学習機械としての パーセプトロン

神経回路システム特論

26/Sep/2006

今回目指すもの

神経回路にインスパイアされたパターン認識系であるパーセプトロンに関する紹介をおこなう.

パーセプトロンの構造は, 神経細胞・回路との類似点があるが, 今回はその計算工学的側面に焦点を絞っている.

構成

パーセプトロンとは？
汎化
サポートベクターマシーン (SVM)

パーセプトロンの定義

Minsky, Papert

パーセプトロンとは，与えられた論理関数の集合Φに関して，すべての線形な論理関数を計算しうる装置である．

パーセプトロンの定義

Hecht-Nielsen

図で示される処理要素が1つ以上で構成されたネットワーク．各処理要素もパーセプトロンと呼ばれる．

パーセプトロンの定義

wikipedia

a type of artificial neural network invented in 1957 by Rosenblatt. It can be seen as the simplest kind of feedforward neural network.LinkBack Item from LaTeXiT

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

パーセプトロンの構造

入力層（S層）：sensory layer 外部入力の受容

中間（隠れ）層（A層）：association layer

出力層（R層）：response layer 入力に対する最終結果出力

パーセプトロンの構造

パーセプトロンの要素

- 荷重
- 素子, ニューロン

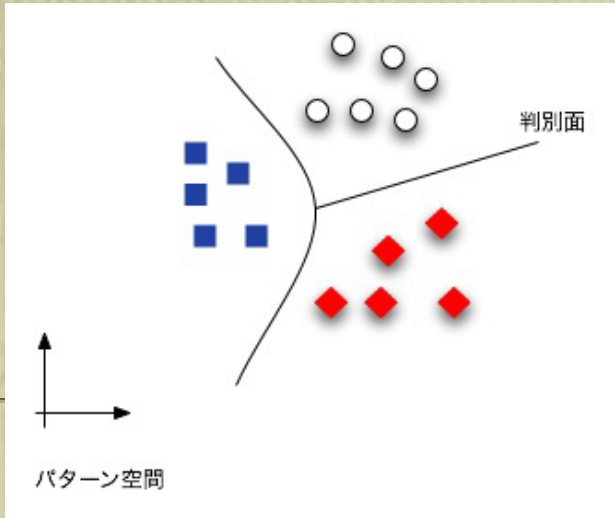
接続

- 縦続接続(feedforward)
- 同じ層内部での相互結合なし.
- 隣接する層間のみ接続.

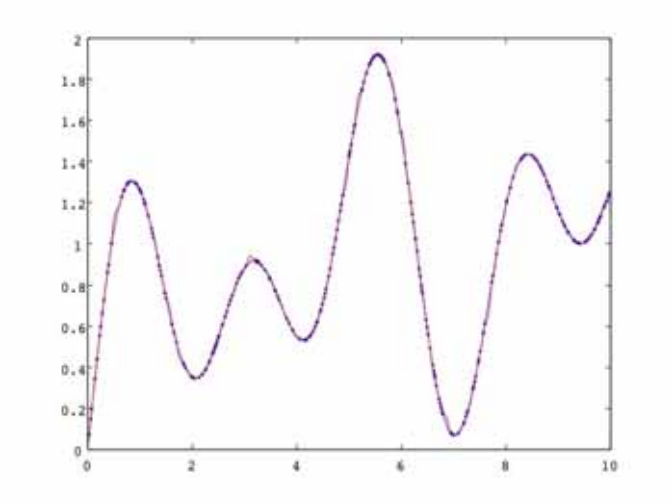
パーセプトロンが扱う問題

パターン認識問題

入力パターンの分類をおこなう.



関数近似

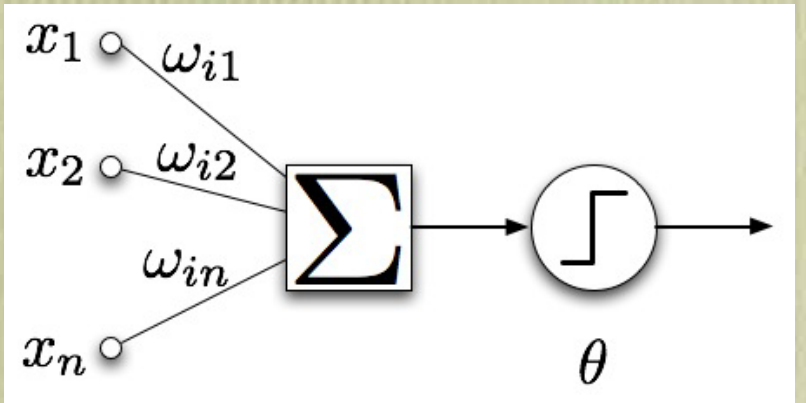


パーセプトロンの要素

McCulloch-Pitts model

入力：x, **出力**：y=0,1

多数決による意思決定システム



今回の目標

例えば，生命の本質である動的変動を内在しない系は脳研究と距離をおいて議論するべきものである．

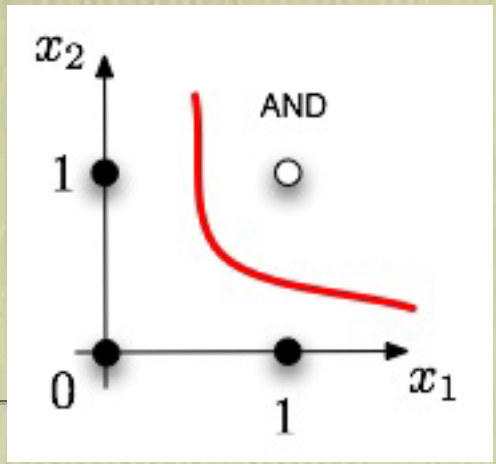
実際のニューロンモデルを単純化したパーセプトロン（ニューラルネットワーク）に関する研究は，いくらか脳の働き・知能に関する示唆を与えるかもしれないが，本目的はパーセプトロンを用いた情報処理に関する議論という工学的観点に主眼をおいている．

パーセプトロンの能力と限界

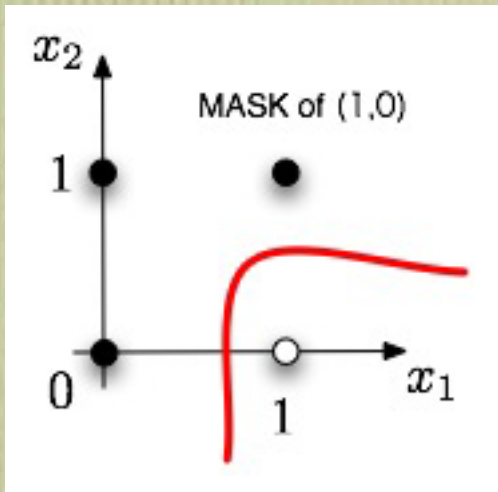
単純パーセプトロン

例 (McCulloch-Pitts model)

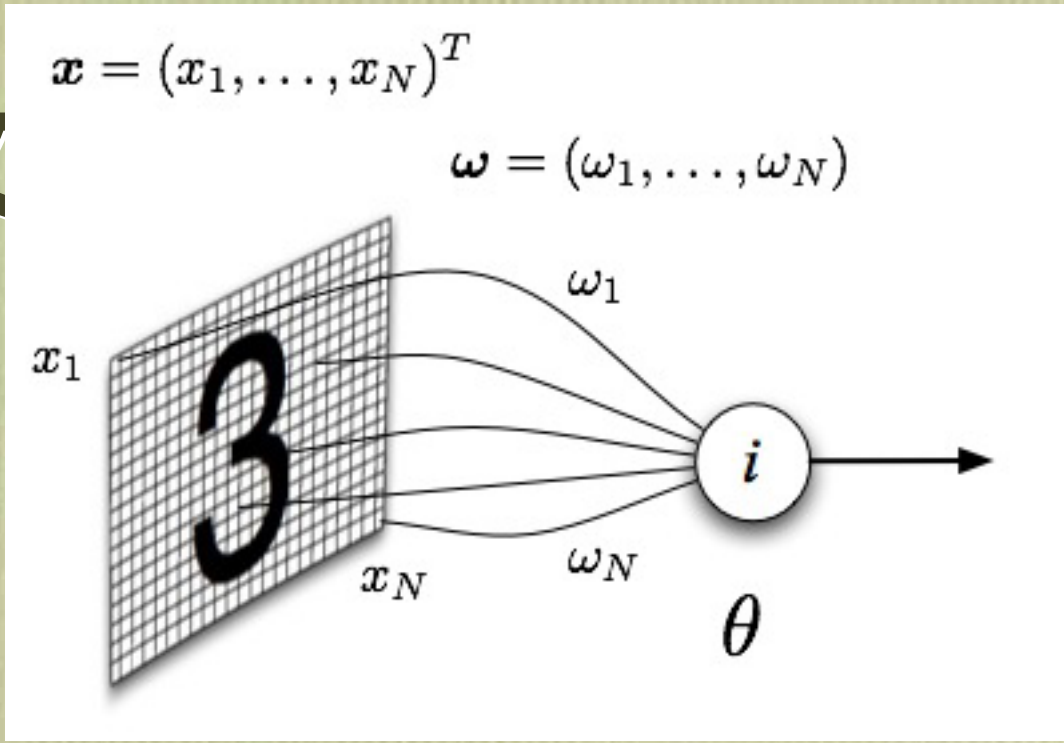
AND



Match filter (MASK)



文　　コン



"覚える"パーセプトロンを構築することは可能.

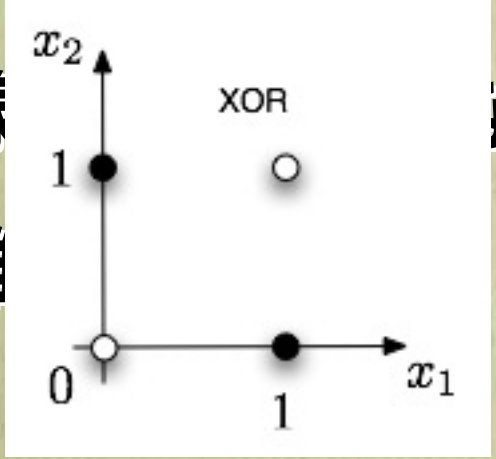
しかし，パターン認識系においては，異なるパターンを「区別する」ことが必要になることから，「学習」のためには対となるパターンも必要である.

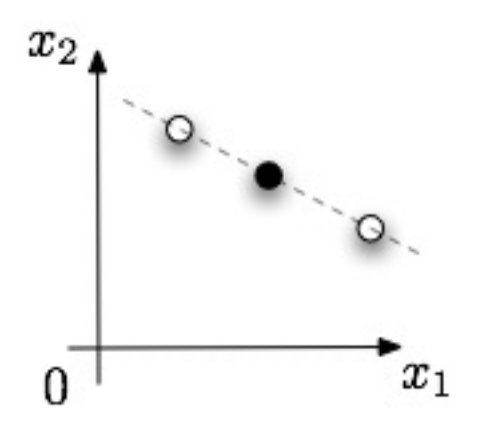
単純パーセプトロンの表現能力

線形分離性

線 なパタ 可能

分離





単純パーセプトロンの学習

学習とは？

入力と出力を別に持つパーセプトロンに対して，正しい入出力の対のリスト（訓練セット）を提示し，（パーセプトロンの実現精度の範囲 内で）正しい入出力関係を実現する関数を構築.

単純パーセプトロンの学習(ニューロサイエンスの視点から)

Hebb則

心理学者Hebbの仮説：同時に発火したニューロン間のシナプス結合は強められる

When an axon of cell A is near enough to excite cell B and repeatedly or consistently takes part in firing it, some growth process or metabolic changes takes place in one or both cells such that A's efficiency, as one of the cell B, is increased. (Hebb,D.O.,(1949), The Organization of Behavior)

状況証拠

例えば，生後間もなく異常な視覚環境におかれた猫や猿は，成長後視覚における特徴抽出に異常がみられる．

勾配降下法 単純パーセプトロンの学習

パラメータ（荷重）空間中における最適解の探索.

（工学的視点から）

勾配降下法

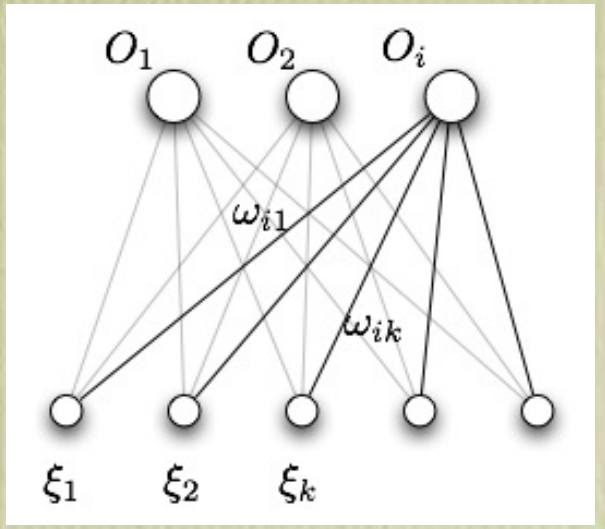
手段

パラメータ (荷重) 空間中における最適解の探索. 学習目標である教師パターンと実際のパーセプトロンのパフォーマンスとの差を評価し, 改善する.

荷重ベクトルが定義域となるエネルギー曲面を滑り落ちていき, 最終的に極小解へ到達する方法.

評価関数

$$E(\omega) = \frac{1}{2}\sum_i \left[\zeta_i^\mu - g\left(\sum_k \omega_{ik}\xi_k^\mu\right)\right]^2$$



出力の自乗誤差和

$$\omega_{ik} \to \omega_{ik} + \Delta\omega_{ik}$$

$$\Delta\omega_{ik} = -\eta\frac{\partial E}{\partial \omega_{ik}}$$

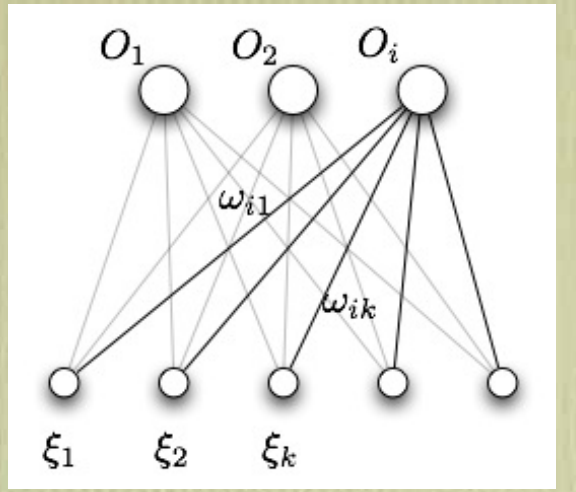
パラメータ修正

問題点と改善

McCulloch-Pitts modelのようにstep function
では，エネルギー曲面が滑らかにならず，
修正の手がかりとなる勾配が発生しない．

そこで，ステップ関数を連続に修正したtanhあ
るいはシグモイド関数を用いる．

$$y(x) = \tanh(x) \qquad y(x) = \frac{1}{1+\exp(-2\beta h)}$$



$$\Delta\omega_k = -\eta\frac{\partial E^\mu}{\partial\omega_k}$$ の算出

$(\xi_1^\mu, \ldots, \xi_k^\mu, \ldots)$ をパターンμ の入力とすると，パターンμ および期待出力ζ^μが与えられたとき，誤差エネルギー関数 $E^\mu = \frac{1}{2}\sum_i(\zeta^\mu - O^\mu)^2$

$$O^\mu = g(\sum_k \omega_k \xi_k^\mu)$$

次式が成立することを確認.

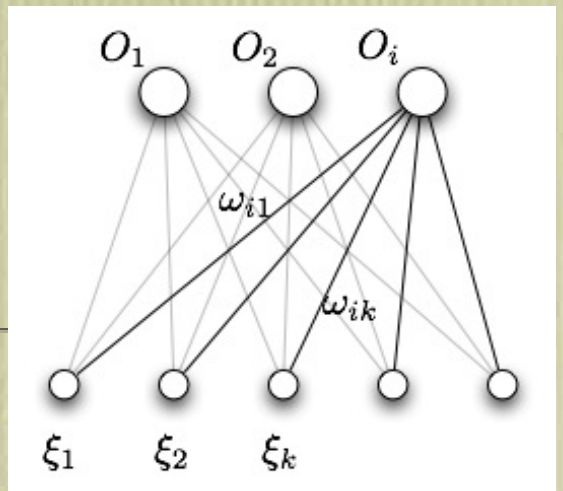
$$g(h) = \tanh(\beta h) \rightarrow \quad g'(h) = \beta(1 - g^2)$$

$$g(h) = (1 + \exp(-2\beta h))^{-1} \rightarrow \quad g'(h) = 2\beta g(1 - g)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

出力　入力

$$\Delta\omega_k = \eta\delta^\mu\xi_k^\mu$$



$$\delta^\mu = [\zeta^\mu - O^\mu]g'(h^\mu)$$

$$h^\mu = \sum_k \omega_k\xi_k^\mu$$

$$g(h) = \tanh(\beta h) \qquad g'(h) = \beta(1-g^2)$$

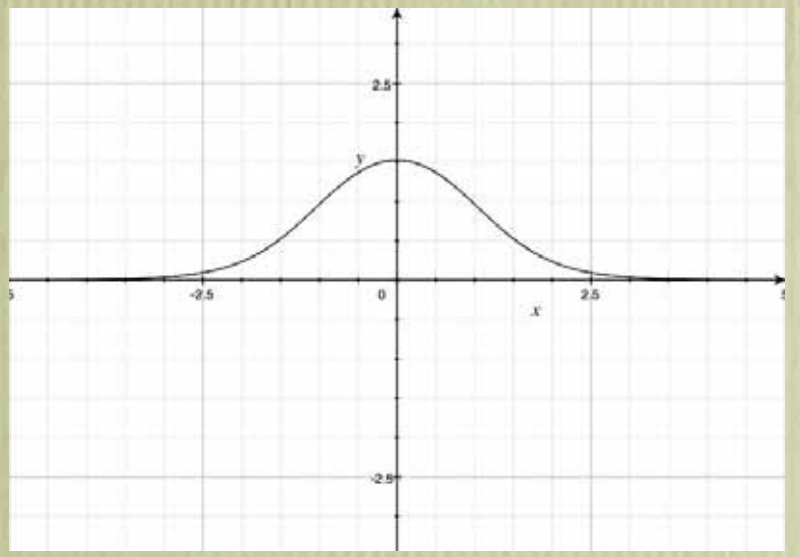
$$g(h) = (1+\exp(-2\beta h))^{-1} \qquad g'(h) = 2\beta g(1-g)$$

関数近似

多層化による能力の改善

2つの活性化関数の組み合わせ（差分）により，局所的なピークを持つ関数が発生する．

$$y = \tanh(x+1) - \tanh(x-1)$$



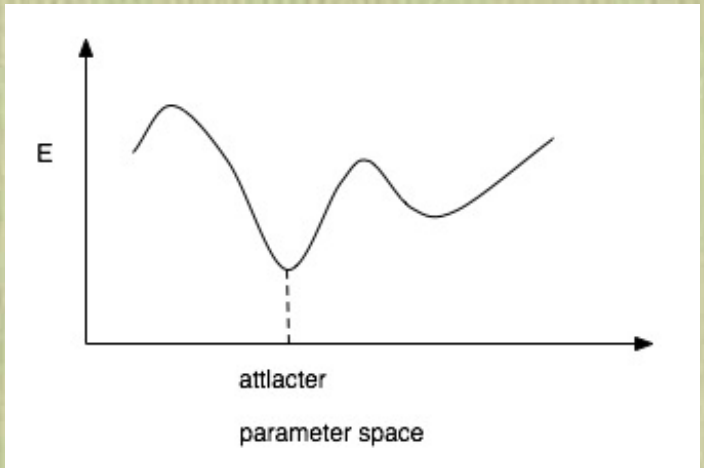
関数近似

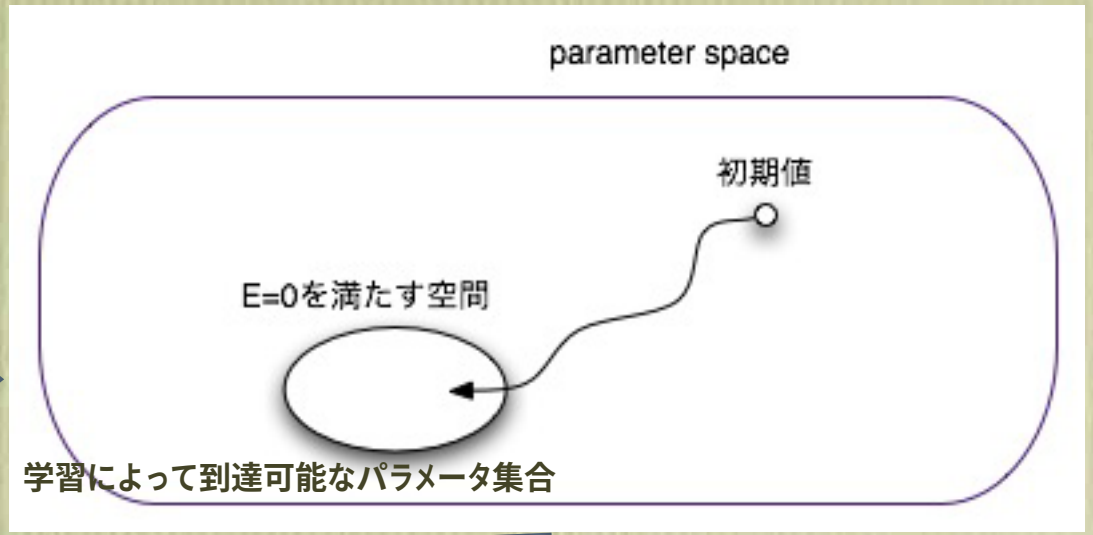
任意の連続関数は，局所的なピークを持つ関数の線形和により近似可能である．

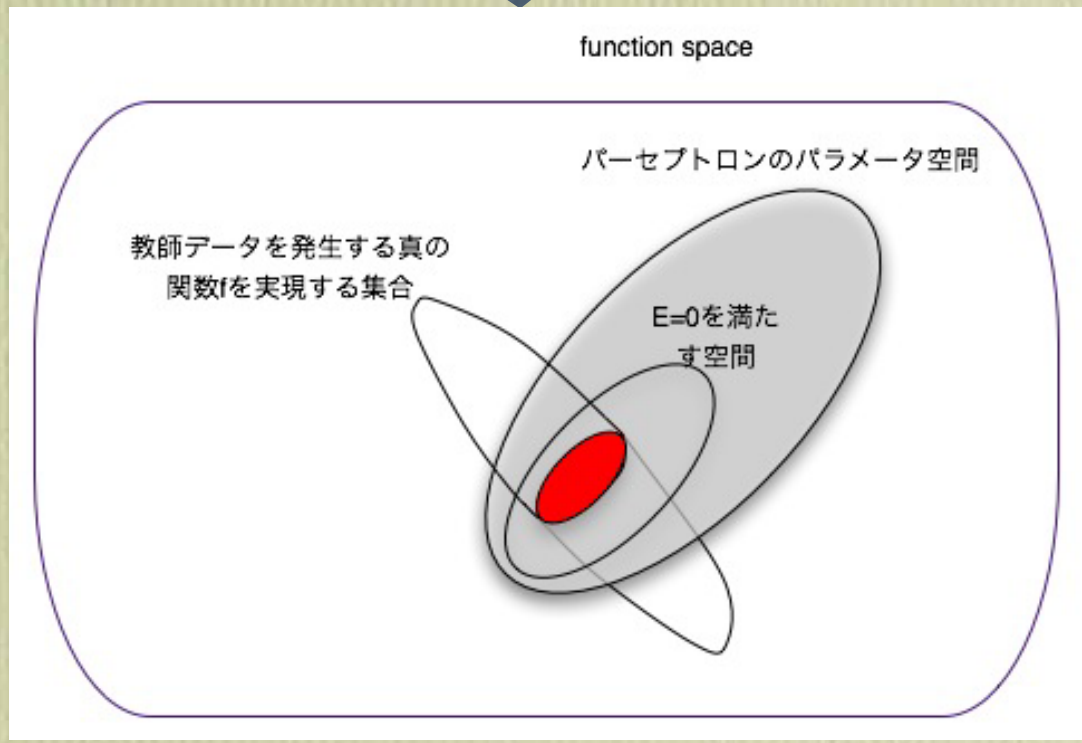
Kolmogorovの定理

任意の連続関数 $f:[0,1]^n \to \boldsymbol{R}, f(\boldsymbol{x})$ が与えられたとき，fは，第一層（すなわちx入力層）にn個のファンアウト処理要素，中間層に(2n+1)個の処理要素，そして最上層（すなわちy出力層）にm個の処理要素を持つ3層フィードフォワードニューラルネットワークにより正確に実現できる．

汎化の定義







学習によって到達したパラメータ集合が，真の関数集合に含まれるか否かに関して議論されるべきだが，前述のKolmogorovの定理より，十分な中間素子数を与えられた多層パーセプトロンを想定する．

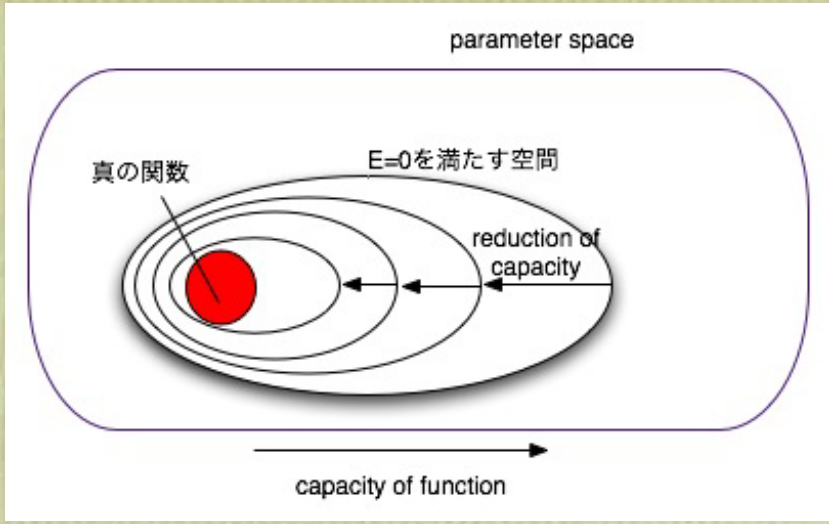
汎化とは？

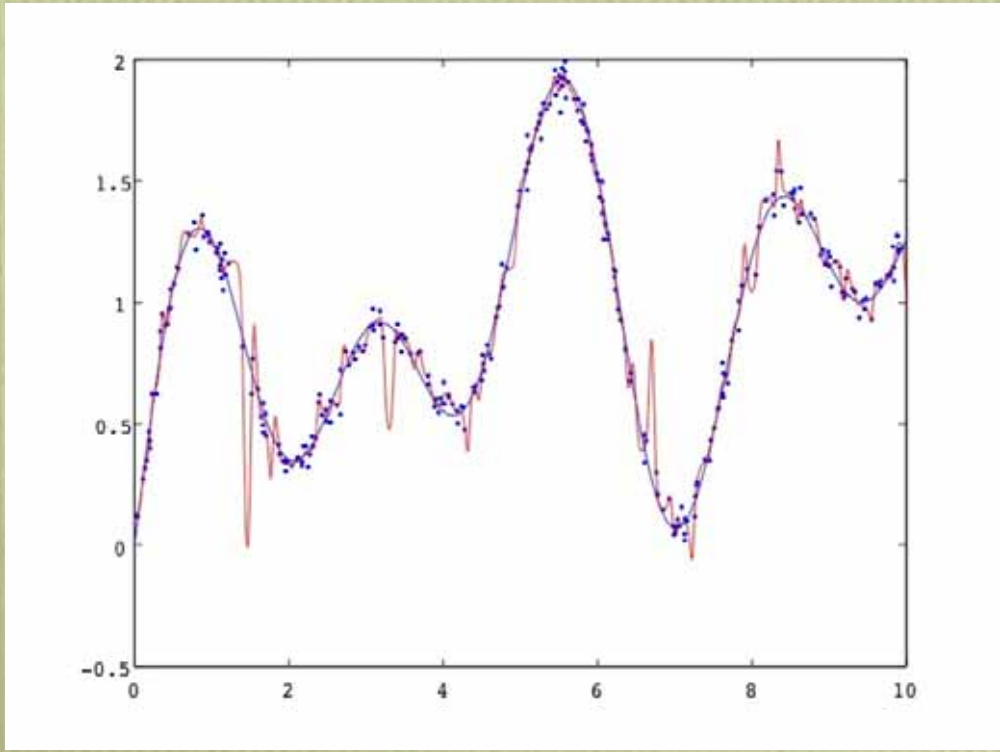
上述の勾配降下法で目指した”学習”は，empirical riskの最小化により，expected riskの最小化を期待するものであるが，ここまでのところ，その保証に関しては議論されていない．

直感的に，例えば，データ数の増大，モデルの簡素化などが汎化能力向上に効果的であると予想される．

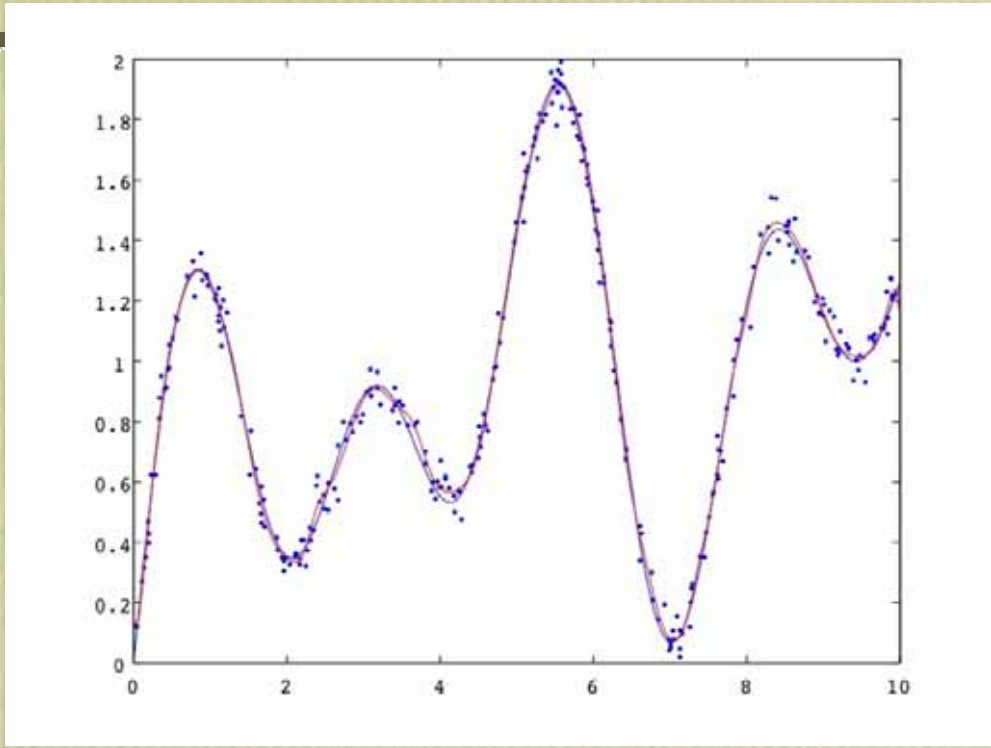
汎化とは？

overfitting





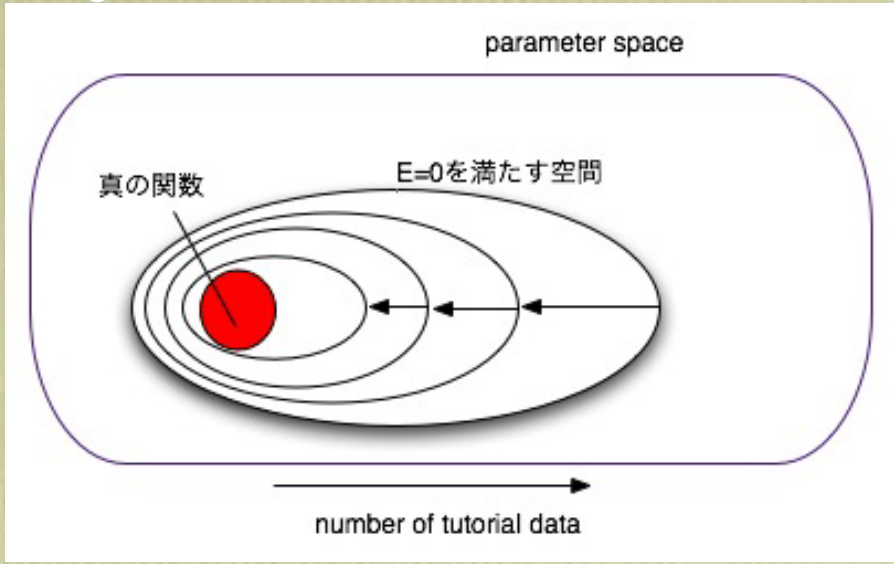
中間素子数：100

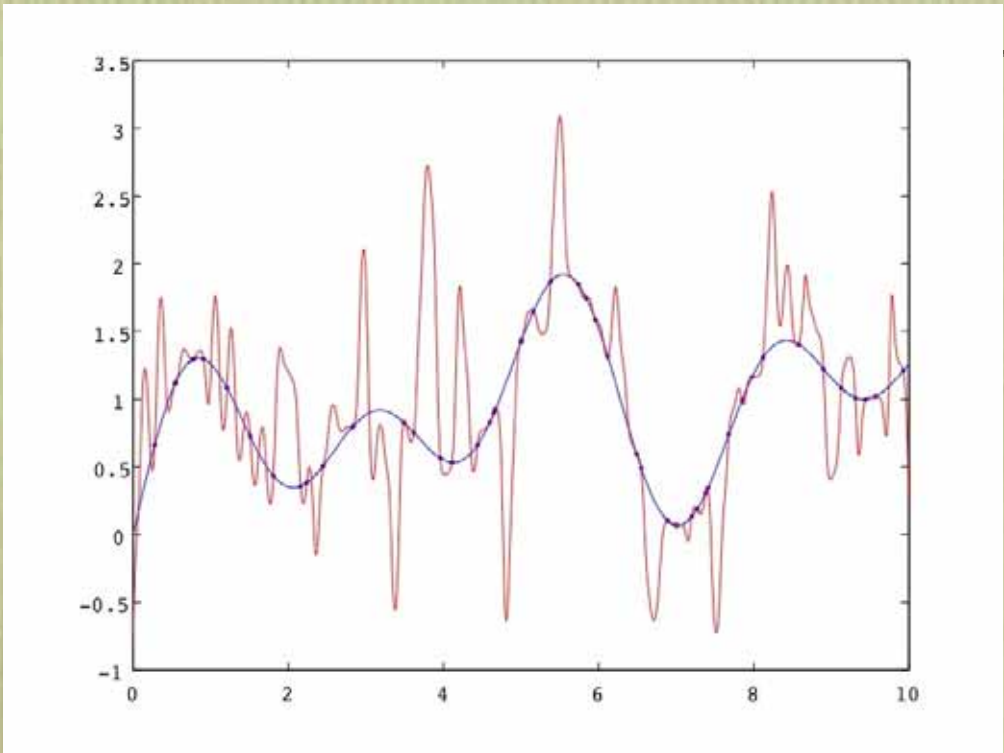


中間素子数：20

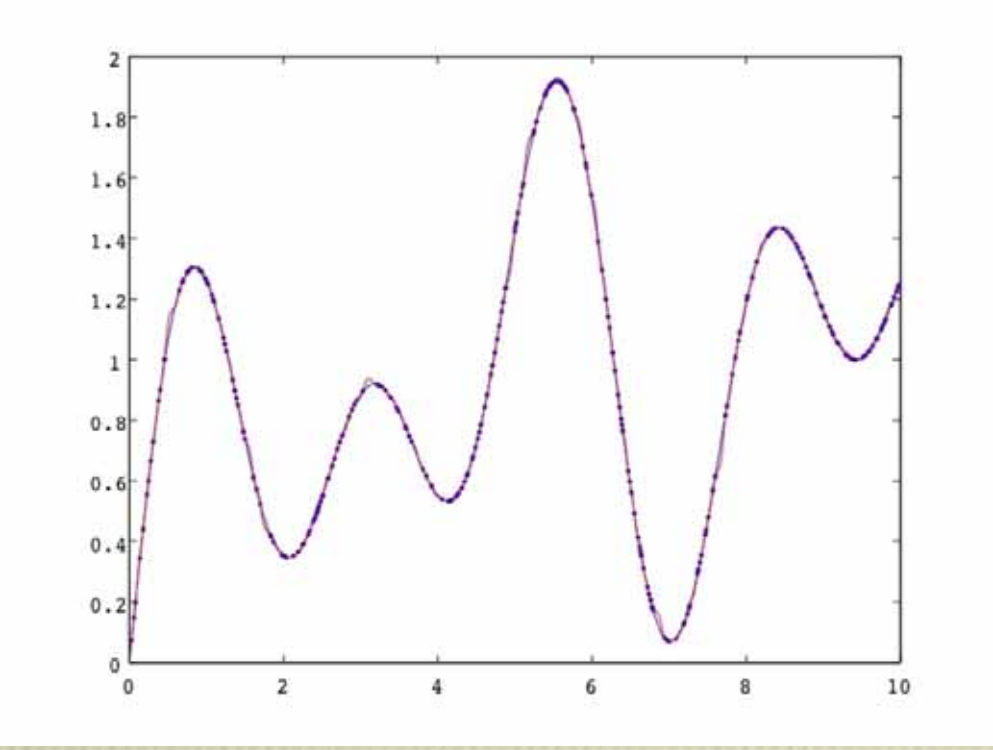
汎化とは？

データ数増大（中間素子数100）





データ数：50



データ数：300

学習において期待すること

term

observation: $Z = \{\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in \mathbb{R}^n, y_i = \{-1, 1\}, i = 1, \ldots, \ell\}$

Empirical risk: $R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(\boldsymbol{z}_i, \alpha)$

Expected risk: $R(\alpha) = \int Q(\boldsymbol{z}, \alpha) dF(\boldsymbol{z})$

loss:

regression: $Q_1(\boldsymbol{z}, \alpha) = (y_i - f(\boldsymbol{x}_i, \alpha))$

, pattern recognition: $Q_2(\boldsymbol{z}, \alpha) = |y_i - f(\boldsymbol{x}_i, \alpha)|$

parameter of model: α

学習において期待すること

Empirical リスク Remp と 推定リスク R

パーセプトロンの学習とは, Empirical riskをゼロにすることである.

これにより, 真の関数へ収束することを期待している.

$$\lim_{\ell \to \infty} P\{\sup_f (R[f] - R_e mp[f]) > \epsilon\} = 0, \forall \epsilon > 0$$

上述の勾配降下法で目指した”学習”は，empirical riskの最小化により，expected riskの最小化を期待するものであるが，ここまでのところ，その保証に関しては議論されていない．

直感的に，例えば，データ数の増大，モデルの簡素化などが汎化能力向上に効果的であると予想される．

期待リスクとEmpiricalリスク

pattern recognition

すべての $f \in \mathcal{F}$, h**に関して，少なくとも確率** $1-\eta$ **で，次の不等式が成立する．**

$$R[f] \leq R_{emp}[f] + \phi\left(\frac{h}{\ell}, \frac{\log(\eta)}{\ell}\right)$$

VC confidne

$$\phi\left(\frac{h}{\ell}, \frac{\log(\eta)}{\ell}\right) = \sqrt{\frac{h\left(\log\frac{2\ell}{h}+1\right) - \log(\eta/4)}{\ell}}$$

h**が小さいほど，サンプル数**l**が多いほど**VC confidence**は小さくなる．**

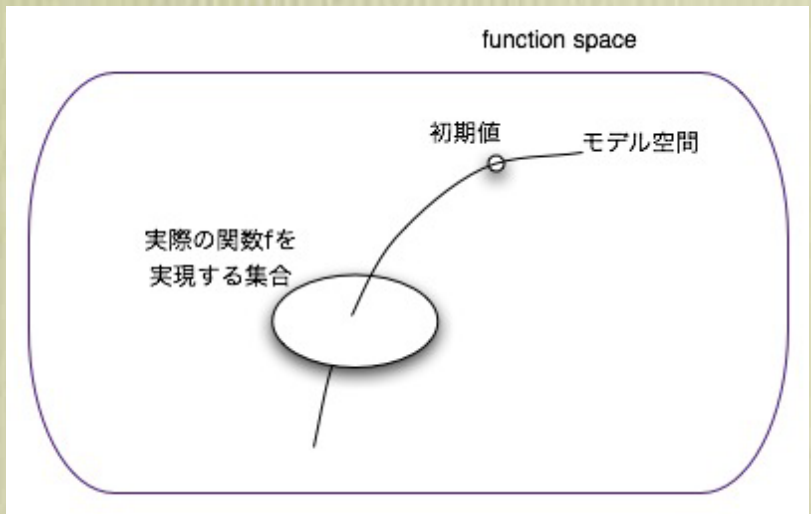
汎化性能に関して

SVMにおける汎化性能で論じられていることは，”教師データ以外に何の手がかりもなく”，”如何なる関数もありうる”場合である．

もし，データを発生するソースに関して何らかの情報が分かっている場合，特にその発生メカニズムなどが分かっている場合などは，モデル化により強い空間の制限がおこなえるはずである．

例：3次スプライン関数f(x)は，次式の意味で”最も滑らかな”曲面を実現する．

$$J(f) = \int |f''(x)|^2 dx$$



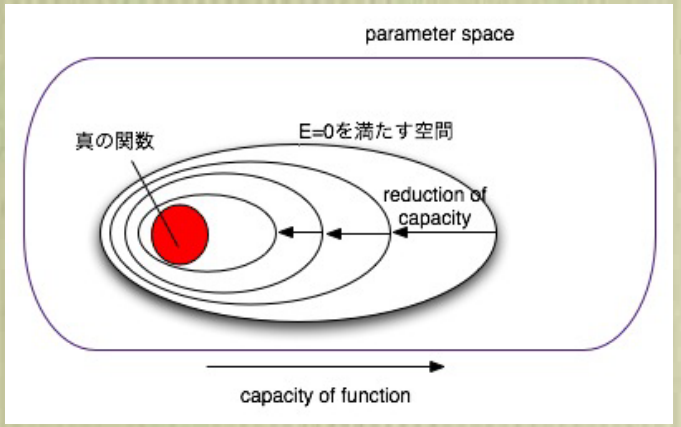
VC次元

m: VC次元 (VC: Vapnik and Chervonenkis)

与えられた関数族　の表現能力

例：2次元入力空間を定義域とする単純パーセプトロンでは，3つのデータに対して，全ての組み合わせ（8通り）にラベル付け可．

如何なる4つのデータに対しても，不可能　→　VC dim = 3



VC次元

個の点に関しては, 2　通りのラベル付け有. 与えられた判別機　が, 全ての組み合わせのラベル付け可能な最大の　の値をVC次元と定義する.

ちなみに, R^nにおける超平面族 (単純パーセプトロン) のVC dim=n+1

Structural risk minimization

与えられた学習法がRempの最小化のみに限定されているのなら唯一の方法.

SVM

- 線形分離可能なデータセットの分類
  - 線形判別器としての単純パーセプトロンの汎化能力に関して
  - 直感：マージンが最大となる判別面が最適

SVM

マージンとVC次元の関係

定理：ベクトル $x \in \mathcal{X}$ が半径Rの超球内に属しているとき，マージン の判別超平面集合のVC次元hは次式の不等式が成り立つ．

$$h \leq \min\left(\left[\frac{R^2}{\Delta^2}\right], n\right) + 1$$

→マージンが大きいほどVC次元は小さくなる．

マージン最大化

全ての教師データに関して次式が成立するwで，||w||が最小となるものを探索する．

minimize $\|\omega\|$ under constraints of $y_i(\boldsymbol{x}_i \cdot \boldsymbol{\omega} + b) - 1 \geq 0 \quad \forall i$

(← $\boldsymbol{x}_i \cdot \boldsymbol{\omega} + b \geq +1$ for $y_i = +1$
$\boldsymbol{x}_i \cdot \boldsymbol{\omega} + b \leq -1$ for $y_i = -1$)

Kuhn-Tucher Theorem

Optimization problem :

minimize $f_0(x)$

under constraints of (1) $f_k(x) \leq 0, \quad k = 1, \ldots, m$ and (2) where

A is a convex domain of $f_k(x), k = 0, \ldots, m$

Kuhn-Tucher Theorem

Theorem : if x* minimizes function $f_0(x)$ under constraints of (1) and (2), then there exist Lagrange multipliers that are simultaneously not equal to zero and such that the following three conditions hold true:

The minimum principle : $\min_{x \in A} L(x, \lambda_0^*, \lambda^*) = L(x^*, \lambda_0^*, \lambda^*)$

The nonnegativeness conditions: $\lambda_k^* \geq 0, \quad k = 0, \ldots, m$

The Kuhn-Tucker conditions: $\lambda_k^* f_k(x^*) = 0$

Langrange function $L(x, \lambda_0, \lambda) = \sum_{k=0}^{m} \lambda_k f_k(x)$

系

Kuhn-Tucher Theremにおける Lagrange関数を次式(3)に置き換えることが出来る場合，(4)が成立する．

Kuhn-Tucher Theorem

$$L(x, 1, \lambda) = f_0(x) + \sum_{k=1}^{m} \lambda_k f_k(x) \qquad (3)$$

$$\min_{x \in A} L(x, 1, \lambda^*) = L(x, 1, \lambda^*) = \max_{\lambda > 0} L(x^*, 1, \lambda) \qquad (4)$$

最小化のための関数

$$L = \lambda_0 \|\boldsymbol{\omega}\|^2 + \sum_{k=1}^{m} \lambda_k (-y_i(\boldsymbol{x}_i \cdot \boldsymbol{\omega} + b) + 1)$$

$$L_P = \frac{1}{2}\|\boldsymbol{\omega}\|^2 - \sum_{k=1}^{m} \alpha_k y_k (\boldsymbol{x}_k \cdot \boldsymbol{\omega} + b) + \sum_{k=1}^{m} \alpha_k \qquad (5)$$

演習

式(5) (式(4)の左辺に対応) を，ω, b を含まない式(式(4)の右辺に対応) に変換せよ.

$$\frac{\partial}{\partial w} L_P = \frac{\partial}{\partial b} L_P = 0 \qquad (6)$$

$$\min_{x \in A} L(x, 1, \lambda^*) = L(x, 1, \lambda^*) = \max_{\lambda > 0} L(x^*, 1, \lambda) \qquad (4)$$

最大のマージンを持つ判別面の導出

$$\alpha_i \geq, \quad i = 1, \ldots, m \tag{8}$$

$$W(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2} \sum_{k=1}^{m} y_i y_j \alpha_i \alpha_j x_i \cdot x_j \tag{9}$$

式(8)で示される制約条件の下で，式(9)を最大化するパラメータαの決定．

α^0

ω

決定された最適パラメータ　を用いて次式に基づき，パラメータ　を決定．

$$\omega = \sum_{k=1}^{m} \alpha_k^0 y_k x_k$$