PRML4.1-4.2

@masa_kzm

4章 線形識別モデル

4.1 識別関数（判別関数）

- 4.1.1 2クラス
- 4.1.2 多クラス
- 4.1.3 分類における最小二乗法
- 4.1.4-4.1.6 フィッシャーの判別
- 4.1.7 パーセプトロン

分類問題

分類の目的は、ある入力ベクトルxをK個の離散クラスC_kの１つに割り当てること

入力空間は決定領域に分離される

決定領域の境界を決定境界または決定面と呼ぶ

線形識別モデル

決定面が、入力ベクトルxの線形関数であり、D次元入力空間に対して、その決定面がD-1次元の超平面で定義される。

目的変数

２クラス分類

２値表現

クラスC_1をt=1、クラスC_2をt=0で表現する。

$t \in \{0,1\}$

多クラス分類

1-of-K符号化法

例えば、K=5クラスの場合、クラス２のパターンは

$$\mathbf{t} = (0, 1, 0, 0, 0)^{\mathrm{T}}$$

分類問題に対するアプローチ

識別関数 4.1章

□入力ベクトルxから直接クラスを推定する識別関数を構築する。

確率的識別モデル 4.3章

□条件付き確率分布$p(C_k|x)$を直接モデル化する。

確率的生成モデル 4.2章

□クラスに対する事前確率$p(C_k)$とともに、
クラスで条件付けられた確率密度$p(x|C_k)$を考え、
ベイズ定理より、事後確率$p(C_k|x)$を求める

分類問題に対するアプローチ2

一般線形化モデル

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right)$$

分類問題では、領域(0,1)の値をとる事後確率を予測したい

fは活性化関数

決定面はy(x)=定数に相当し $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$ が定数となる。つまり、関数fが非線形でも、決定面はxの線形関数である。

4.1.1 識別関数

線形識別関数　$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$

wは重みベクトル、w_0はバイアスパラメータ。
$-w_0$はしきい値パラメータと呼ばれる。

y(x)≧0ならば、入力ベクトルxはクラスC_1に割り当てられる。
それ以外は、クラスC_2に割り当てられる。

決定境界は、y(x)=0で定義される。

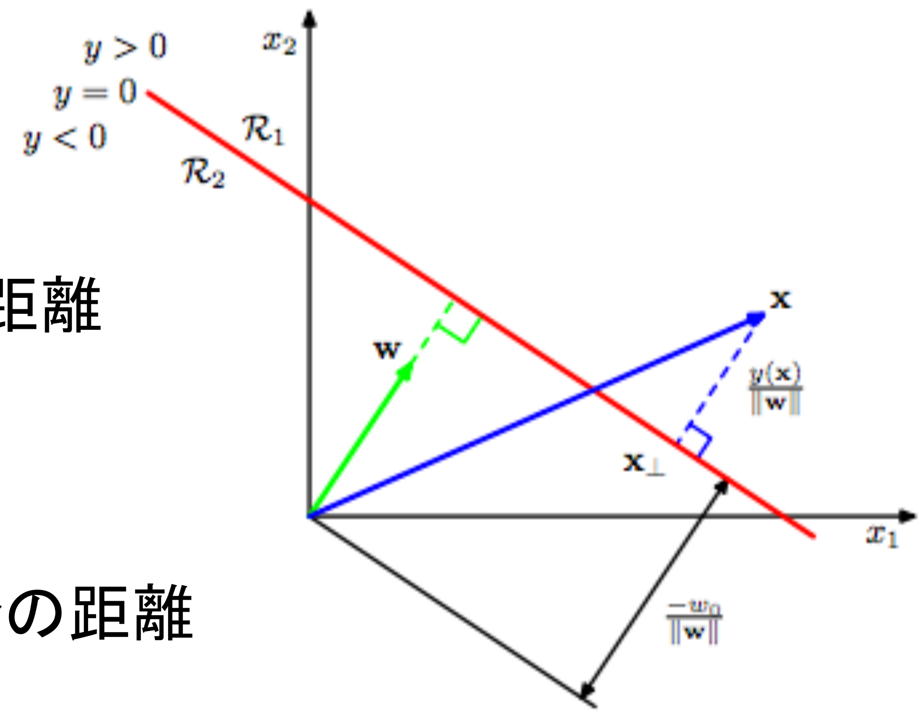
D次元入力空間中のD-1次元超平面に対応する。

決定面の性質

決定面上にあるx_Aとx_Bを考えると下の式が成り立つ。

$$y(\mathbf{x}_\mathrm{A}) = y(\mathbf{x}_\mathrm{B}) = 0,$$

$$\mathbf{w}^\mathrm{T}(\mathbf{x}_\mathrm{A} - \mathbf{x}_\mathrm{B}) = 0$$



原点から決定面までの距離

$$\frac{\mathbf{w}^\mathrm{T}\mathbf{x}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}$$

任意のxから決定面までの距離

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$$

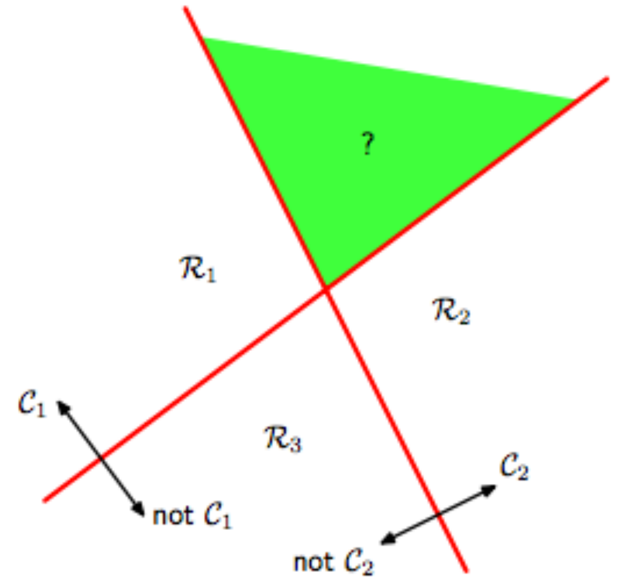
4.1.2 多クラス

1対他分類器

ある特定のクラスC_kに入る点とそのクラスに入らない点とに分類する2クラス問題を解く分類器をK-1個利用する。

緑の部分は、クラスC_1とクラスC_2の両方に所属している。

曖昧な分類領域が出てしまう。

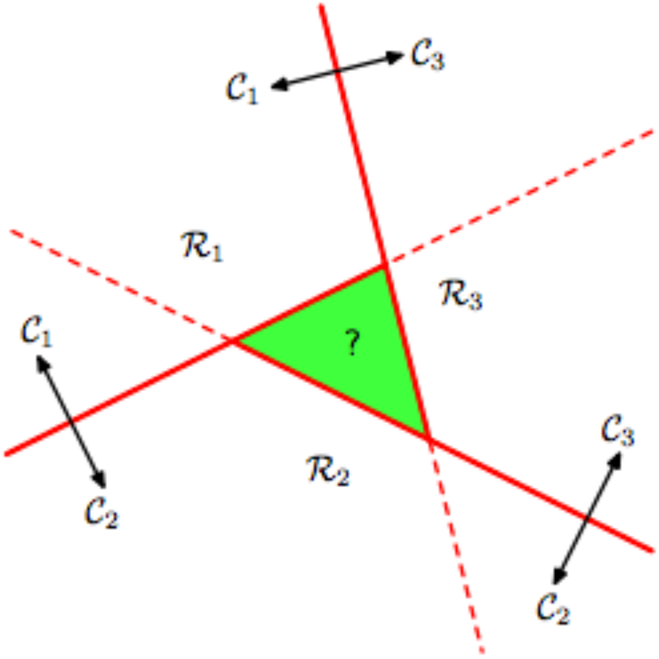


4.1.2 多クラス

1対1分類器

すべての可能なクラスの組の2クラス識別関数を考え、K(K-1)/2個の2クラス識別関数を利用する。

緑の部分は、
「クラスC_1ではなくクラスC_2」
「クラスC_2ではなくクラスC_3」
「クラスC_3ではなくクラスC_1」
である。



曖昧な分類領域が出てしまう。

4.1.2 多クラス

K個の識別関数を考える。

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

最大の$y_k(x)$のクラスC_kに割り当てる。

境界は以下の式で定義される(D-1)次元の超平面に相当する。

$$y_k(\mathbf{x}) = y_j(\mathbf{x})$$

$$(\mathbf{w}_k - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{k0} - w_{j0}) = 0.$$

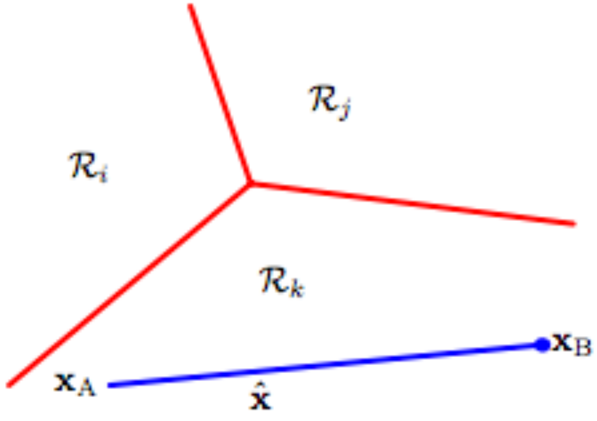
4.1.2 多クラス

凸領域

2点x_Aとx_Bが同じ決定領域R_kにあるとき、
2点x_Aとx_Bを結ぶ直線上にある任意の点も
決定領域R_kにある。

$$y_k(\mathbf{x}) = y_j(\mathbf{x})$$

$$(\mathbf{w}_k - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{k0} - w_{j0}) = 0.$$



凸性の証明

$$\hat{x} = \lambda x_A + (1-\lambda) x_B \qquad 0 \le \lambda \le 1$$

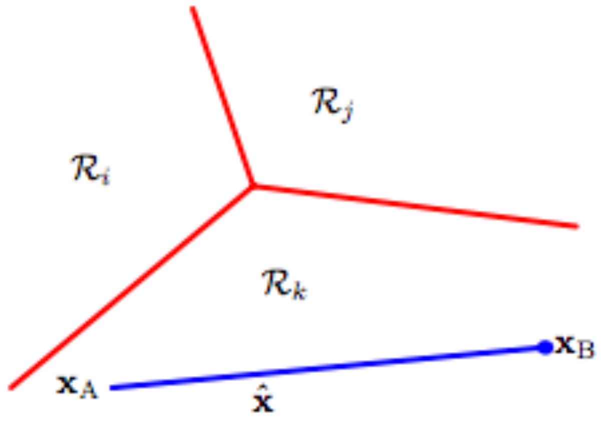
識別関数の線形性より

$$y_k(\hat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_\mathrm{A}) + (1-\lambda) y_k(\mathbf{x}_\mathrm{B})$$

$$y_k(x_B) > y_j(x_B)$$

$$y_k(x_A) > y_j(x_A)$$

$$y_k(\hat{x}) > y_j(\hat{x})$$



よって、2点x_Aとx_Bを結ぶ直線上にある任意の点も決定領域R_kにある。

識別関数のパラメータを学習する方法

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

- 最小二乗　4.1.3章
- フィッシャーの線形判別　4.1.4章
- パーセプトロンアルゴリズム　4.1.7章

4.1.3 分類における最小二乗

目的変数ベクトルtは1-of-K符号化法

$$y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0}$$

$$y(\mathbf{x}) = \widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}$$

$$\widetilde{\mathbf{W}} = \begin{bmatrix}\widetilde{\mathbf{w_1}} & \widetilde{\mathbf{w_2}} & \cdots & \widetilde{\mathbf{w_K}}\end{bmatrix} \qquad \tilde{\mathbf{w}}_k = \begin{bmatrix} w_{k0} \\ w_{k1} \\ \cdots \\ w_{kD} \end{bmatrix} \qquad \tilde{\mathbf{x}} = \begin{bmatrix} 1 \\ x_1 \\ \cdots \\ x_D \end{bmatrix}$$

4.1.3 分類における最小二乗

二乗和誤差関数 $E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$

Wの導関数=0の解 $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \widetilde{\mathbf{X}}^{\dagger}\mathbf{T}$

識別関数 $\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}} = \mathbf{T}^{\mathrm{T}}\left(\widetilde{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}\widetilde{\mathbf{x}}.$

目的変数ベクトルがある定数aとbに対して $\mathbf{a}^{\mathrm{T}}\mathbf{t}_n + b = 0$ を満たすとき、モデルの予測も以下のように同じ制約を満たすことがある。

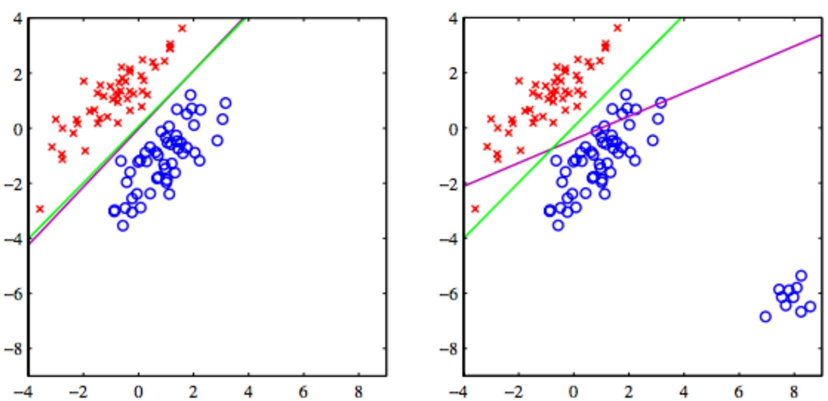
$$\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}) + b = 0$$

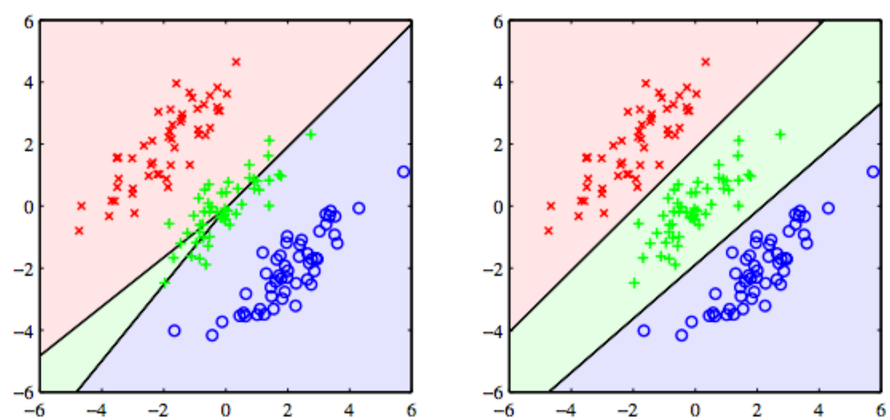
最小二乗法の問題点

外れ値に敏感。頑健性が弱い。

最小二乗法は条件付き確率に
ガウス分布を仮定した場合の最尤法

2値目的変数ベクトルは
ガウス分布からかけ離れている。





4.1.4 フィッシャーの線形判別

2クラス問題について

D次元入力ベクトルを、１次元に射影する。

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

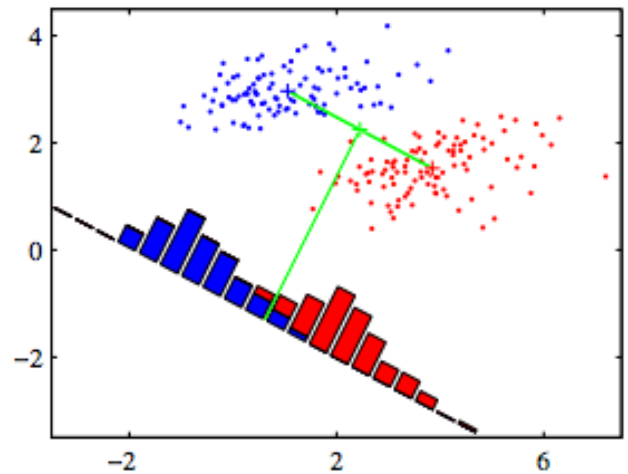
２つのクラスの平均ベクトルは

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}\mathbf{x}_n \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}\mathbf{x}_n$$

射影されたクラスの平均の差を最大にしたい。

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$$

4.1.4 フィッシャーの線形判別



最大化 $\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)$

制約条件 $\mathbf{w}^T\mathbf{w} = 1$

ラグランジュの未定乗数法

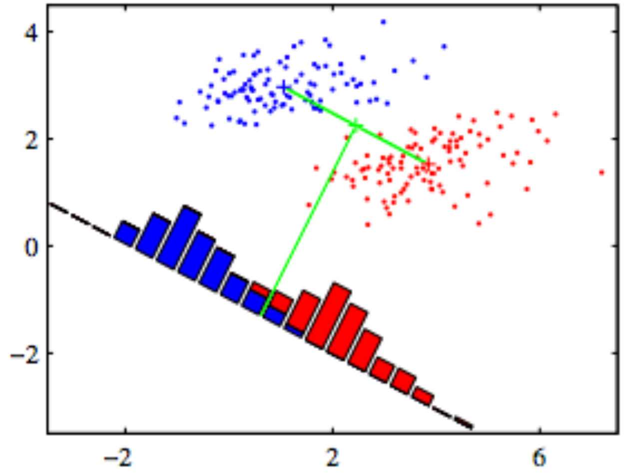
$$L = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1) - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

$$\frac{\partial L}{\partial \mathbf{w}}\mathbf{L} = (\mathbf{m_2} - \mathbf{m}_1) - 2\lambda\mathbf{w}$$

よって $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$

4.1.4 フィッシャーの線形判別

クラス平均を結んだ直線上への射影
重なりあう部分が多い。



フィッシャーの方法

射影されたクラス平均間の分離度を大きくすると同時に、各クラス内では小さな分散を与える関数を最大化する。

フィッシャーの判別基準 = クラス間分散／クラス内分散

4.1.4 フィッシャーの線形判別

クラス内分散は

$$s_1^2 = \sum_{n \in C_1} (y_n - m_1)^2 \qquad s_2^2 = \sum_{n \in C_2} (y_n - m_2)^2$$

フィッシャーの判別基準 = クラス間分散／クラス内分散

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

4.1.4 フィッシャーの線形判別

フィッシャーの判別基準 = クラス内分散／クラス間分散

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\begin{aligned}(m_2 - m_1)^2 &= \left(\frac{1}{N_1} \sum_{n \in C_1} \mathbf{w}^T \mathbf{x_n} - \frac{1}{N_2} \sum_{n \in C_2} \mathbf{w}^T \mathbf{x_n} \right)^2 \\ &= \left(\mathbf{w}^T \left(\frac{1}{N_1} \sum_{n \in C_1} \mathbf{x_n} - \frac{1}{N_2} \sum_{n \in C_2} \mathbf{x_n} \right) \right)^2 \\ &= \left(\mathbf{w}^T (\mathbf{m_2} - \mathbf{m_1}) \right)^2 \\ &= \mathbf{w}^T (\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_B \mathbf{w}\end{aligned}$$

$$\begin{aligned}s_1^2 + s_2^2 &= \sum_{n \in C_1} (y_n - m_1)^2 - \sum_{n \in C_2} (y_n - m_2)^2 \\ &= \sum_{n \in C_1} (\mathbf{w}^T \mathbf{x_n} - \mathbf{w}^T \mathbf{m_1})^2 - \sum_{n \in C_2} (\mathbf{w}^T \mathbf{x_n} - \mathbf{w}^T \mathbf{m_2})^2 \\ &= \mathbf{w}^T \mathbf{S}_W \mathbf{w}\end{aligned}$$

$$\mathbf{S}_B = (\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^T$$

クラス間共分散行列

$$\mathbf{S_W} = \sum_{n \in C_1} (\mathbf{x_n} - \mathbf{m_1})(\mathbf{x_n} - \mathbf{m_1})^T + \sum_{n \in C_2} (\mathbf{x_n} - \mathbf{m_2})(\mathbf{x_n} - \mathbf{m_2})^T$$

総クラス内共分散行列

4.1.4 フィッシャーの線形判別

最大化 $$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}) = \frac{(\mathbf{w}^T\mathbf{S}_B\mathbf{w})'(\mathbf{w}^T\mathbf{S}_W\mathbf{w}) - (\mathbf{w}^T\mathbf{S}_B\mathbf{w})(\mathbf{w}^T\mathbf{S}_W\mathbf{w})'}{(\mathbf{w}^T\mathbf{S}_W\mathbf{w})^2}$$

$$2\mathbf{S}_B\mathbf{w}(\mathbf{w}^T\mathbf{S}_W\mathbf{w}) = (\mathbf{w}^T\mathbf{S}_B\mathbf{w})2\mathbf{S}_W\mathbf{w}$$

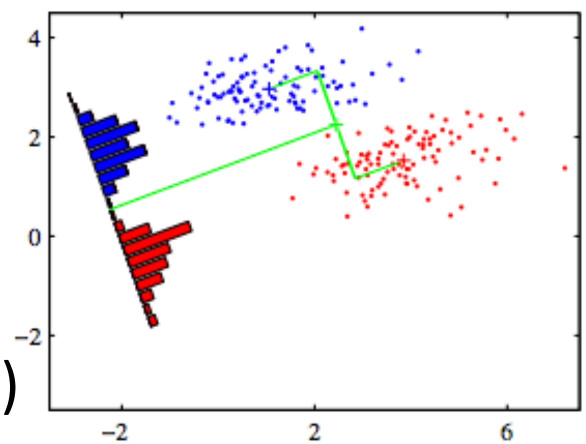
$$(\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w} = (\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w}$$

$$\mathbf{S}_W\mathbf{w} \propto \mathbf{S}_B\mathbf{w} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$$

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

※ $$\frac{\partial}{\partial \vec{x}}(\vec{x}^T\mathbf{A}\vec{x}) = (\mathbf{A} + \mathbf{A}^T)\vec{x}$$

$$= 2\mathbf{A}\vec{x}$$ (Aが対称行列)



4.1.5 最小二乗との関連

最小二乗法

　目的変数値の集合にできるだけ近い予測をすることを目的

フィッシャーの判別基準

　出力空間でのクラス分類を最大にする

２クラス問題において、フィッシャーの判別基準は
最小二乗の特殊な場合である。

4.1.5 最小二乗との関連

クラスC_1に対する目的変数値をN/N_1
クラスC_2に対する目的変数値を$-N/N_2$とする。

N_1はクラスC_1に属するパターンの個数
N_2はクラスC_2に属するパターンの個数

二乗和誤差関数 $$E = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)^2$$

w_0の導関数 $$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right) = 0$$

$$\left(\mathbf{S}_{\mathrm{W}} + \frac{N_1 N_2}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2)$$

wの導関数 $$\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 - t_n\right)\mathbf{x}_n = 0$$

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

フィッシャーの線形判別と同じ

4.1.6 多クラスにおけるフィッシャーの判別

入力空間の次元はD

線形特徴を導入する。k=1,.....,D′ $y_k = \mathbf{w}_k^{\mathrm{T}}\mathbf{x},$

グループ化してベクトル $\mathbf{y}$ で表現する。

重みベクトル{w_k}を列とする行列Wを考える

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$$

DからD′に射影する

4.1.6 多クラスにおけるフィッシャーの判別

クラス内共分散

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \mathbf{S}_k$$

$$\begin{aligned} \mathbf{S}_k &= \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} \\ \mathbf{m}_k &= \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n \end{aligned}$$

総共分散行列

$$\mathbf{S}_{\mathrm{T}} = \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^{\mathrm{T}}$$

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} \sum_{k=1}^{K} N_k \mathbf{m}_k$$

クラス間共分散行列の測度と考えられる行列

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$$

総共分散行列

$$\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{W}} + \mathbf{S}_{\mathrm{B}}$$

4.1.6 多クラスにおけるフィッシャーの判別

射影後

クラス内共分散

$$\mathbf{s}_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (\mathbf{y}_n - \boldsymbol{\mu}_k)(\mathbf{y}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \qquad \boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{y}_n$$

クラス間共分散行列の測度と考えられる行列

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k$$

$$\mathbf{s}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathrm{T}}$$

Fukunaga,1990

クラス間共分散が大きく、クラス内共分散が小さい場合に、大きくなるスカラーを構成。

$$J(\mathbf{W}) = \mathrm{Tr}\left\{\mathbf{s}_{\mathrm{W}}^{-1}\mathbf{s}_{\mathrm{B}}\right\}$$

$$J(\mathbf{w}) = \mathrm{Tr}\left\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\right\}$$

S_Bのランクは高々(K-1)である。(K-1)個以上の線形「特徴」を発見することができない。

4.1.7 パーセプトロンアルゴリズム

入力ベクトルxを特徴ベクトルφ(x)に変換する。

一般化線形モデル　$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)$

非線形活性化関数 f()はステップ関数　$f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0 \end{cases}$

目的変数値 t∈{-1,1}

$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) > 0$ ならC_1

$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) < 0$ ならC_2

すべてのパターンは $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n > 0$ を満たす。

4.1.7 パーセプトロンアルゴリズム

パーセプトロン基準

正しく分類された任意のパターンに対しては誤差0

誤分類された任意のパターンに対しては $-\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)t_n$

$$E_{\mathrm{P}}(\mathbf{w}) = -\sum_{n\in\mathcal{M}} \mathbf{w}^{\mathrm{T}}\phi_n t_n$$

確率的最急降下アルゴリズム

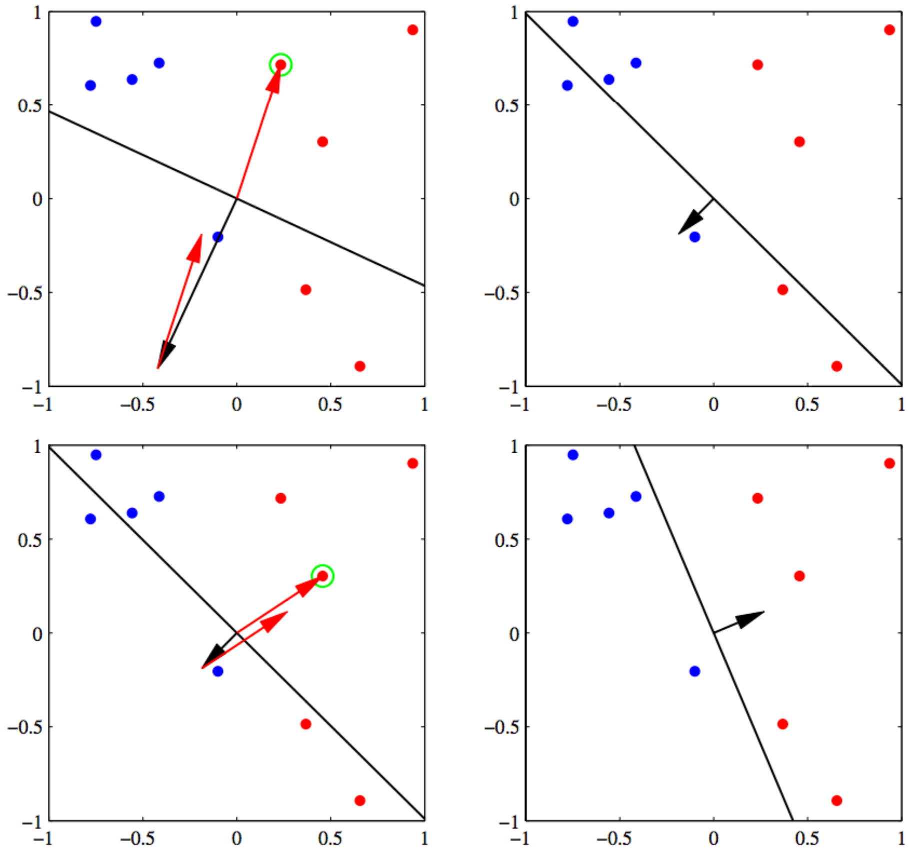
$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla E_{\mathrm{P}}(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta\phi_n t_n$$

パターンが正しく分類されている場合には、重みベクトルに手を加えず、パターンが誤って分類された場合、
誤分類されたパターンがC_1の場合には、ϕ_nを加え
誤分類されたパターンがC_2の場合には、ϕ_nを引く

4.1.7 パーセプトロンアルゴリズム

確率的最急降下アルゴリズム

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_{\mathrm{P}}(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \boldsymbol{\phi}_n t_n$$



4.1.7 パーセプトロンアルゴリズム

一回の更新で、誤分類されたパターンの誤差は減少できる

一回の更新で、新たな誤差が生じることも

一回の更新で、総誤差関数を減少させることを保証していない

パーセプトロンの収束定理

線形分離が可能な場合、パーセプトロン学習アルゴリズムは有限回の繰り返しで厳密解に収束することを保証している。

収束するのに必要な繰り返し回数がかなり多い

初期値やデータの提示順に依存して様々な解に収束してしまう

4.2 確率的生成モデル

クラスの条件付き確率密度$p(x|C_k)$と
クラスの事前確率$p(C_k)$をモデル化して、
ベイズの定理より、
事後確率$p(C_k|x)$を計算する。

4.2 確率的生成モデル

２クラスの場合

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)}$$

$$= \frac{1}{1 + \frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x}|C_1)p(C_1)}}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

ロジスティックシグモイド関数

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

4.2 確率的生成モデル

多クラスの場合

正規化指数関数（ソフトマックス関数）

すべての$j \neq k$に対して、$a_k >> a_j$である場合には、
$p(C_k|x) \approx 1$, $p(C_j|x) \approx 0$となる。

$$
\begin{aligned}
p(\mathcal{C}_k|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} \\
&= \frac{\exp(a_k)}{\sum_j \exp(a_j)}
\end{aligned}
$$

$$
a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)
$$

4.2.1 連続値入力

仮定

クラスの条件付き確率密度がガウス分布

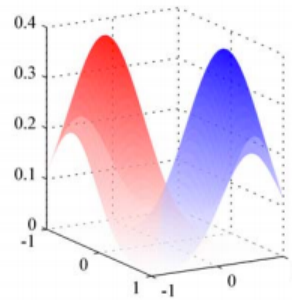
すべてのクラスが同じ共分散行列を共有する

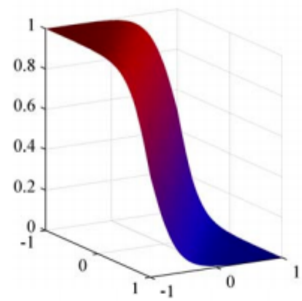
クラスC$_k$の確率密度

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

クラスC$_1$の事後確率

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$





$$\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}
\end{aligned}$$

4.2 確率的生成モデル

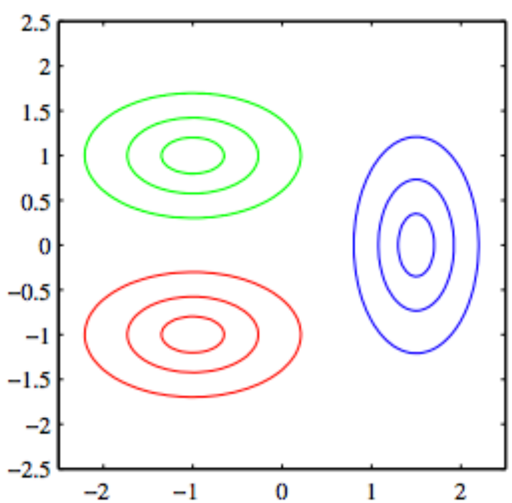
Kクラス分類

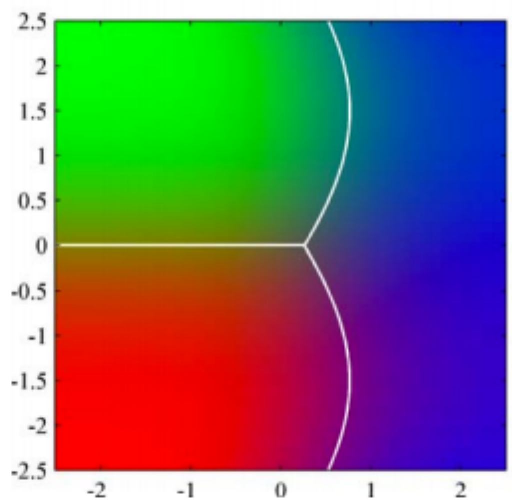
共分散を共有　→　２次の項がキャンセル

一般化線形モデル

$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

$$\begin{aligned} \mathbf{w}_k &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \\ w_{k0} &= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k) \end{aligned}$$





4.2.2 最尤解

クラスの条件付き確率密度$p(x|C_k)$に対するパラメトリックな関数形を決める。

クラスの事前確率$p(C_k)$と、パラメータの値を最尤法で求める。

xの観測値とそれに対応するクラスラベルで構成する
学習データ集合が必要

4.2.2 最尤解

仮定

条件付き確率密度がガウス分布、共通の共分散行列を持つ

データ集合　$\{x_n, t_n\}$　$t \in \{0,1\}$

$t=1$はクラスC_1を表し、$t=0$はクラスC_2を表す。

クラスの事前確率$p(C_1)=\pi$ ，　$p(C_2)=1-\pi$

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

尤度関数

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n} \left[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1-t_n}$$

4.2.2 最尤解

πに関する対数尤度の項 $\sum_{n=1}^{N} \{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)\}$

πに関する最尤推定 $\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$

μ_1に関する対数尤度の項

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = -\frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) + \text{const.}$$

μ_1に関する最尤推定 $\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n$

μ_2に関する最尤推定 $\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1 - t_n) \mathbf{x}_n$

4.2.2 最尤解

∑に関する対数尤度の項

$$-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1)$$

$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_2)$$

$$= -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{N}{2}\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\mathbf{S}\right\}$$

$$\mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^{\mathrm{T}}$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\mathrm{T}}$$

微分すると

$$-\frac{N}{2}(\Sigma^{-1})^T + \frac{N}{2}(\Sigma^{-1}S\Sigma^{-1})^T = 0$$

よって、∑ = Sとなる

※ $$\frac{\partial}{\partial A}\log|A| = (A^{-1})^T \qquad \frac{\partial}{\partial A}tr(A^{-1}B) = -(A^{-1}BA^{-1})^T$$

4.2.3 離散特徴

特徴が離散値x_iの場合を考える。$x_i \in \{0,1\}$

特徴数D個の入力がある場合、一般的な分布は各クラスに対する2^D個の要素の表に相当する。

ナイーブベイズを仮定。特徴値がクラスC_kに対して条件付き独立であるとして扱われる。

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x}$$

$$a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k).$$

$$a_k(\mathbf{x}) = \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1-x_i)\ln(1-\mu_{ki})\} + \ln p(\mathcal{C}_k)$$

入力値x_iの線形関数

4.2.3 指数型分布族

ガウス分布と離散値入力のとき、クラスの事後確率は一般線形化モデルとなる。

クラスの条件付き確率密度$p(x|C_k)$が指数型分布族であるなら、クラスの事後確率は一般線形モデルとなる。

指数型分布族

$$p(\mathbf{x}|\boldsymbol{\lambda}_k) = h(\mathbf{x})g(\boldsymbol{\lambda}_k)\exp\left\{\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

u(x)=xとなるような分布の部分クラスに注目し、
尺度パラメータSを導入する。

$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s}h\left(\frac{1}{s}\mathbf{x}\right)g(\boldsymbol{\lambda}_k)\exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}$$

4.2.3 指数型分布族

クラスの事後確率がxの線形関数a(x)のロジスティックシグモイド関数によって、以下の式のようになる。

2クラス

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2)$$

Kクラス

$$a_k(\mathbf{x}) = \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_k) + \ln p(\mathcal{C}_k)$$

xの線形関数