PRML 5.5.6~5.6

5501 酒井一徳
3/22/18

目次

5.5.6 たたみ込みニューラルネットワーク

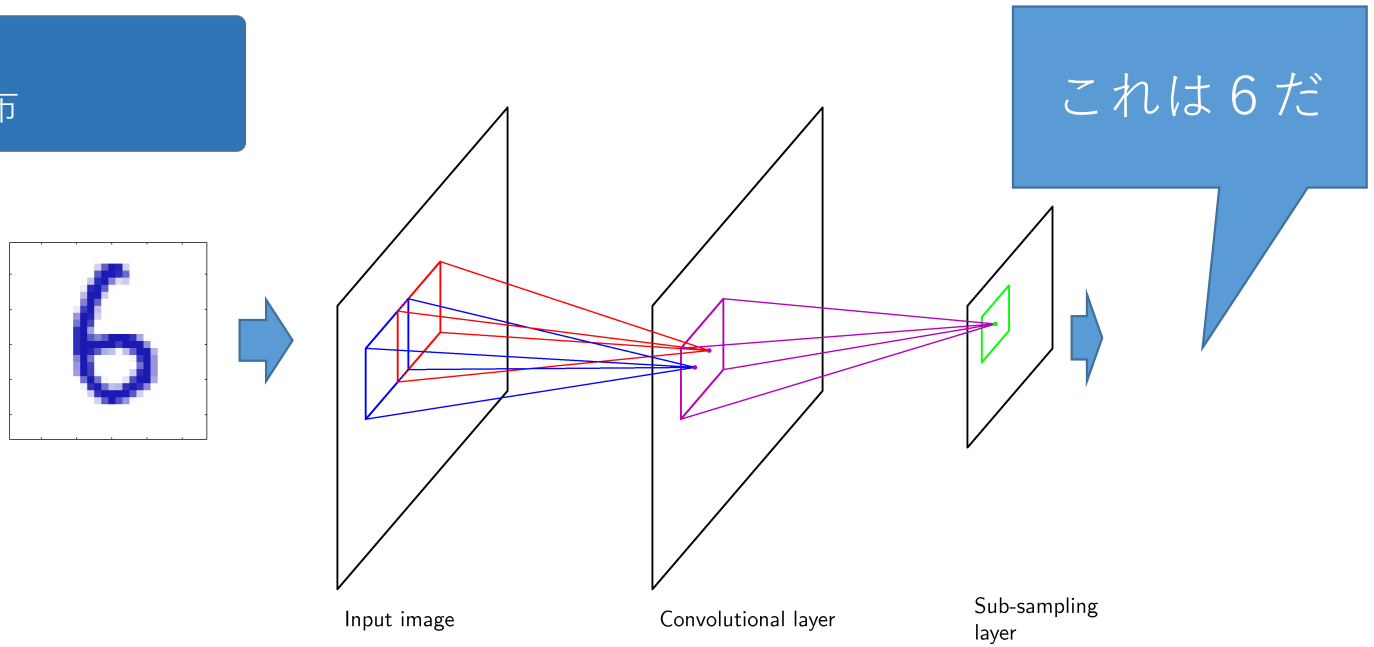
たたみ込みニューラルネットワーク

CNN： Convolutional Neural Network

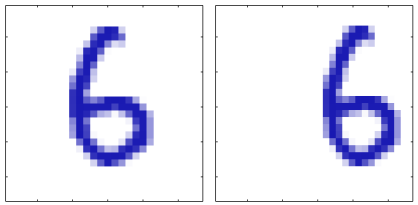
畳み込み層と**プーリング層**，全結合層などから構成されるニューラルネットワーク．

タスク例: 手書き文字認識

入力: 画像の画素の輝度値の集合
出力: 10個の数字のクラスの事後分布



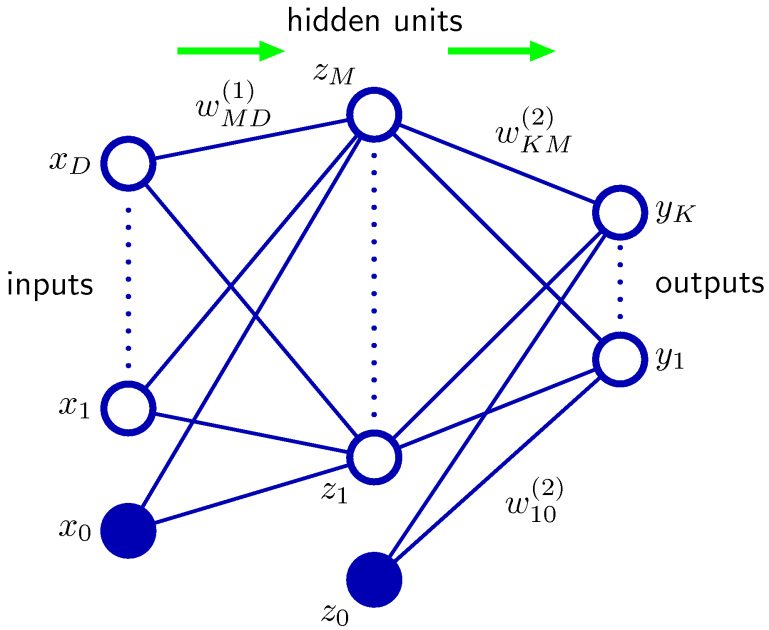
平行移動や拡大縮小，(小さい)
回転などにも
不変であるべき



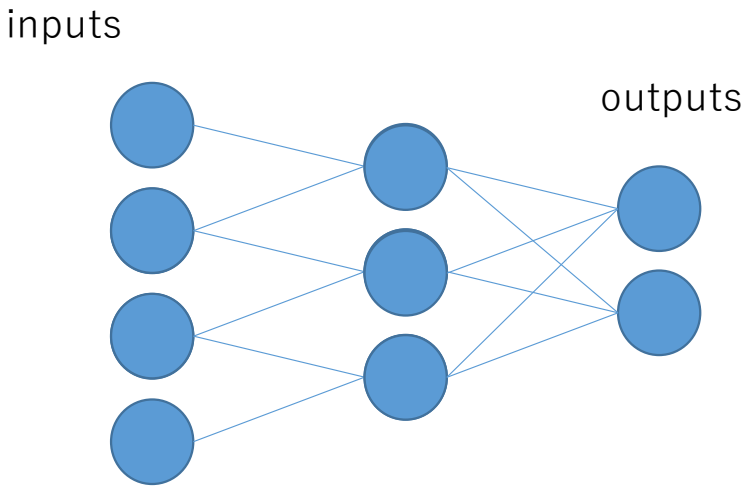
どちらも6だと認識してほしい

CNN

完全結合ネットワーク
入力同士の相関性を無視



入力同士が相関性を持つネットワーク
画像であれば，隣り合う画素同士は相関性が高い等



入力について相関性などの知見があるなら
ネットワークをあらかじめ変形させよう

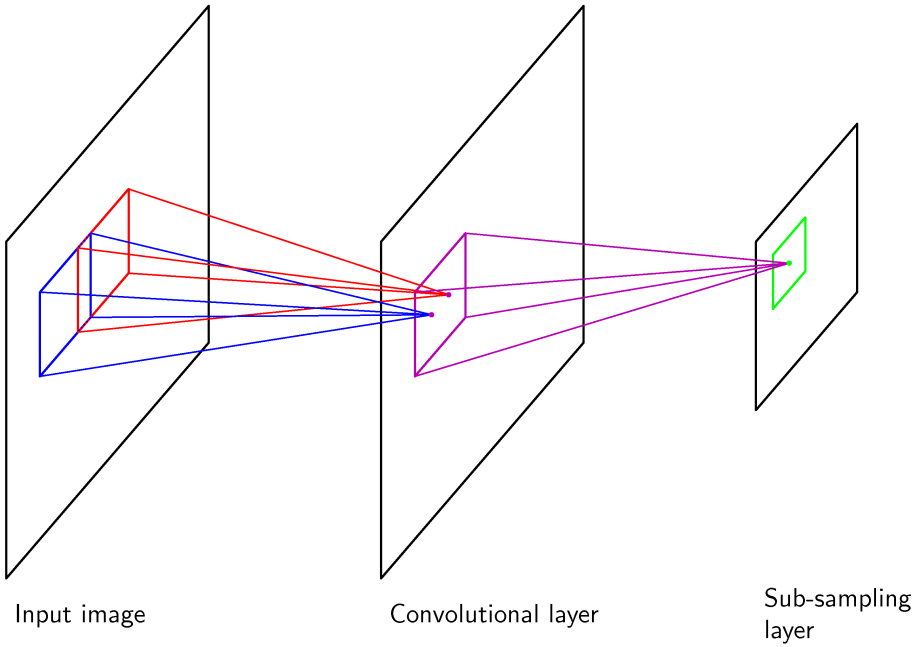
CNNの３つの機構

- 局所的な特徴の抽出
- その特徴の情報を統合することによりさらに高次の特徴の検出

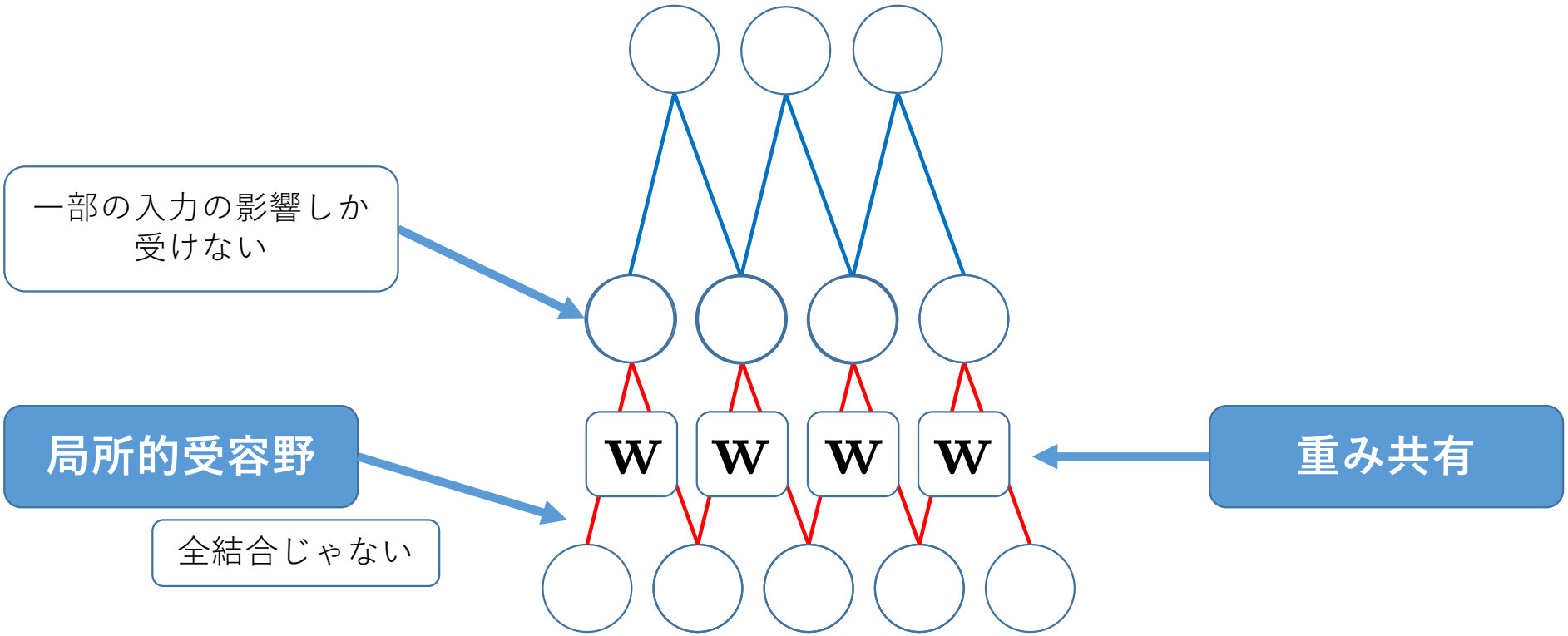
局所的受容野

重み共有

部分サンプリング



局所受容野と重み共有(畳み込み層)



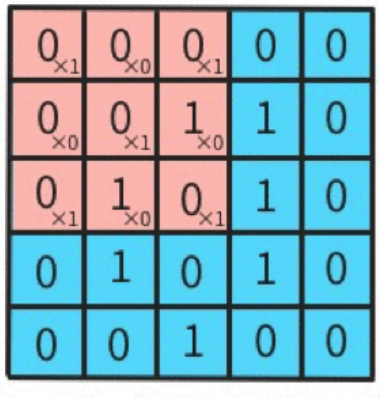
死ぬほど分かりやすい畳み込み演算の様子

0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	0	1	0
0	0	1	0	0

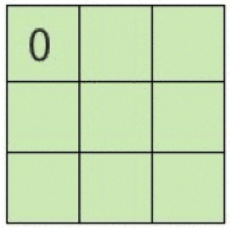
入力画像の画素表記
(簡略のため 2 値)

1	0	1
0	1	0
1	0	1

フィルター(重み)
(簡略のためバイアス 0 を想定)



入力画像



特徴マップ
(たたみ込み層)

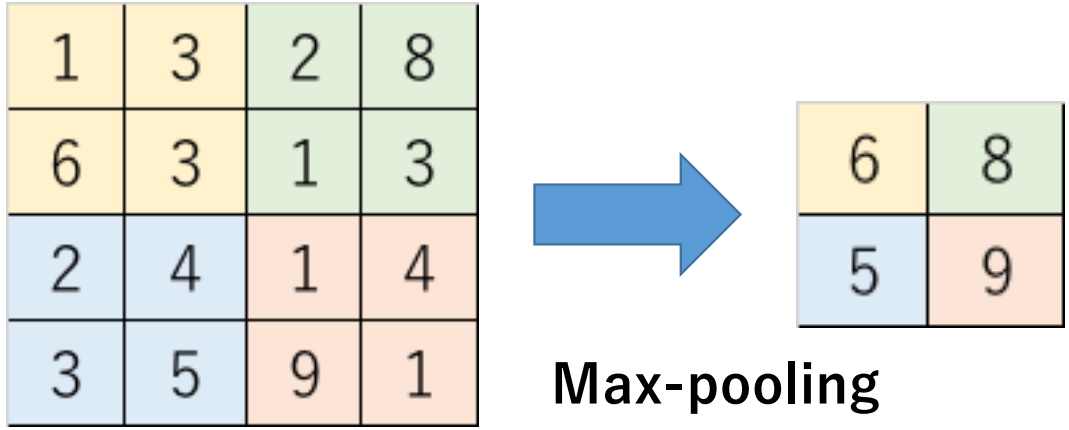
入力画像のあらゆる位置で同じパターン(特徴)を抽出している

入力画像の平行移動は特徴マップの平行移動として現れる

複数の特徴に対応して複数の特徴マップ，重みとバイアスが存在する

部分サンプリング(プーリング操作)

畳み込み層の出力(特徴マップ)

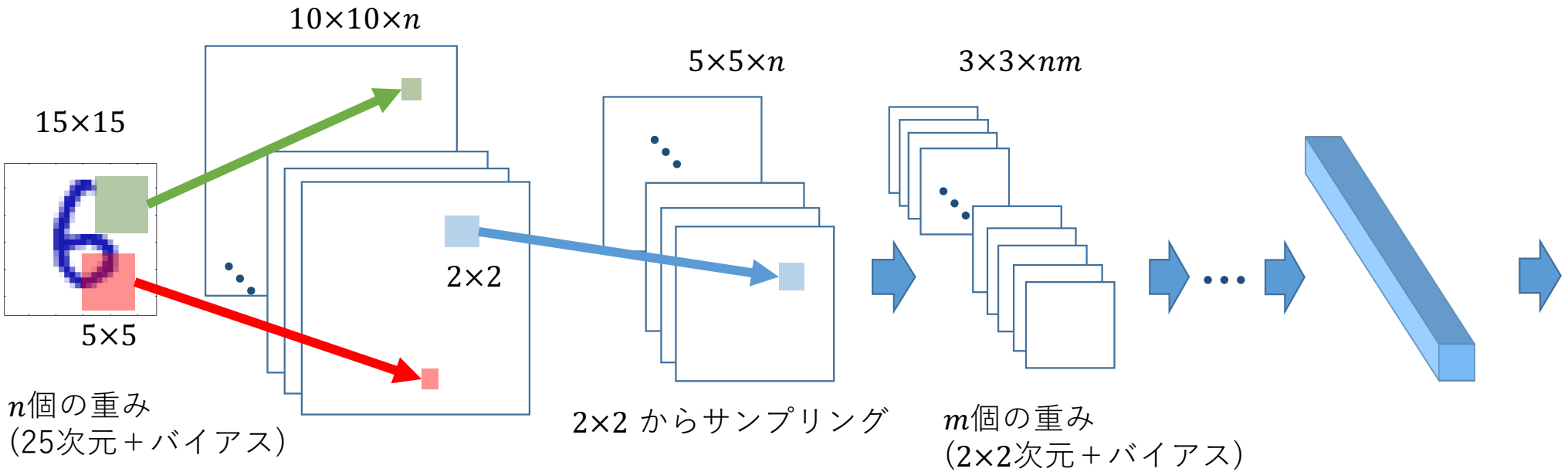


多少の数値の入れ替わりがあっても影響がない

先ほどの畳み込みと合わせて平行移動や歪みに対して頑健になる

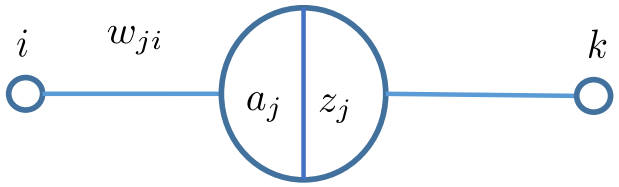
特徴同士がどのような関係であるかが希薄になる問題もある。(⇒カプセルネットワーク)

CNNの概観



層を重ねるごとに空間解像度は減るが特徴数を増やすことで補償する
層を重ねるごとに不変性は強固になる
Conv層とpooling層を合わせて畳み込み層と呼ぶ場合もある
最後に集めた特徴を集約する，多クラス分類問題ならソフトマックス関数など

逆伝播法による誤差最小化(演習5.28)

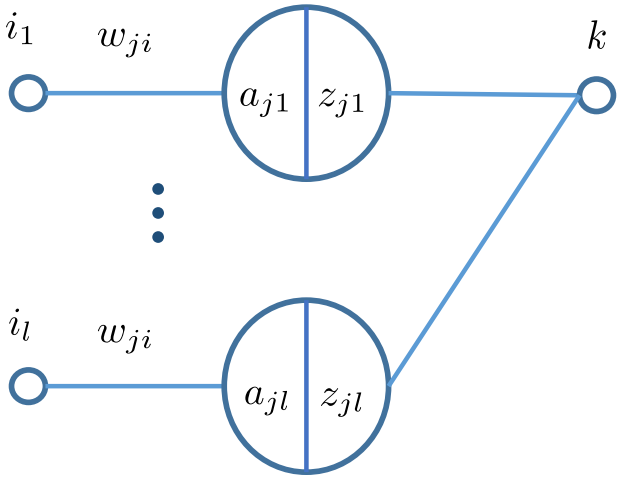


順伝播

$$a_j = \sum_i w_{ji} z_i$$
$$z_j = h(a_j)$$

逆伝播

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = \delta_j z_j$$
$$\delta_j \equiv \frac{\partial E}{\partial a_j} = \sum_k \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial a_j}$$



順伝播

$$a_{jl} = \sum_i w_{ji} z_{il}$$
$$z_{jl} = h(a_{jl})$$

逆伝播

$$\frac{\partial E}{\partial w_{ji}} = \sum_l \frac{\partial E}{\partial a_{jl}}\frac{\partial a_{jl}}{\partial w_{ji}} = \sum_l \delta_j z_j$$
$$\delta_j \equiv \frac{\partial E}{\partial a_{jl}} = \sum_k \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial a_{jl}}$$

5.5.7 ソフト重み共有

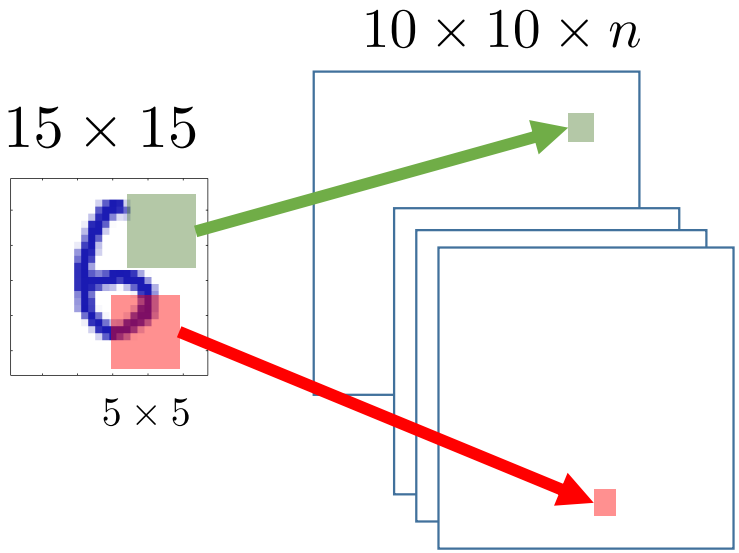
ソフト重み共有

重み共有はネットワークの複雑さを削減する有効なテクニック

しかし，制限の形が特定される特別な問題にしか適用できない

ソフト重み共有

重みが同じグループで似た値を取りやすくなるように正則化項の導入



つまり，緑のグループや赤のグループごとに
重みの平均や分散を設定しよう
さらにそれらを学習の過程で得よう

ソフト重み共有ー荷重減衰の拡張

単純な荷重減衰正則化項(5.112)が, 重みのガウス事前分布の負の対数尤度とみなせる(5.5節，3.1.1〜3.1.4節)

つまり，このときの重みはガウス分布に従う一つのグループであるとみなせる

複数のグループを構成したい

混合ガウス分布を代わりに用いよう(2.3.9)

確率密度関数: $p(\mathbf{w}) = \prod_i p(w_i)$

ただし, $p(w_i) = \sum_{j=1}^{M} \pi_j \mathcal{N}\left(w_i | \mu_j, \sigma_j^2\right)$

負の対数尤度を取る

$$\Omega(\mathbf{w}) = -\sum_i \ln\left(\sum_{j=1}^{M} \pi_j \mathcal{N}\left(w_i | \mu_j, \sigma_j^2\right)\right)$$

誤差関数: $\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \Omega(\mathbf{w})$

ソフト重み共有-誤差関数の最小化

$$\tilde{E}(w) = E(\mathbf{w}) + \Omega(\mathbf{w}) = E(\mathbf{w}) - \sum_i \ln\left(\sum_{j=1}^{M} \pi_j \mathcal{N}\left(w_i|\mu_j, \sigma_j^2\right)\right)$$

調整可能なパラメータ
中心: μ_j
分散: σ_j
混合係数: π_j

パラメータである平均，分散，混合係数についても最小化し，決定する

- 重みが定数なら，EMアルゴリズム
- 重みと同時最適化が必要な場合，共役勾配法，準ニュートン法

2.3.9節に従って負担率の導入(2.192),
$\{\pi_j\}$を事前分布とみなすとベイズの定理から

事後分布: $\gamma_j(w) = \dfrac{\pi_j \mathcal{N}\left(w|\mu_j, \sigma_j^2\right)}{\sum_k \pi_k \mathcal{N}\left(w|\mu_k, \sigma_k^2\right)}$

誤差関数の最小化-重み, 中心について

重み(演習5.29)

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \sum_j \gamma_j(w_i)\frac{(w_i - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial \Omega}{\partial w_i} = \frac{\sum_j \pi_j \frac{(w_i-\mu_j)}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)} = \sum_j \gamma_j(w_i)\frac{(w_i-\mu_j)}{\sigma_j^2}$$

中心(演習5.30)

$$\frac{\partial \tilde{E}}{\partial \mu_j} = \sum_i \gamma_j(w_i)\frac{(\mu_j - w_i)}{\sigma_j^2}$$

$$\frac{\partial \Omega}{\partial \mu_j} = -\sum_i \frac{\pi_j \frac{(w_i-\mu_j)}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)} = \sum_i \gamma_j(w_i)\frac{(\mu_j-w_i)}{\sigma_j^2}$$

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \Omega(\mathbf{w})$$

$$\Omega(\mathbf{w}) = -\sum_i \ln\left(\sum_{j=1}^{M}\pi_j \mathcal{N}\left(w_i|\mu_j,\sigma_j^2\right)\right)$$

$$\frac{\partial \mathcal{N}(x|\mu,\sigma^2)}{\partial x} = -\frac{(x-\mu)}{\sigma^2}\mathcal{N}\left(x|\mu,\sigma^2\right)$$

$$\frac{\partial \mathcal{N}(x|\mu,\sigma^2)}{\partial \mu} = \frac{(x-\mu)}{\sigma^2}\mathcal{N}\left(x|\mu,\sigma^2\right)$$

事後分布: $\gamma_j(w) = \frac{\pi_j \mathcal{N}\left(w|\mu_j,\sigma_j^2\right)}{\sum_k \pi_k \mathcal{N}\left(w|\mu_k,\sigma_k^2\right)}$

誤差関数の最小化-分散について

分散(演習5.31)

$$\frac{\partial \tilde{E}}{\partial \sigma_j} = \sum_i \gamma_j(w_i) \left\{ \frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\}$$

$$\frac{\partial \tilde{E}}{\partial \sigma_j} = -\sum_i \left\{ \frac{\pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)} \left\{ -\frac{1}{\sigma_j} + \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\} \right\}$$

$$= \sum_i \gamma_j(w_i) \left\{ \frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\}$$

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \Omega(\mathbf{w})$$

$$\Omega(\mathbf{w}) = -\sum_i \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}\left(w_i|\mu_j, \sigma_j^2\right) \right)$$

$$\frac{\partial \mathcal{N}(x|\mu, \sigma^2)}{\partial \sigma} = \left\{ -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3} \right\} \mathcal{N}(x|\mu, \sigma^2)$$

事後分布: $\gamma_j(w) = \dfrac{\pi_j \mathcal{N}\left(w|\mu_j, \sigma_j^2\right)}{\sum_k \pi_k \mathcal{N}\left(w|\mu_k, \sigma_k^2\right)}$

混合ガウスの対数尤度の注意

実際の実装では　$\sigma_j^2 = \exp(\xi_j)$　と定義される変数ξ_j を導入し，この変数について最小化を行う

なぜか

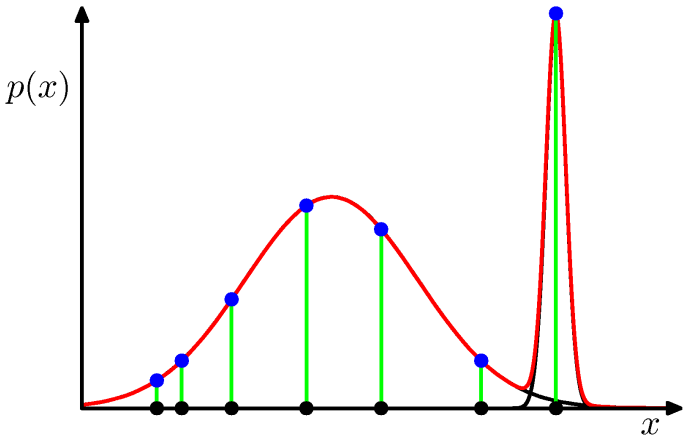
対数尤度：　$\Omega(\mathbf{w}) = -\sum_i \ln\left(\sum_{j=1}^{M} \pi_j \mathcal{N}\left(w_i|\mu_j, \sigma_j^2\right)\right)$
（正則化項）

- パラメータσ_jの正値性の保証
- σ_jがゼロに近づくことで，ガウス要素が重みのパラメータの値の一つに収縮するという病的な解を避けるため

例えば…　あるn について $\mu_j = w_n$となる重みが存在した時，その重みは対数尤度に対して

$$\mathcal{N}(w_n|w_n, \sigma_j^2) = \frac{1}{(2\pi\sigma_j^2)^{\frac{1}{2}}}$$

の形で寄与する，この項は$\sigma_j \to 0$ の極限を考えると無限大に発散する．
つまり，対数尤度も無限大に発散する．
混合要素の分布が特定の点に収束すれば対数尤度は常に増大する方向へ働いてしまう．



誤差関数の最小化-混合係数について

復習(2.3.9節-混合ガウス分布)

前述の通り，$\{\pi_j\}$を事前分布として解釈するため
制約を考慮し，右記の補助変数$\{\eta_j\}$を導入

制約

$$\sum_j \pi_j = 1 \qquad 0 \leq \pi_j \leq 1$$

$$\pi_j = \frac{\exp(\eta_j)}{\sum_k \exp(\eta_k)}$$

$\{\eta_j\}$に関する微分(演習5.32)　$\dfrac{\partial \tilde{E}}{\partial \eta_j} = \sum_i \{\pi_j - \gamma_j(w_i)\}$

$$\frac{\partial \Omega}{\partial \pi_j} = -\sum_i \left\{ \frac{\mathcal{N}(w_i|\mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)} \right\} = \sum_i \frac{\gamma_j(w_i)}{\pi_j}$$

$$\frac{\partial \pi_k}{\partial \eta_j} = \frac{\delta_{kj} \exp(\eta_k) \sum_l \exp(\eta_l) - \exp(\eta_k)\exp(\eta_j)}{\{\sum_l \exp(\eta_l)\}^2} = \delta_{kj}\pi_j - \pi_k\pi_j$$

$$\frac{\partial \Omega}{\partial \eta_j} = \sum_k \frac{\partial \Omega}{\partial \pi_k}\frac{\partial \pi_k}{\partial \eta_j} = \sum_k \left\{ (\delta_{kj}\pi_j - \pi_k\pi_j) \left(\sum_i \frac{\gamma_j(w_i)}{\pi_k} \right) \right\} = \sum_i \{\pi_j - \gamma_j(w_i)\}$$

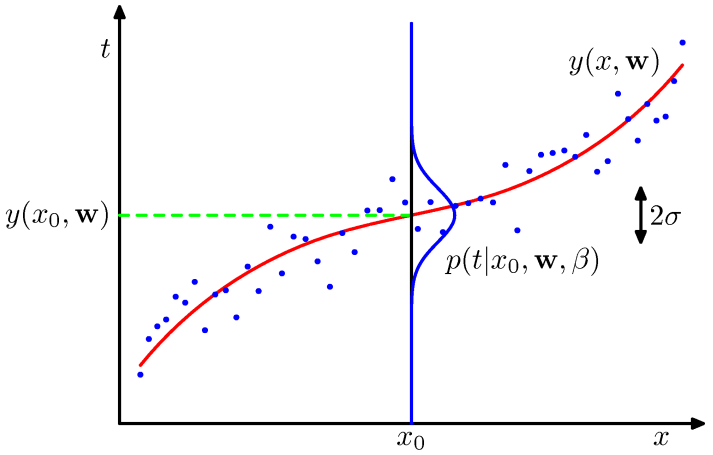
5.6 混合密度ネットワーク

混合密度ネットワーク

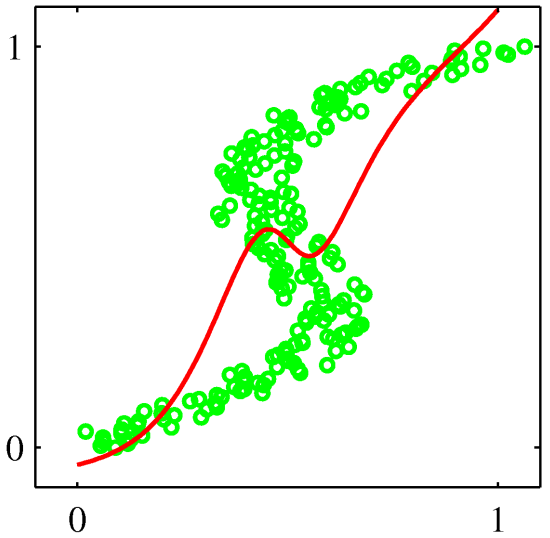
教師あり学習の目標: $p(\mathbf{t}|\mathbf{x})$のモデル化

単純な回帰問題ではガウス分布を仮定することが多い

例えば，PRML 1 章の二乗和誤差の最小化が最尤推定と等価であった話もノイズがガウス分布に従う仮定があった

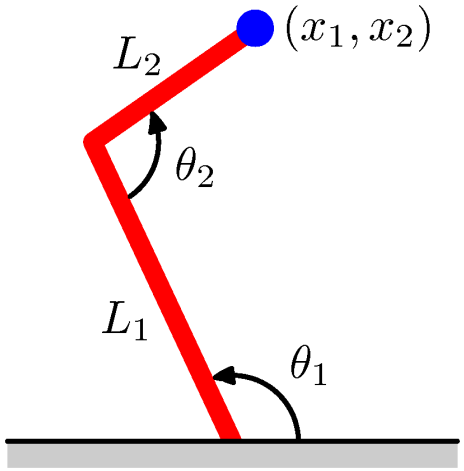


もし，分布が多峰性を持つならば？

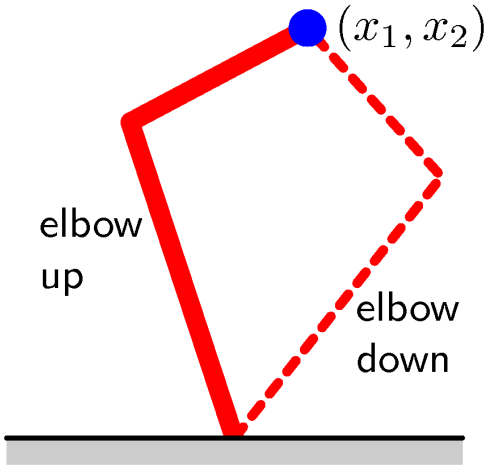


順問題と逆問題

順問題	逆問題
関節角から終端位置	終端位置から関節角
特定の疾患から特定の症状	ある症状の集合から疾患



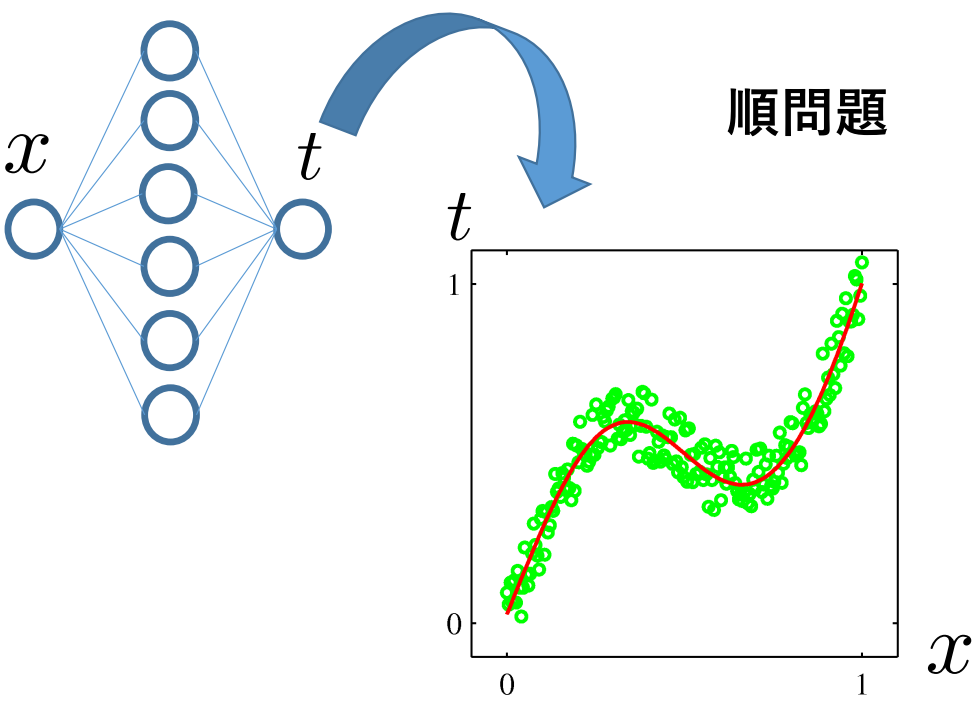
高次元の機械学習では多峰性の存在は明白ではない

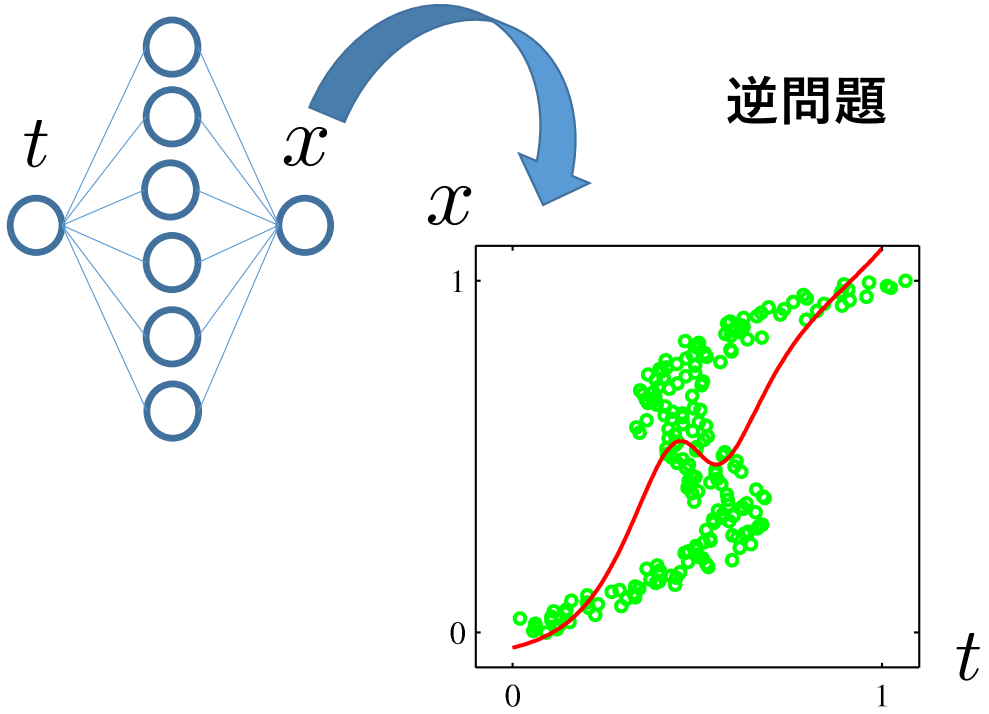


順問題と逆問題

$\{x_n\}$：一様分布 $\mathrm{U}(0,1)$ に従う確率変数$\boldsymbol{x}$からサンプリングしたもの

$\{t_n\}$：関数 $x_n + 0.3\sin(2\pi x_n)$ に $\mathrm{U}(-0.1, 0.1)$ から生成される乱数を加えたもの





混合密度ネットワーク

任意の条件付き密度関数をモデル化するための枠組み
もしガウス分布を要素に持つならば,

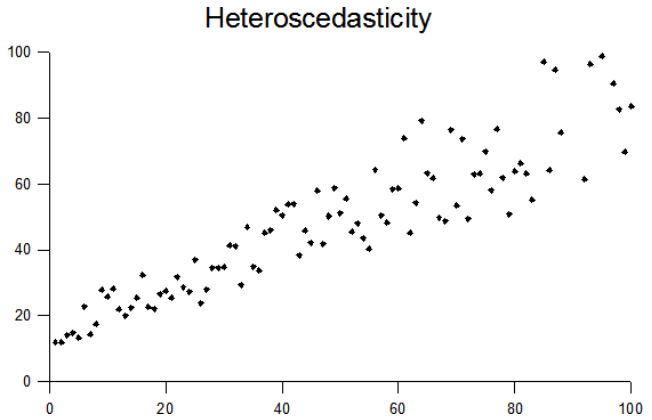
$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I})$$

異分散(heteroscedastic)のモデルの一例

データに対するノイズの分散が$\mathbf{x}$の関数になっている

等方性共分散を持つことを仮定しない
Tの要素に分解できることを仮定しない

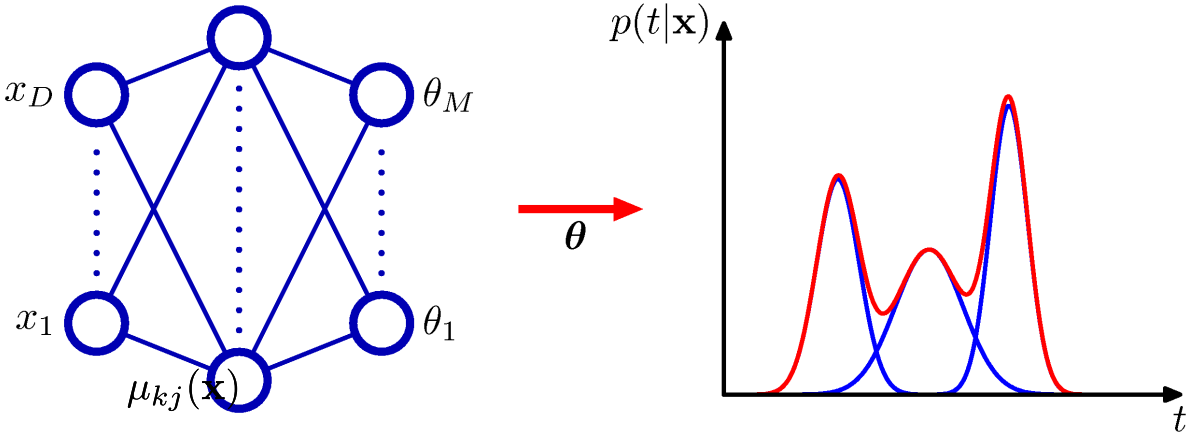
Wikiより



混合密度ネットワーク

パラメータをNNの出力に
支配されるようにとる.

出力するパラメータ
平均: $\mu_k(\mathbf{x})$
分散: $\sigma_k^2(\mathbf{x})$
混合係数: $\pi_k(\mathbf{x})$



例えば, $p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I})$ のモデルは要素がK個であり，$\mathbf{t}$の要素がL個ならば

出力	決定するもの	個数
a_k^{π}	混合係数 $\pi_k(\mathbf{x})$	K
a_k^{σ}	カーネルの幅 $\sigma_k^2(\mathbf{x})$	K
a_{kj}^{μ}	カーネルの中心 $\boldsymbol{\mu}_k(\mathbf{x})$ の要素 $\mu_{kj}(\mathbf{x})$	K×L

図5.19の例であれば
K=3, L=1でいい感じであることが
後に書かれている
つまり出力の総数は9個となる

混合密度ネットワーク

パラメータについての注意
ソフト重み共有の時と同じように，混合係数は制約を満たすために

$$\pi_k(\mathbf{x}) = \frac{\exp(a_k^\pi)}{\sum_{l=1}^{K} \exp(a_l^\pi)}$$

制約

$$\sum_j \pi_j = 1 \quad 0 \leq \pi_j \leq 1$$

上記のソフトマックス関数を導入．同じく分散も制約から，

$$\sigma_k^2(\mathbf{x}) = \exp(a_k^\sigma)$$

制約

$$\sigma_k^2(\mathbf{x}) \geq 0$$

を導入する．平均は実数の値を取るので，

$$\mu_{kj}(\mathbf{x}) = a_{kj}^\mu$$

最適化

ニューラルネットの重みとバイアスであるベクトル$\mathbf{w}$を誤差関数の最小化によって得る.

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k(\mathbf{x}_n, \mathbf{w}) \mathcal{N}(\mathbf{t}_n | \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w}) \mathbf{I}) \right\}$$

ただし，データが独立である．(2.193)の負の対数尤度.

各出力ユニットの出力の誤差の微分が与えられるならば，通常の逆伝播で評価できる.

各訓練データ点についての和で構成されている.

事後分布: $\gamma_{nk}(\mathbf{t}_n|\mathbf{x}_n) = \dfrac{\pi_k \mathcal{N}_{nk}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}}$　ただし，$\mathcal{N}_{nk} = \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n), \sigma_k^2(\mathbf{x}_n))$

各出力に関する微分

混合係数について
(演習5.34)

$$\frac{\partial E_n}{\partial a_k^{\pi}} = \pi_k - \gamma_{nk}$$

平均について
(演習5.35)

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = \gamma_{nk}\left\{\frac{\mu_{kl} - t_{nl}}{\sigma_k^2}\right\}$$

分散について
(演習5.36)

$$\frac{\partial E_n}{\partial a_k^{\sigma}} = \gamma_{nk}\left(L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right)$$

分散のみ後述

分散による微分(演習5.36)

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \sigma_k}\frac{\partial \sigma_k}{\partial a_k^\sigma}$$

$$\frac{\partial \sigma_k}{\partial a_k^\sigma} = \exp(a_k^\sigma) = \sigma_k$$

$$\begin{aligned}\frac{\partial \mathcal{N}_{nk}}{\partial \sigma_k} &= -\frac{L}{\sigma_k}\mathcal{N}_{nk} + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3}\mathcal{N}_{nk} \\ &= \frac{\mathcal{N}_{nk}}{\sigma_k}\left(-L + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial E_n}{\partial \sigma_k} &= -\frac{\pi_k \frac{\partial \mathcal{N}_{nk}}{\partial \sigma_k}}{\sum_j \pi_j \mathcal{N}_{nj}} \\ &= -\frac{\pi_k \frac{\mathcal{N}_{nk}}{\sigma_k}\left(-L + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right)}{\sum_l \pi_l \mathcal{N}_{nl}} \\ &= \frac{\gamma_{nk}}{\sigma_k}\left(L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right)\end{aligned}$$

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}_{nk}\right\}$$

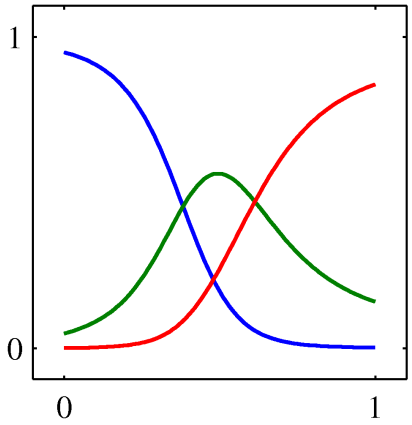
$$\mathcal{N}_{nk} = \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w})\mathbf{I})$$

$$\begin{aligned}\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k, \sigma_k^2\mathbf{I}) &= \frac{1}{(2\pi)^{\frac{L}{2}}}\frac{1}{|\sigma_k^2\mathbf{I}|^{\frac{1}{2}}}\exp\{\} \\ &= \frac{1}{(2\pi)^{\frac{L}{2}}}\frac{1}{\sigma_k^L}\exp\{\}\end{aligned}$$

$$\sigma_k(x) = \exp(a_k^\sigma)$$

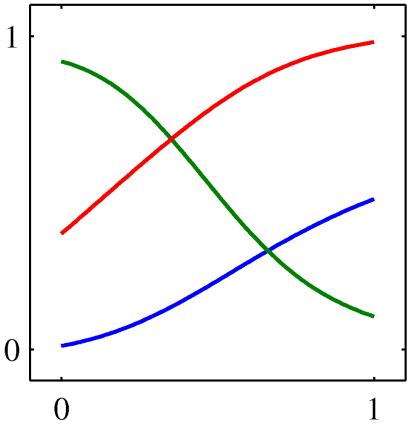
$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \sigma_k}\frac{\partial \sigma_k}{\partial a_k^\sigma} = \gamma_{nk}\left(L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right)$$

混合密度ネットワーク-図5.19の例



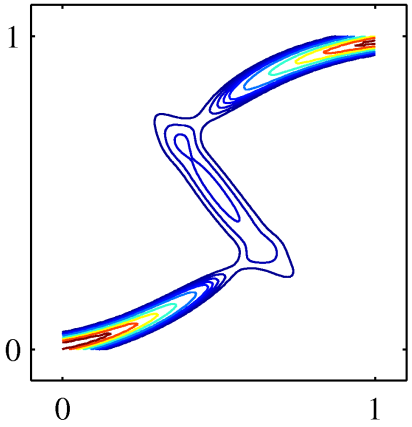
(a)

混合係数
どの要素が選ばれるかという
事前分布とみなせる



(b)

平均
各要素の分布の
xに応じた平均



(c)

条件付き密度分布
混合係数のグラフからも
xの大きさに応じて単峰か多峰か
変わることがわかる

今回のモデルはガウス分布を３つ持つ混合モデル
５個のシグモイド隠れユニットと
９個の出力からなる多層パーセプトロン

予測モデルを作ったら

いろんな有用な量を計算したい，例えば，

平均: $\mathbb{E}[\mathbf{t}|\mathbf{x}] = \int \mathbf{t}p(\mathbf{t}|\mathbf{x})d\mathbf{t} = \sum_{k=1}^{K} \pi_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})$

分散: $s^2(\mathbf{x}) = \mathbb{E}\left[\|\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 |\mathbf{x}\right]$

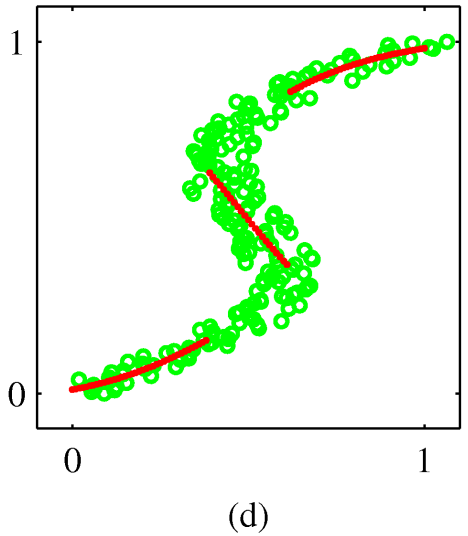
$$= \sum_{k=1}^{K} \pi_k(\mathbf{x}) \left\{ \sigma_k^2(\mathbf{x}) + \left\| \boldsymbol{\mu}_k(\mathbf{x}) - \sum_{l=1}^{K} \pi_l(\mathbf{x})\boldsymbol{\mu}_l(\mathbf{x}) \right\|^2 \right\}$$

条件付き平均が多峰性の分布では貧弱な表現になる(最小二乗と大差ない).
分散は最小二乗よりは一般的.

条件付き平均よりモードの方が重要

ただし解析解はなく数値的反復法が必要.

右図はxの値に対して混合係数の大きい(事前確率の高い)要素の平均を
プロットしたもの



(d)