（標本再抽出）法を用いて，M 個のブーツストラップ標本を取り出し，学習データ $D^{*}_{(i)}$ $\left(i=1,2,\cdots,M\right)$ を作る．

ブーツストラップ法を簡単に紹介する（詳細は，Efron & Tibshirani（1993）[4] 参照）．群 G_1 と G_2 の学習データを $D=\left\{\left(\mathbf{x}_i,y_i\right),i=1,2,\cdots,n\right\}$ とする．y_i はデータの属する群を示すラベル変数である．ここで，$D=\left\{\left(\mathbf{x}_i,y_i\right),i=1,2,\cdots,n\right\}$ を用いた離散分布

$$\Pr\left(\left(\mathbf{X},y\right)=\left(\mathbf{x}_i,y_i\right)\right)=\Pr\left(\mathbf{X}=\mathbf{x}_i,y=y_i\right)=\frac{1}{n}\quad\left(i=1,2,\cdots,n\right)$$

を考えて，$\left(\mathbf{X},y\right)$ についての大きさ n の独立標本 $D^{*}=\left\{\left(\mathbf{x}_i^{*},y_i^{*}\right),i=1,2,\cdots,n\right\}$ を抽出する．標本 D^{*} は標本 $D=\left\{\left(\mathbf{x}_i,y_i\right),i=1,2,\cdots,n\right\}$ からのブーツストラップ標本と呼ばれる．

学習データ $D=\{(\mathbf{x}_i, y_i), i=1,2,\cdots,n\}$ からの M 個のブーツストラップ標本 $D^*_{(i)}$ $(i=1,2,\cdots,M)$ を用いて，M 個の予測式 $\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}{}^*\right)$ $(i=1,2,\cdots,M)$ を求める．新しい予測式として

$$\bar{\varphi}^*(\mathbf{x})=\frac{1}{M}\sum_{i=1}^{M}\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}{}^*\right)$$

または

$\varphi_{\max}{}^*(\mathbf{x})=$ $\{\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}{}^*\right)$ $(i=1,2,\cdots,M)$ により **x** が属すると判定された個数の最も多い群$\}$

を利用する．

上の方式は，ブーツストラップ（bootstrap）を用いて予測式を統合する（aggregating）ことから，バギング（bagging）と呼ばれている．統合した予測式 $\bar{\varphi}^*(\mathbf{x})$ の平均２乗誤差は，個々の予測式 $\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}{}^*\right)$ の平均２乗誤差を M 個足し合わせて M で平均したものより小さくなることが知られている．詳細は，Breiman（1996）[2] 参照．

例７．バギング

バギングを我々の調査の学習データに適用して予測式を推定した．得られた予測式を用いて，学習データの学習精度（誤判別率）と，検証データの予測精度（誤判別率）を調べた結果が表７である．ブーツストラップ標本の個数は $M=40$ としている．

表７．バギングの予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	306	0	306
良品	0	344	344
誤判別率=0/650=0			

予測精度

実測	予測		
	不良品	良品	合計
不良品	114	0	114
良品	0	136	136
誤判別率=0/250=0			

表７の結果を見ると，今までの手法の中で最良の結果になっている．

８．ブースティング

前節のバギングでは，データ **x** に対する予測式 $\varphi(\mathbf{x},\boldsymbol{\theta})$ を推定するのに，M 個のブーツストラップ標本を用いて，M 個の予測式 $\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(i)}\right)$ (i=1,…,M)を統合することで，新しい予測式を構成した．ここでは，学習データを繰り返し利用することで，段階的により良い予測式を構成し，最終的に

は得られた予測式を全て利用して，それらの一次結合で統合した予測式を求める適応型ブースティング（adaptive boosting，略して AdaBoost）を紹介する．

学習データを $D=\{(\mathbf{x}_i,y_i),i=1,2,\cdots,n\}$ とする．ラベル変数 y は，

$$y_i=\begin{cases}+1\Rightarrow \mathbf{x}_i\in G_1\\ -1\Rightarrow \mathbf{x}_i\in G_2\end{cases}$$

とする．データ $\mathbf{x}$ に対する予測式 $\varphi(\mathbf{x},\boldsymbol{\theta})$ は， y の値を予測する，すなわち， $y(\mathbf{x})=\varphi(\mathbf{x},\boldsymbol{\theta})$（ $y(\mathbf{x})=+1$ または $y(\mathbf{x})=-1$ ）と仮定する． $\varphi(\mathbf{x},\boldsymbol{\theta})$ はベース学習器（分類器）とも呼ばれる．

次の誤差関数

$$E=\sum_{i=1}^{n}\exp\left\{-y_i\varphi^{(m)}(\mathbf{x}_i)\right\}$$

を考える．ここで， $\varphi^{(m)}(\mathbf{x})$ は

$$\varphi^{(m)}(\mathbf{x})=\frac{1}{2}\sum_{j=1}^{m}\alpha_{(j)}\varphi\left(\mathbf{x},\boldsymbol{\theta}_{(j)}\right)$$

で定義される予測式である．目標は，誤差関数 E を最小にするような $\left\{\alpha_{(j)},\boldsymbol{\theta}_{(j)}\right\}(j=1,2,\cdots,m)$ を見つけることである．

誤差関数 E の $\left\{\alpha_{(j)},\boldsymbol{\theta}_{(j)}\right\}(j=1,2,\cdots,m)$ に関する最小化を行うのに，同時最適化を行わずに，逐次最適化を行うことを考える．そこで， $\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(1)}\right),\cdots,\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(m-1)}\right)$ と $\hat{\alpha}_{(1)},\cdots,\hat{\alpha}_{(m-1)}$ が与えられたとして， $\alpha_{(m)}$ と $\boldsymbol{\theta}_{(m)}$ について最小化を行う．誤差関数を書き直して

$$\begin{aligned}E&=\sum_{i=1}^{n}\exp\left\{-y_i\hat{\varphi}^{(m-1)}(\mathbf{x}_i)-\frac{1}{2}y_i\alpha_{(m)}\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\right\}\\&=\sum_{i=1}^{n}c_i^{(m)}\exp\left\{-\frac{1}{2}y_i\alpha_{(m)}\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\right\}\end{aligned}$$

を得る． ただし， $c_i^{(m)}=\exp\left\{-y_i\hat{\varphi}^{(m-1)}(\mathbf{x}_i)\right\}$， $\hat{\varphi}^{(m-1)}(\mathbf{x}_i)=\frac{1}{2}\sum_{j=1}^{m-1}\hat{\alpha}_{(j)}\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(j)}\right)$ $(i=1,2,\cdots,n)$ と置いた．ここで，重み $\left\{c_i^{(m)}\right\}$ は定数である．

$\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)$ と y_i に関する標示関数

$$I\left\{\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\neq y_i\right\}=\begin{cases}+1\Leftrightarrow \varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\neq y_i\\ 0\Leftrightarrow \varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)= y_i\end{cases}$$

を導入すると，誤差関数 E はさらに簡単になって

$$E=\left(\exp\left(\frac{\alpha_{(m)}}{2}\right)-\exp\left(-\frac{\alpha_{(m)}}{2}\right)\right)\sum_{i=1}^{n}c_i^{(m)}I\left\{\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\neq y_i\right\}+\exp\left(-\frac{\alpha_{(m)}}{2}\right)\sum_{i=1}^{n}c_i^{(m)}$$

となる．誤差関数 E の式の形から，$\alpha_{(m)}$ と $\boldsymbol{\theta}_{(m)}$ に関する最小化は次のように行う．

まず，誤差関数 E の第 1 項の係数

$$J^{(m)}=\sum_{i=1}^{n}c_i^{(m)}I\left\{\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\neq y_i\right\}$$

を，$\boldsymbol{\theta}_{(m)}$ に関して最小化する．その値を $\hat{\boldsymbol{\theta}}_{(m)}$ とする．続いて，誤差関数 E を $\alpha_{(m)}$ について最小化すると，$\hat{\alpha}_{(m)}$ は

$$\hat{\alpha}_{(m)}=\log\left(\frac{1-\hat{\varepsilon}_{(m)}}{\hat{\varepsilon}_{(m)}}\right),$$

$$\hat{\varepsilon}_{(m)}=\frac{\sum_{i=1}^{n}c_i^{(m)}I\left\{\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)\neq y_i\right\}}{\sum_{i=1}^{n}c_i^{(m)}}$$

で与えられる．

$\hat{\alpha}_{(m)}$ と $\hat{\boldsymbol{\theta}}_{(m)}$ が見つかったので，重み $\left\{c_i^{(m)}\right\}$ が更新できる．更新式は

$$c_i^{(m+1)}=c_i^{(m)}\exp\left\{-\frac{1}{2}y_i\hat{\alpha}_{(m)}\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)\right\}$$

で与えられる．ここで，恒等式

$$y_i\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)=1-2I\left\{\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)\neq y_i\right\}$$

を用いて，更新式を書き直せば，

$$c_i^{(m+1)}=c_i^{(m)}\exp\left(-\frac{1}{2}\hat{\alpha}_{(m)}\right)\exp\left[\hat{\alpha}_{(m)}I\left\{\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)\neq y_i\right\}\right]$$

を得る．詳細は，Freund & Schapire（1996）[5] を参照．

以上のことから，AdaBoost のアルゴリズムは，次のようにまとめられる．

(AdaBoost)

1．データの重み c_i $(i=1,2,\cdots,n)$ の初期化．$c_i^{(1)}=\dfrac{1}{n}$ $(i=1,2,\cdots,n)$ と置く．

2．$m=1,2,\cdots,M$ について，以下を繰り返す．

(i) 予測式 $\varphi\left(\mathbf{x},\boldsymbol{\theta}_{(m)}\right)$ について，次の量

$$J^{(m)}=\sum_{i=1}^{n}c_i^{(m)}I\left\{\varphi\left(\mathbf{x}_i,\boldsymbol{\theta}_{(m)}\right)\neq y_i\right\}$$

を最小にするような $\hat{\boldsymbol{\theta}}_{(m)}$ を求める．

(ii) 次の値を，順次計算する．

$$\hat{\varepsilon}_{(m)}=\frac{\sum_{i=1}^{n}c_i^{(m)}I\left\{\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)\neq y_i\right\}}{\sum_{i=1}^{n}c_i^{(m)}},$$

$$\hat{\alpha}_{(m)}=\log\left(\frac{1-\hat{\varepsilon}_{(m)}}{\hat{\varepsilon}_{(m)}}\right)$$

(iii) 以下の式で，データの重みの更新を行う．

$$c_i^{(m+1)}=c_i^{(m)}\exp\left[\hat{\alpha}_{(m)}I\left\{\varphi\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_{(m)}\right)\neq y_i\right\}\right]\ \left(i=1,2,\cdots,n\right)$$

3．最終の予測式の構成．

$$\hat{\varphi}_M\left(\mathbf{x}\right)=\sum_{m=1}^{M}\hat{\alpha}_{(m)}\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(m)}\right)\begin{cases}\geq 0\Rightarrow\mathbf{x}\in G_1\\ <0\Rightarrow\mathbf{x}\in G_2\end{cases}$$

$\mathbf{x}$ が与えられたとき，$y=1$ となる確率を $\Pr\left(y=1|\mathbf{x}\right)$，$y=-1$ となる確率を $\Pr\left(y=-1|\mathbf{x}\right)$ とすれば，$\mathbf{x}$ が群 G_1 か群 G_2 のいずれに属するのかの判定は，対数オッズ

$$\log\left\{\frac{\Pr\left(y=1|\mathbf{x}\right)}{\Pr\left(y=-1|\mathbf{x}\right)}\right\}\begin{cases}\geq 0\Rightarrow\mathbf{x}\in G_1\\ <0\Rightarrow\mathbf{x}\in G_2\end{cases}$$

に従って行うことが出来る．

AdaBoost で求められる最終の予測式 $\hat{\varphi}_M\left(\mathbf{x}\right)=\sum_{m=1}^{M}\hat{\alpha}_{(m)}\varphi\left(\mathbf{x},\hat{\boldsymbol{\theta}}_{(m)}\right)$ は，上の対数オッズの近似関数になることが判っている．

例８．ブースティング

AdaBoost を我々の調査の学習データに適用して予測式を推定した．得られた予測式を用いて，学習データの学習精度（誤判別率）と，検証データの予測精度（誤判別率）を調べた結果が表８である．反復数（すなわち，一連の予測式の個数）は $M = 40$ としている．

表８．AdaBoost の予測精度

学習精度

実測	予測		
	不良品	良品	合計
不良品	306	0	306
良品	0	344	344
誤判別率=0/650=0			

予測精度

実測	予測		
	不良品	良品	合計
不良品	114	0	114
良品	0	136	136
誤判別率=0/250=0			

表８の結果を見ると，バギングの表７の結果と同様に，今までの手法の中で最良の結果になっている．

Ⅳ．まとめ

我々の調査データに対しての分析結果を表９にまとめる．

表９．各手法の予測精度のまとめ

例	手法	学習精度 (%)	予測精度 (%)
1	線形判別	3.23	26.4
2	ロジスティック判別	0	30
3	MTシステム	0	53.2
4	ニューラルネット	2.77	16.4
5	サポートベクターマシーン	3.07	18.4
6	k近傍法	9.84	21.2
7	バギング	0	0
8	ブースティング	0	0

（注：精度は誤判別率を示し，小さい方が良い．）

表９の結果を見ると，予測精度によって手法は３つのタイプの群に分かれる．まずは，伝統的手法群（線形判別，ロジスティック判別，MT システム），次に，機械学習でおなじみの手法群（ニューラルネット，サポート・ベクター・マシーン，k 近傍法），最後は，統合的手法群（バギング，ブースティング）になっている．学習データの情報を上手に利用しているのが，統合的手法群だと考えられる．

謝辞　調査データの利用を快く許可してくださったユニバーサル製缶株式会社と、調査に携われた打田浩明氏ならびに関係諸氏に深く感謝の意を表する．また，本研究は JSPS 科研費（C）

No.24500266 の助成を受けていることに謝意を表する．

参考文献

[1] Bishop, C. M.（2006）, *"Pattern Recognition and Machine Learning"*, Springer-Verlag New York.（村田昇監訳，「パターン認識と機械学習（上,下）」，2012, 丸善出版）．

[2] Breiman, L.（1996）, *"Bagging predictors"*, Machine Learning, Vol. 26, pp.123-140.

[3] Cristianini, N. and Shawe-Taylor, J.（2000）, *"An Introduction to Support Vector Machines"*, Cambridge University Press.（大北剛訳，「サポートベクターマシーン入門」，2005, 共立出版）．

[4] Efron, B. and Tibshirani, R. J.（1993）, *"An Introduction to the Bootstrap"*, Chapman & Hall.

[5] Freund, Y. and Schapire, R. E.（1996）, *"Experiments with a new boosting algorithm"*, 13th International Conference on Machine Learning, L. Saitta（Ed.）, pp.148-156., Morgan Kaufmann.

[6] Giudici, P. and Figini, S.（2009）, *"Applied Data Mining for Business and Industry"*, 2nd Edition, Wiley.

[7] Hastie, T., Tibshirani, R. and Friedman, J.（2009）, "The Elements of Statistical Learning", 2nd Edition, Springer, New York.（杉山將　他監訳，井尻善久　他訳，「統計的学習の基礎」, 2014, 共立出版）．

[8] 小西貞則，（2010）,「多変量解析入門」，岩波書店．

[9] 夏木崇・打田浩明・磯貝恭史，（2012）,「検査工程における高次元データの統計的分類法に関する研究」，神戸大学大学院海事科学研究科紀要，第 9 号，pp.8-19.

[10] 立林和夫・長谷川良子・手島昌一,（2008）,「入門 MT システム」，日科技連出版社．

[11] Vapnik, V.（1998）, *"Statistical Learning Theory"*, Wiley.