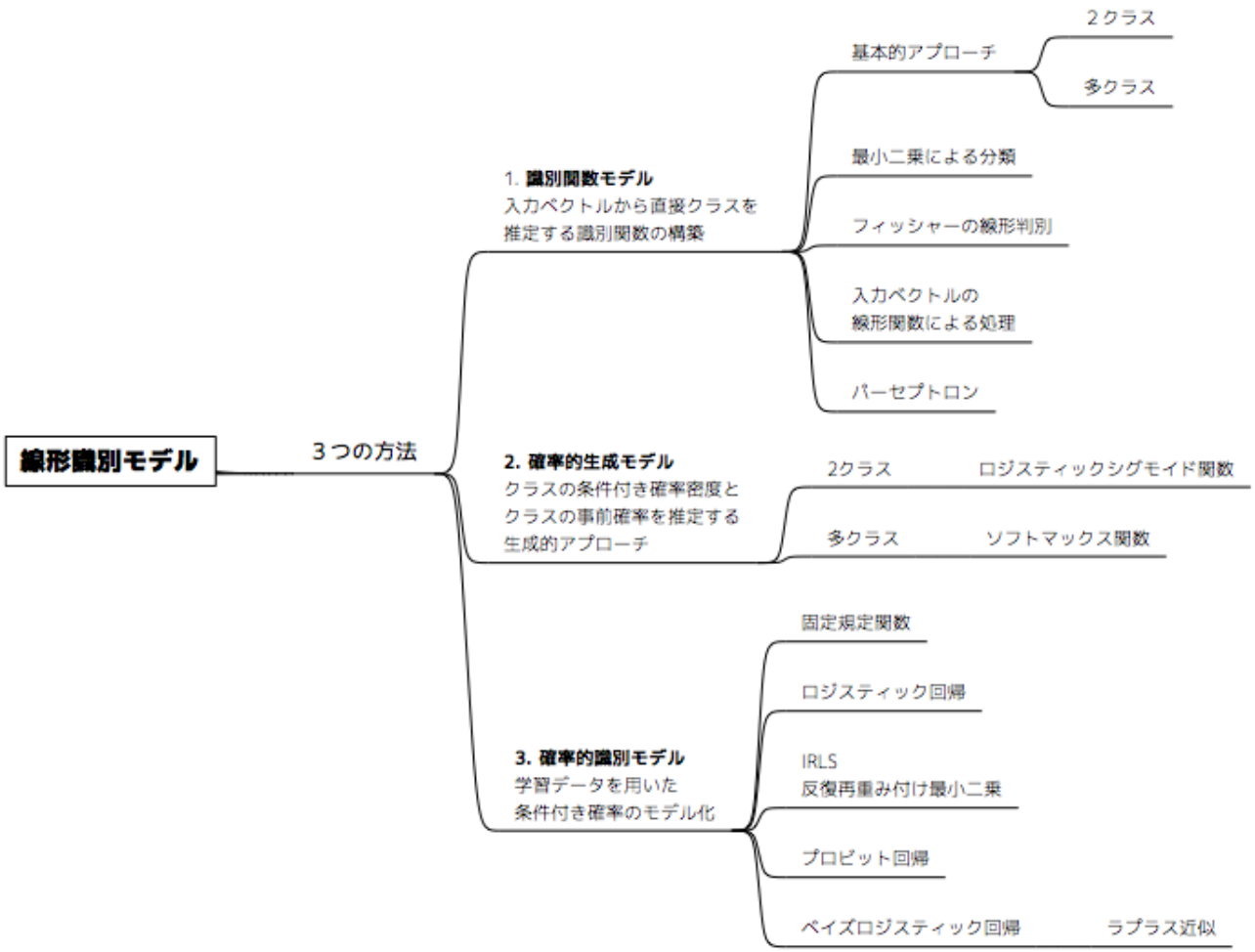
Pattern Recognition and Machine Learning

第４章　線形識別モデル

修士２年
松村草也

BehaviorInNetworks
都市生活学・ネットワーク行動学研究室

4章の流れ


線形識別モデル
3つの方法
1. 識別関数モデル
入力ベクトルから直接クラスを
推定する識別関数の構築
基本的アプローチ
2クラス
多クラス
最小二乗による分類
フィッシャーの線形判別
入力ベクトルの
線形関数による処理
パーセプトロン
2. 確率的生成モデル
クラスの条件付き確率密度と
クラスの事前確率を推定する
生成的アプローチ
2クラス
ロジスティックシグモイド関数
多クラス
ソフトマックス関数
3. 確率的識別モデル
学習データを用いた
条件付き確率のモデル化
固定規定関数
ロジスティック回帰
IRLS
反復再重み付け最小二乗
プロビット回帰
ベイズロジスティック回帰
ラプラス近似

線形識別モデルとは

- ある**入力ベクトルx**の要素を，**K個の離散クラスC_k**に分類することを目的とする.
- 一般的に各クラスは互いに重ならず、各入力は一つのクラスに割り当てられる。
- 分類先を**決定領域**と呼ぶ.
- 決定領域の境界を**決定境界・決定面**と呼ぶ.
- **線形識別モデル**とは
  - 決定面が入力ベクトルxの線形関数で,
  - D次元の入力空間に対して，決定面はD-1次元のモデル.
- 線形決定面によって正しく各クラスに分類できるデータ集合を**線形分離可能**であるという.

分類問題の表記方法について

- 回帰問題では目的変数tは実数値ベクトルだった.
- 分類問題では離散的なクラスラベルを表現するための方法がいろいろある.
- ２クラス分類問題における確率モデルの場合，**２値表現**が一般的である.
- K>2クラスの場合は，目的変数に対して1-of-k表記法を使用するのが便利である.
- クラスがCjの場合，j番目の要素を除くtの要素がすべて0であるような長さKのベクトルが使用される.

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

識別関数モデル

パラメータについて線形な関数

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

さらに一般的に事後確率を予測するため，非線形関数f(・)によって一般化する.
f(・)を**活性化関数(activation function)**とよぶ.

$$y(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + w_0) \tag{4.3}$$

識別関数モデル – 2クラス

まず，単純な線形識別関数についてクラス分類方法を考える.

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \qquad (4.4)$$

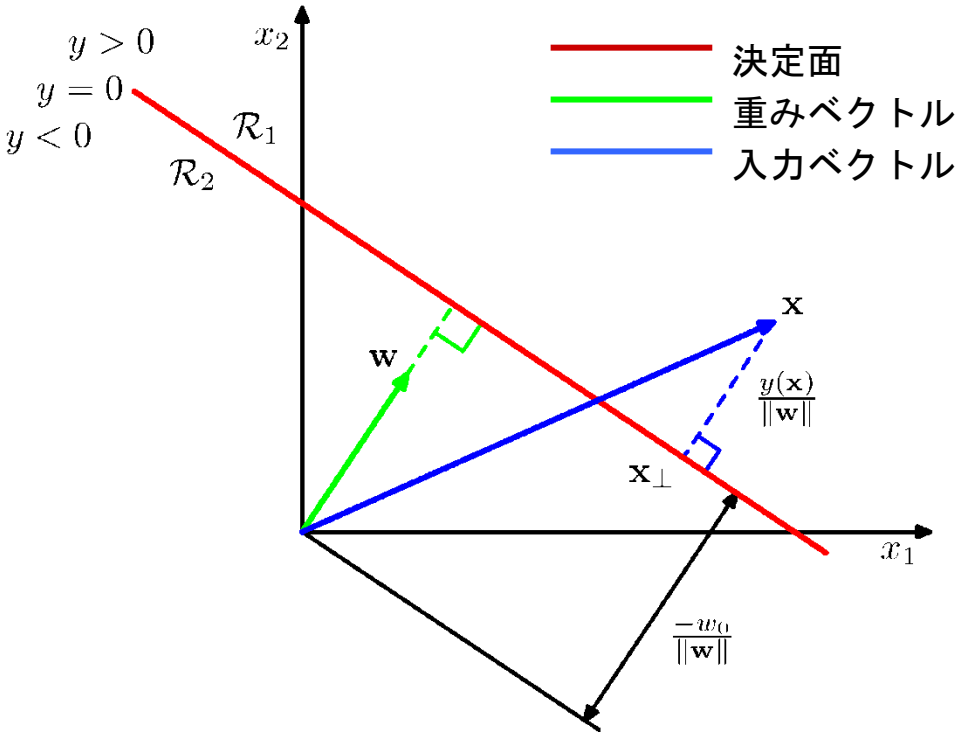
重みベクトル

入力ベクトル

バイアスパラメータ
（マイナスの場合は
閾値パラメータ）

- y(x)>0ならば，**x**はクラスC_1に割り当てられ，それ以外はC_2に割り当てられる.
- wは**重みベクトル**と呼ばれ，決定境界の傾きを決める.
- w_0は**バイアスパラメータ**と呼ばれ，原点からの境界のずれを決める.
- 関係を図示するとわかりやすい.

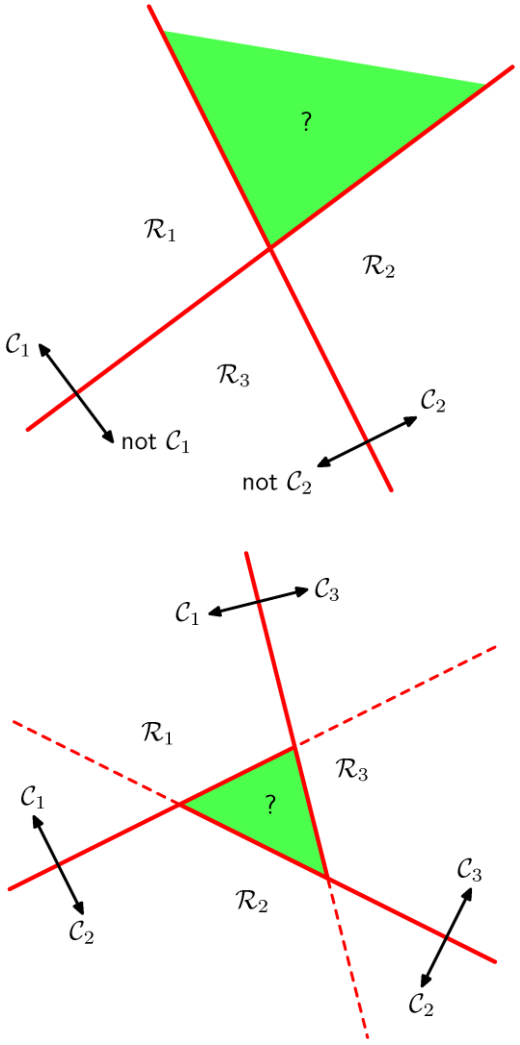
識別関数モデル – 2クラス



- 2次元線形識別関数の幾何的表現.
- 赤で示された決定面はwに垂直である.
- 原点から面までの距離はバイアスパラメータw_0によって制御される.
- 決定面（境界）のどちら側にあるかによって，入力ベクトルのクラスを判別する.

多クラスへの拡張・問題点

- 次に，K=2クラスの線形識別をK>2のクラスへ拡張することを考える.
- 多くの2クラス識別関数の組み合わせでKクラスの識別が構成可能だが，単純に行うと曖昧な領域が生まれてしまう.
- **1対他分類器(one-versus-the-rest classifier)**
  - ある特定のクラスに入る点と入らない点を識別する2クラスをK-1個用意する方法.
- **1対1分類器(one-versus-one classifier)**
  - すべてのクラスの組み合わせを考え，K(K-1)/2個の2クラスを用意する方法

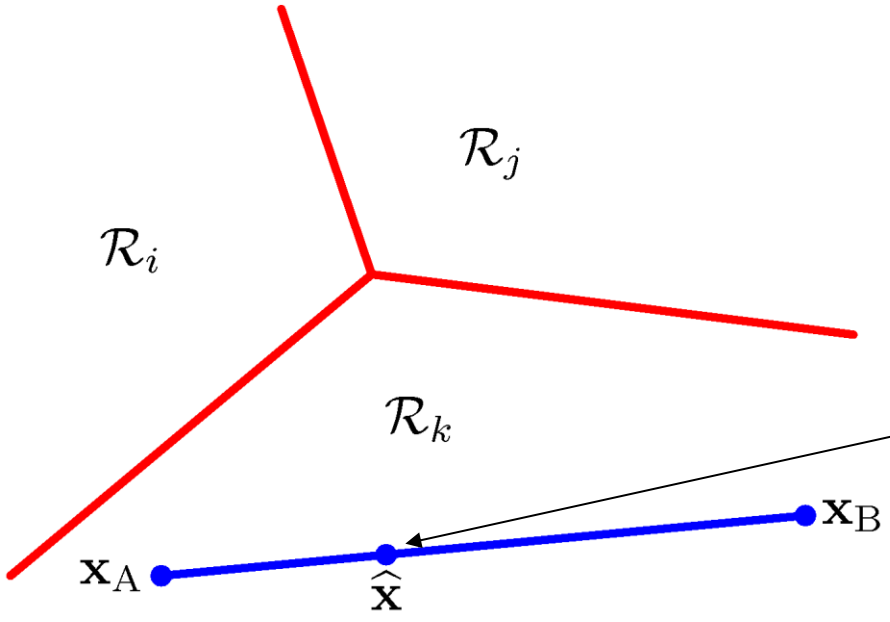


多クラスの決定方法

そこで,

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

というクラスをK個用意する．各**x**については$y_k(\mathbf{x})$の大小を比較することでどのクラスに分類するか決まる．値が等しい時は決定領域になる.



決定領域Rは単一接続しており，凸領域である.

ベクトル $\hat{\mathbf{x}}$ については，下記が成立.

$$\hat{\mathbf{x}} = \lambda \mathbf{x}_A + (1-\lambda)\mathbf{x}_B$$

$$y_k(\hat{\mathbf{x}}) = \lambda y_k(\mathbf{x}_A) + (1-\lambda) y_k(\mathbf{x}_B)$$

最小二乗法を用いた分類

- 3章ではパラメータに関する線形モデルを考え，二乗和誤差の最小化により，最適なパラメータが解析的に求められることを確認した．そこで同じ定式化を分類問題にも適用してみる．
- 一般的なKクラス分類問題についても最小二乗を使用する理由は，入力ベクトルが与えられた際の目的変数値の条件付き期待値を近似するから（？）
- しかし，推定された確率は一般的に非常に近似が悪く，線形モデルの柔軟性が低いために，確率の値が(0,1)の範囲を超えてしまうこともある．

最小二乗法を用いた分類

3章では

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

の二乗和誤差関数を最小にすることを考えた．二乗和誤差関数は，

$$E_D(\tilde{\mathbf{W}}) = \frac{1}{2} tr\{(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^T (\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})\}$$

と，書くことができる．ただし，$\mathbf{T}=t_n^T$　　Wに関する導関数を0とおくと

$$\tilde{\mathbf{W}} = \tilde{\mathbf{X}}^{\dagger}\mathbf{T}$$

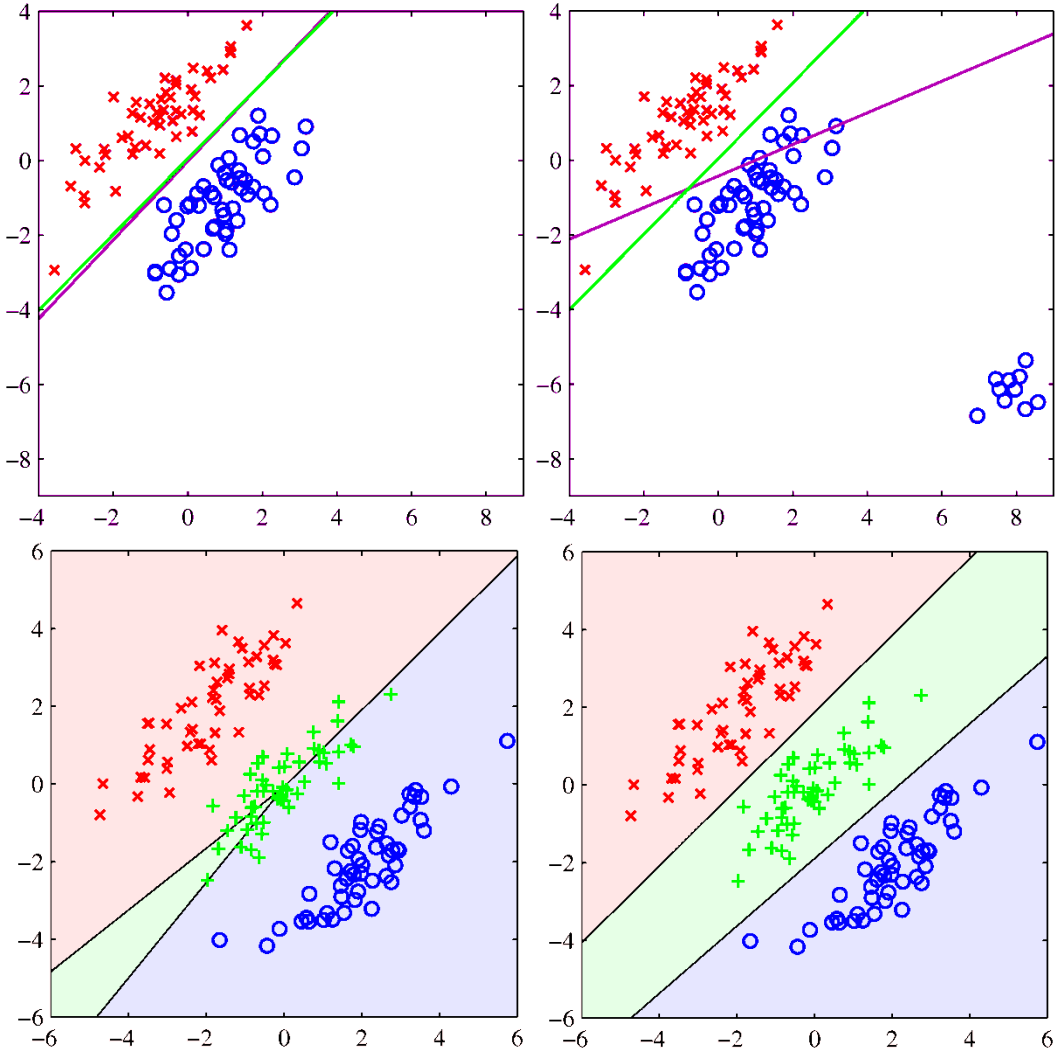
$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^T \tilde{\mathbf{x}} = \mathbf{T}^T (\tilde{\mathbf{X}}^{\dagger})^T \tilde{\mathbf{x}}$$

最小二乗法を用いた分類

- 最小二乗法は識別関数のパラメータを求めるための解析解を与えるが，いくつかの難しい問題を抱えている.
- 2.3.7節で，最小二乗法は**外れ値に対する頑健さが欠けている**ことを見た.
- 3クラスの分類に対しても十分なクラスを集合に対して与えられない.
- これは，**最小二乗法は条件付き確率分布にガウス分布を仮定した場合の最尤法**であるが，**２値目的変数ベクトルは明らかにガウス分布からかけ離れている**ことが原因である.

最小二乗法の脆弱性

- 緑色はロジスティック回帰モデル，紫は最小二乗によって得られる決定面.
- 外れ値が右下にある場合，最小二乗は過敏に反応していることがわかる.
- 下段は３クラスの分類.
- 左図は最小二乗による分類．緑色のクラスについては誤識別が大きい.
- 右図はロジスティック回帰モデルで，うまく分類できていることがわかる.

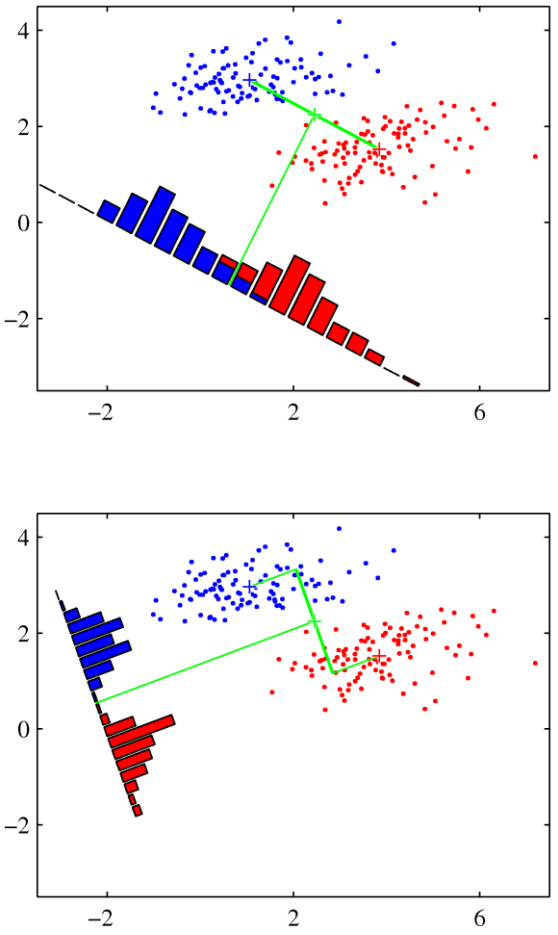


次元の削減

- 次元の削減，という観点から線形識別モデルを見ることができる．まず2クラスの場合を考える．D次元の入力ベクトルを，1次元に射影するとする．yにある閾値を設定した，2クラス分類.
- 一般的に1次元への射影は相当量の情報の損失を発生させるので，**元のD次元空間では分離されていたクラスが1次元空間では大きく重なってしまう可能性がある**.
- このとき，**重みベクトルを調整することでクラスの分離を最大にする射影を選択する**ことができる.
- この手法は「**フィッシャーの線形判別(fisher's linear discriminant)**」として知られている.

フィッシャーの線形識別

- 図のプロットは2クラスからのサンプルを示している.
- また，クラス平均を結ぶ直線上に射影された結果をヒストグラムで示している.
- 射影空間では無視できないくらいにクラスが重なり合っていることがわかる.
- 下の図のプロットはフィッシャーの線形判別に基づく射影を示す.
- クラス分離度を大きく改善していることがわかる.

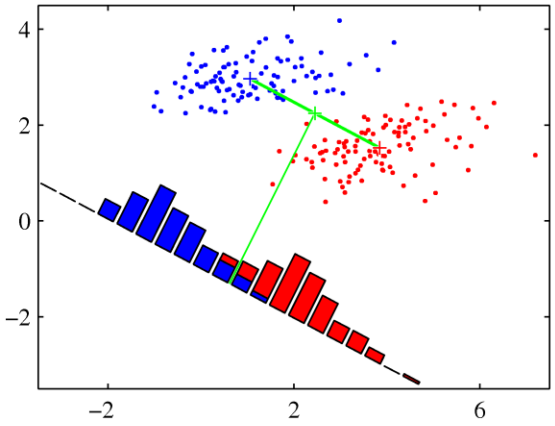


フィッシャーの線形識別

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in C_1}\mathbf{x}_n, \quad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in C_2}\mathbf{x}_n \tag{4.21}$$

$$m_2 - m_1 = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1) \tag{4.22}$$

$$m_k = \mathbf{w}^T\mathbf{m}_k \tag{4.23}$$



- 境界面を1次元の射影とみたときの解決方法．wを調整することで，クラス間の分離度，式(4.23)を最大にする射影を選ぶ.
- wの長さは単位長であるという制限を加える.
- この方法では，射影結果に重なりが生じている.

フィッシャーの線形判別

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2 \quad (4.24)$$

ラベルづけされたデータに対するクラス内分散

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \quad (4.25)$$

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \quad (4.26)$$

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \quad (4.27)$$

$$\mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T \quad (4.28)$$

(4.26)を最大にすることを考える．wで微分して，

$$(\mathbf{w}^T \mathbf{S}_B \mathbf{w}) \mathbf{S}_W \mathbf{w} = (\mathbf{w}^T \mathbf{S}_W \mathbf{w}) \mathbf{S}_B \mathbf{w} \quad (4.29)$$

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \quad (4.30)$$

フィッシャーの線形判別

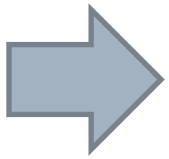
最小二乗との関連

- 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
- フィッシャーの判別基準
  - 出力空間でのクラス分離を最大にすること
- これまでは1-of-K表記法を考えてきたが，それとは異なる**目的変数値の表記法**を使うと，
- **重みに対する最小二乗解がフィッシャーの解と等価になる**． (Duda and Hart, 1973)

最小二乗との関連

二乗和誤差関数

$$E = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n + w_0 - t_n)^2$$



w_0とwに関する
導関数を0とする

$$\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n + w_0 - t_n) = 0$$

$$\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}_n + w_0 - t_n)\mathbf{x}_n = 0$$

$$w_0 = -\mathbf{w}^T\mathbf{m}$$

$$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}(N_1\mathbf{m}_1 + N_2\mathbf{m}_2)$$

$$(\mathbf{S}_W + \frac{N_1N_2}{N}\mathbf{S}_B)\mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2)$$

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

パーセプトロンアルゴリズム

- パーセプトロンはあるきまった非線形変換を用いて，入力ベクトルxを変換して特徴ベクトルを得て，以下の式で表わされる一般化線形モデルを構成する．(4.52)
- 2クラス分類問題では目的変数値の表記法として，t∈{0,1}を用いていたが，パーセプトロンではステップ関数で与えられる．(4.53)
- 今回は誤差関数として，**誤識別したパターンの総数**を選択する．

$$y(\mathbf{x}) = f(\mathbf{w}^T \phi(\mathbf{x})) \quad (4.52)$$

$$f(a) = \begin{cases} +1 & (a \geqslant 0) \\ -1 & (a < 0) \end{cases} \quad (4.53)$$

パーセプトロンアルゴリズム

- 誤差がwの区分的な定数関数であり，wの変化に伴い変化する決定境界が，データ点を横切るたびに不連続となるため，誤差関数の勾配を使ってwを変化させる方法が使えない.
- そこで，パーセプトロン規準として知られている別の誤差関数を考える.
- t∈{-1,+1}という目的変数値の表記方法をもちいると，すべてのパターンは正の値を取る.

$$y(\mathbf{x}) = f(\mathbf{w}^T\phi(\mathbf{x})) \qquad (4.52)$$

$$f(a) = \begin{cases} +1 & (a \geqslant 0) \\ -1 & (a < 0) \end{cases} \qquad (4.53)$$

$$\begin{cases} \mathbf{w}^T\phi(\mathbf{x}_n) < 0 \\ \mathbf{w}^T\phi(\mathbf{x}_n) > 0 \end{cases}$$

$$\mathbf{w}^T\phi(\mathbf{x}_n)t_n > 0$$

パーセプトロンアルゴリズム

■パーセプトロン基準

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

Mは誤分類されたすべてのパターンの集合を表す．ただし$\varphi_n=\varphi(x_n)$.
w空間でパターンが誤分類される領域内では，誤分類されたパターンの誤差への寄与は0である．よって総誤差関数は区分的に線形.

■確率的最急降下アルゴリズムの適用

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

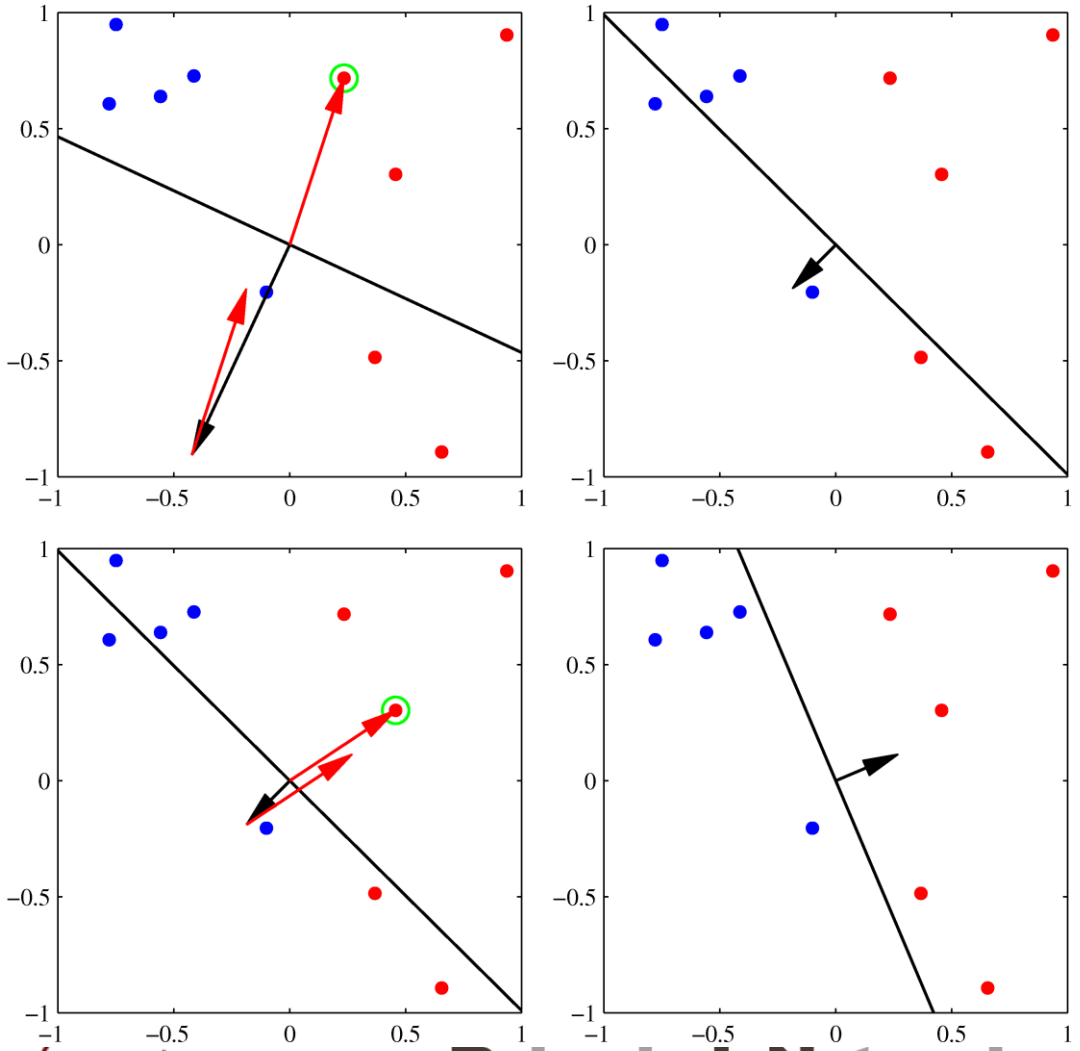
ηは学習率パラメータ，τはアルゴリズムのステップ数を表す.
ηは1にしても一般性は失われない.

$$-\mathbf{w}^{(\tau+1)T}\phi_n t_n = -\mathbf{w}^{(\tau)T}\phi_n t_n \boxed{-(\phi_n t_n)^T \phi_n t_n} < -\mathbf{w}^{(\tau)T}\phi_n t_n$$

誤差の減少

パーセプトロンアルゴリズムの学習特性

- 初期のパラメータベクトルを決定境界とともに黒矢印で表示.
- 緑の円で囲まれたデータは誤分類されている.
- その特徴ベクトルが現在の重みベクトルに追加される.
- さらに考慮すべき次の誤分類点を示す.
- 誤分類点の特徴ベクトルをまた重みベクトルに追加，右下の決定領域を得る.



パーセプトロンの弱み

- 更新の対象としていない誤分類パターンからの誤算関数への寄与は減少しているとは限らない.
- 重みベクトルの変化は，以前正しく分類されていたパターンを誤分類させるようなこともあり得る.
- しかし，**パーセプトロンの収束定理**では，厳密解が存在する場合（学習データ集合が線形に分離可能な場合）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.
- しかし，必要な繰り返し回数はかなり多くて実用的かどうかは怪しい.
- パラメータの初期値，データの提示順に依存してさまざまな解に収束する.
- 線形分離可能でないデータ集合に対しては決して収束しない.

確率的生成モデル

- 分類を確率的な視点からとらえ，線形決定境界を持つモデルがどのように生成されるかを示す.
- クラスの条件付き確率密度$p(\mathbf{x}|C_k)$とクラスの事前確率$p(C_k)$をモデル化する生成的アプローチについて議論する.

ロジスティックシグモイド関数

- 2クラスの場合の事後確率を**ロジスティックシグモイド関数**を用いて表現する.
- K>2クラスの場合は，**正規化指数関数**で表現され，これは**ソフトマックス関数**としても知られている.
- 入力変数**x**が連続値をとる場合と離散値をとる場合について，クラスの条件付き確率密度が特定の形で与えられる時の結果を調べる.

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

ロジスティックシグモイド関数

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

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_j p(\mathbf{x}|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln(p(\mathbf{x}|C_k)p(C_k))$$

連続値の入力

- クラスの条件付き確率密度$p(\mathbf{x}|C_k)$がガウス分布であると仮定して，事後確率の形をみてみる.
- まずすべてのクラスが同じ共分散行列$\sum$を共有すると仮定.

離散値の特徴

- 特徴が離散値xiの場合を考える．簡単のため，xi∈{0,1}から
- 特徴数がD個入力がある場合は，一般的な分布は各クラスに対する2D個の要素の表に相当する．
- そこには2D-1個の独立変数が含まれている．これでは特徴数が指数関数的に増加してしまうので，より制限的な表現を考える．
- ここで，ナイーブベイズを仮定する．

指数型分布族

- ガウス分布と離散値入力の両方に対して，クラスの事後確率がロジスティックシグモイド関数もしくはソフトマックス活性化関数の一般化線形モデルで与えられることがわかった.
- これらは，クラスの条件付き確率が指数型分布族のメンバーであるという仮定によって得られる一般的な結果の特殊解.

確率的識別モデル

- 2クラス分類問題では，多くのクラスの条件付き確率分布$p(\mathbf{x}|C_k)$に対してその事後確率がxのロジスティックシグモイド関数で書ける.
- 同様に多クラスの場合はxの線形関数のソフトマックス変換によって与えられることを見てきた.
- 特定のクラスの条件付き確率密度に対して，そのパラメータと事前確率を最尤法によって決定でき，ベイズの定理を用いて事後確率が求められることを示してきた.

- 別のアプローチとして，一般化線形モデルの関数形式を陽に仮定し，最尤法を利用して一般化線形モデルのパラメータを直接決定する方法もある. （確率的識別モデル）