オンライン学習（１５章）

杉山将・本多淳也
sugi@k.u-tokyo.ac.jp, jhonda@k.u-tokyo.ac.jp
http://www.ms.k.u-tokyo.ac.jp

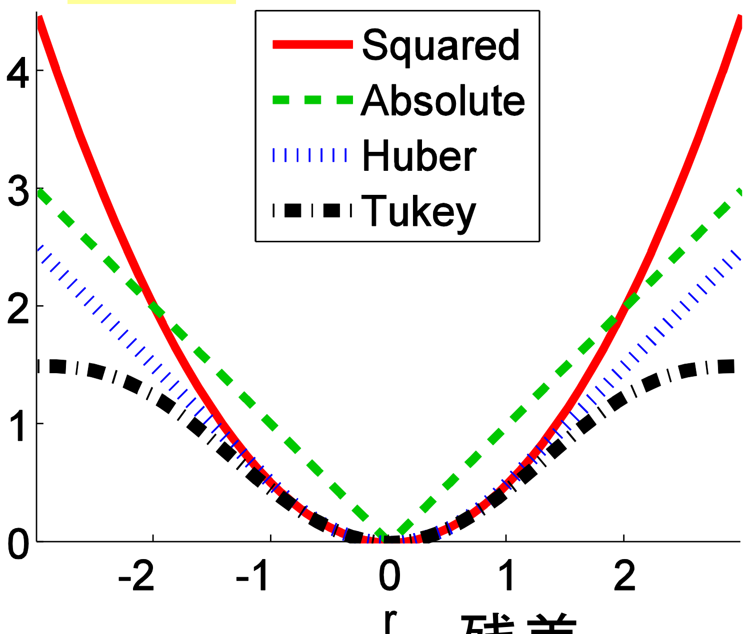
一括学習(バッチ学習)

- 全ての訓練標本を用いてパラメータを学習
- 標本が多いとき，一括で学習するのは困難

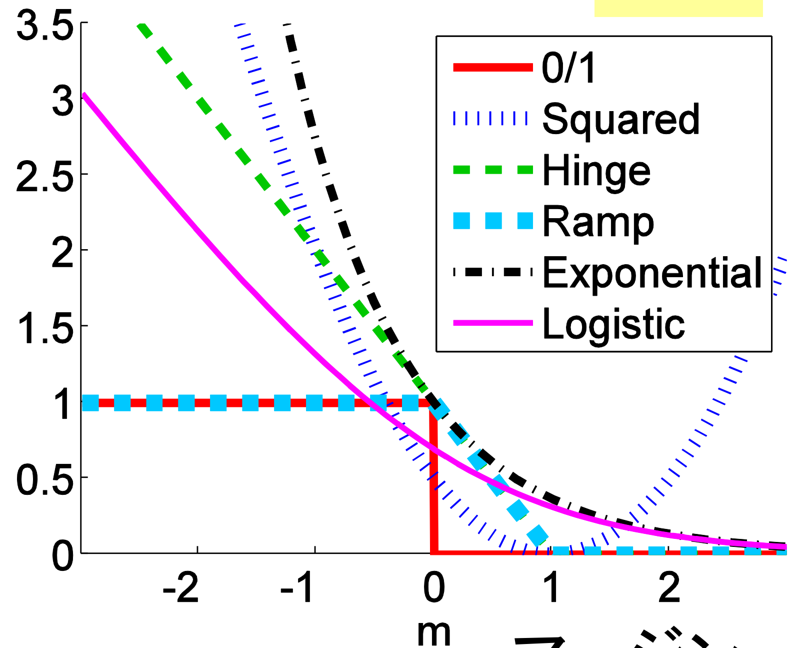
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathrm{loss}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i\Big)$$

$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$

回帰

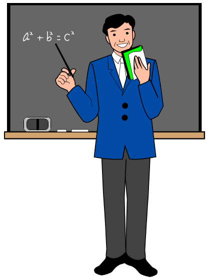


分類



オンライン学習

- データを一つ一つ（あるいはいくつかずつ）学習していく
- 一回一回の解の更新は非常に簡単に行える
- データを生成する確率分布が変化するような場面にも適応できる



講義の流れ

1. 確率的勾配法
2. PA学習
3. 適応正則化学習

確率的勾配法

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathrm{loss}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i\Big)$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$$

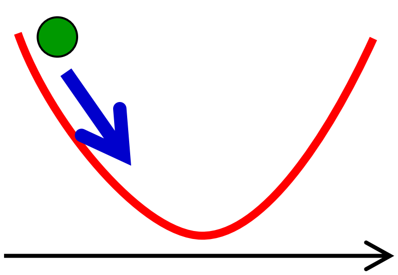
1. パラメータ $\boldsymbol{\theta}$ を適当に初期化
2. 標本 $(\boldsymbol{x}, y)$ をランダムに選び，勾配を少し降下

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} - \varepsilon \nabla \mathrm{loss}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y\Big)$$

ステップ幅

$\varepsilon > 0$

3. 収束するまで 2. を繰り返す



■実用上は，一つでなく複数の標本(ミニバッチ)を用いることが多い

実行例

- ガウスカーネルモデル

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}\|^2}{2h^2}\right)$$

に対する最小二乗法

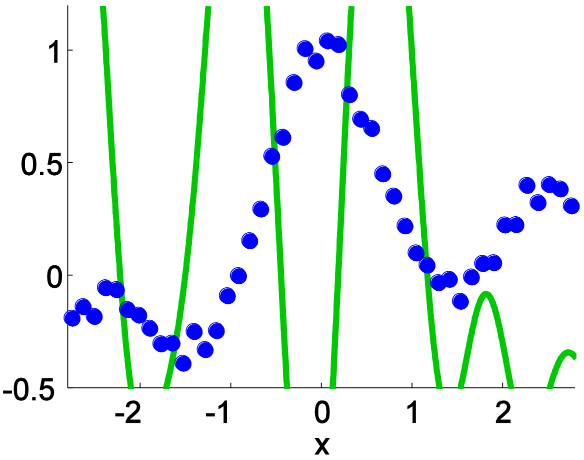
$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2$$

の確率的勾配法

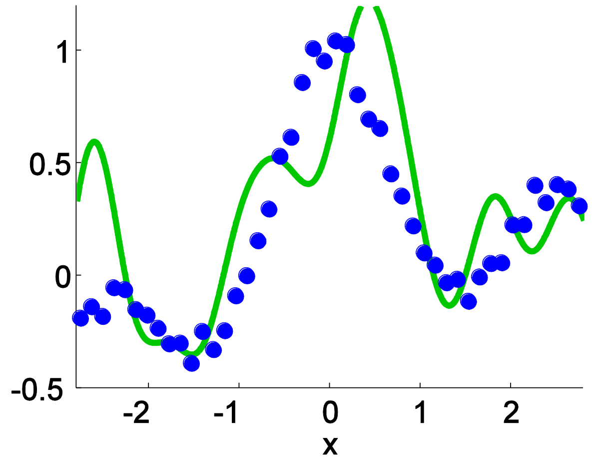
- 勾配は $\boldsymbol{k}\left(\boldsymbol{k}^\top \boldsymbol{\theta} - y\right)$ $\boldsymbol{k} = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n))^\top$

実行例(続き)

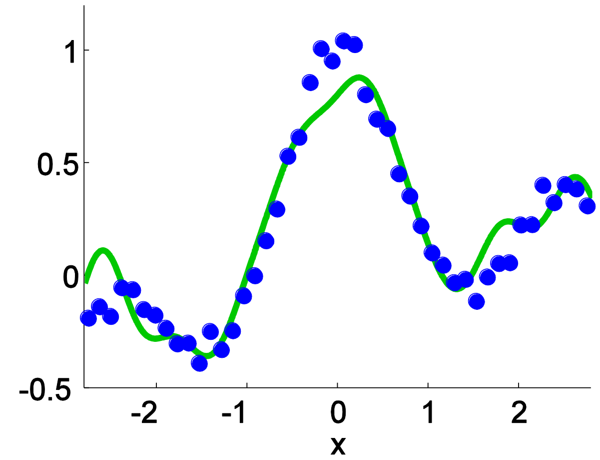
初期値



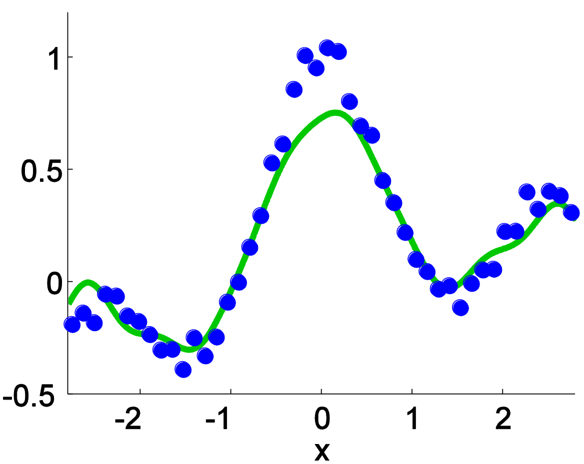
50反復後



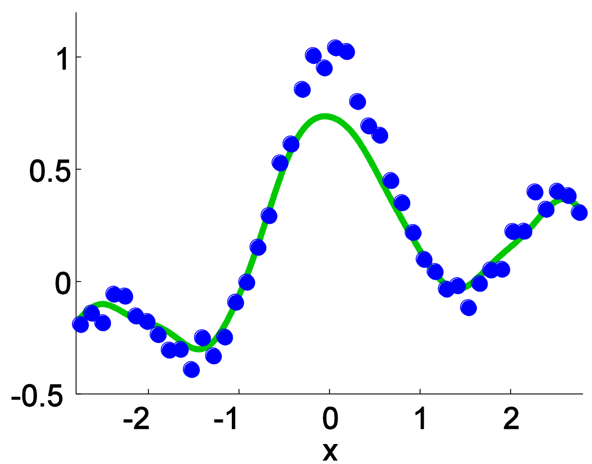
100反復後



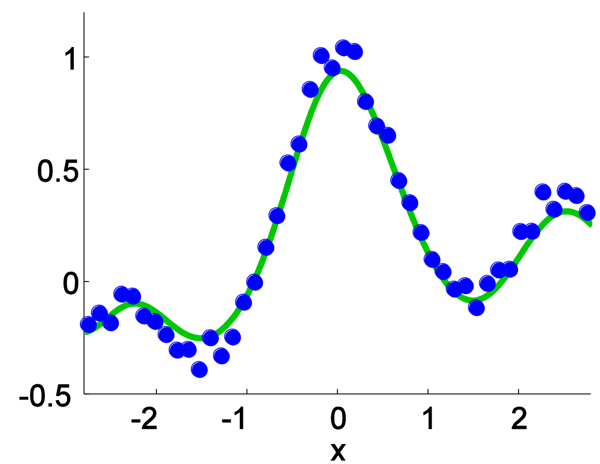
150反復後



200反復後



約5000反復後

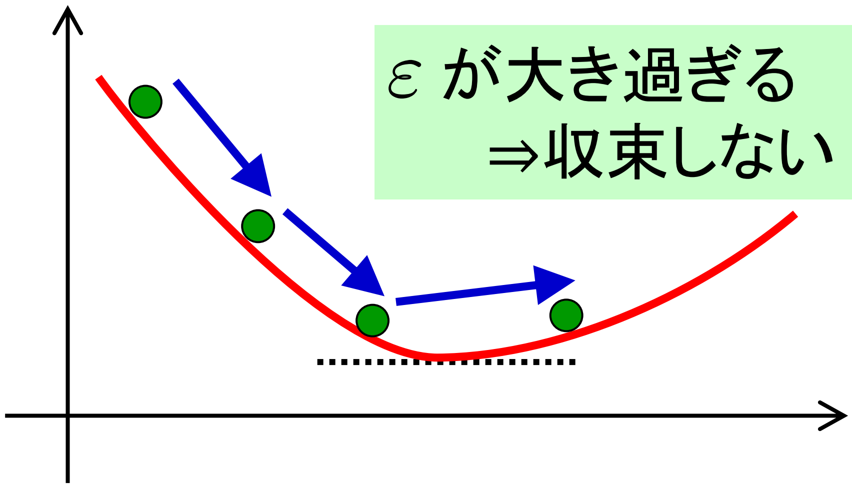


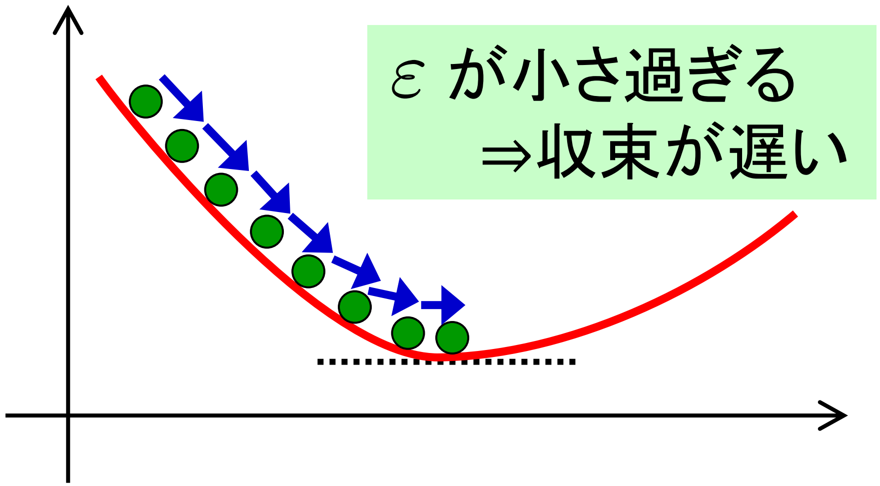
実装例

```
clear all; rand('state',0); randn('state',0);
n=50; N=1000; x=linspace(-3,3,n)';
X=linspace(-3,3,N)';
pix=pi*x; y=sin(pix)./(pix)+0.1*x+0.05*randn(n,1);

hh=2*0.3^2; t0=randn(n,1); e=0.1;
for o=1:n*100
  i=ceil(rand*n); ki=exp(-(x-x(i)).^2/hh);
  t=t0-e*ki*(ki'*t0-y(i));
  t0=t;
end
K=exp(-(repmat(X.^2,1,n)+repmat(x.^2',N,1)-2*X*x')/hh);
F=K*t;

figure(1); clf; hold on; axis([-2.8 2.8 -0.5 1.2]);
plot(X,F,'g-'); plot(x,y,'bo');
```

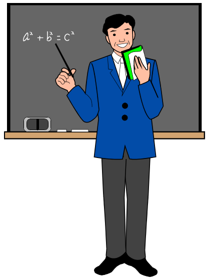
確率的勾配法の問題点

- ステップ幅 ε の設定が難しい





- 急激な学習を行うと，過去の学習で得られた知識が損なわれる恐れがある



講義の流れ

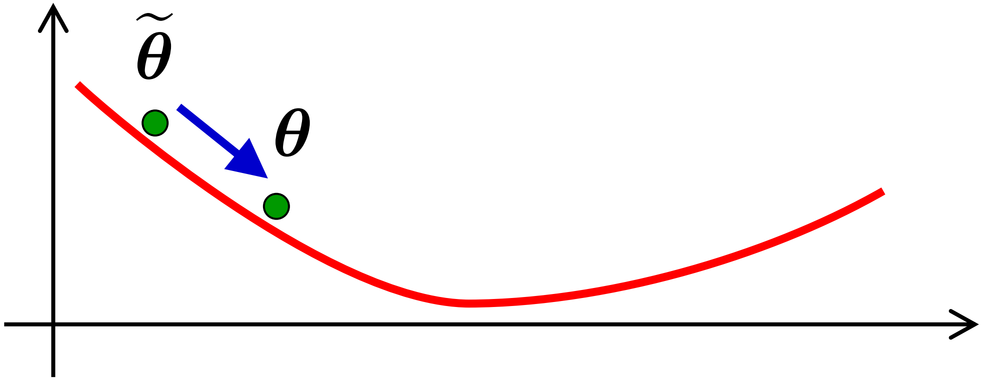
1. 確率的勾配法
2. PA学習
3. 適応正則化学習

PA (Passive-Aggressive) 学習

- 現在の解 $\widetilde{\boldsymbol{\theta}}$ からの変化量を抑制する

$$\min_{\boldsymbol{\theta}} \underbrace{\mathrm{loss}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y\Big)}_{\text{訓練出力に対する適合の良さ}} + \gamma \underbrace{\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|^2}_{\text{パラメータの変化量を抑制}}$$

受動係数 $\gamma > 0$



基底関数 $\phi(\boldsymbol{x})$

- 以下，線形モデルを考える: $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x})$

確率的勾配法との比較

- 二乗損失に対するPA回帰の解の更新式:

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y}{\|\boldsymbol{\phi}(\boldsymbol{x})\|^2 + \gamma} \boldsymbol{\phi}(\boldsymbol{x})$$

$\gamma > 0$

導出は後述

$$\mathrm{loss}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y\Big) = \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y\Big)^2$$

- 確率的勾配法の解の更新式:

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} - \varepsilon \boldsymbol{\phi}(\boldsymbol{x}) \Big(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y\Big)$$

- PA回帰では，ステップ幅をデータ $\boldsymbol{x}$ に合わせて適応的に決定:

$$\varepsilon = \frac{1}{\|\boldsymbol{\phi}(\boldsymbol{x})\|^2 + \gamma}$$

二乗ヒンジ損失

■ 二乗ヒンジ損失に対するPA分類

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

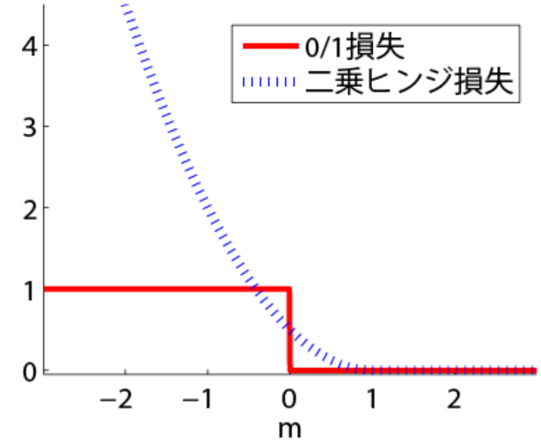
$\gamma > 0$

$$J(\boldsymbol{\theta}) = \frac{1}{2}\left(\max\left(0, 1 - \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) y\right)\right)^2 + \frac{\gamma}{2}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|^2$$

の解は次式で与えられる

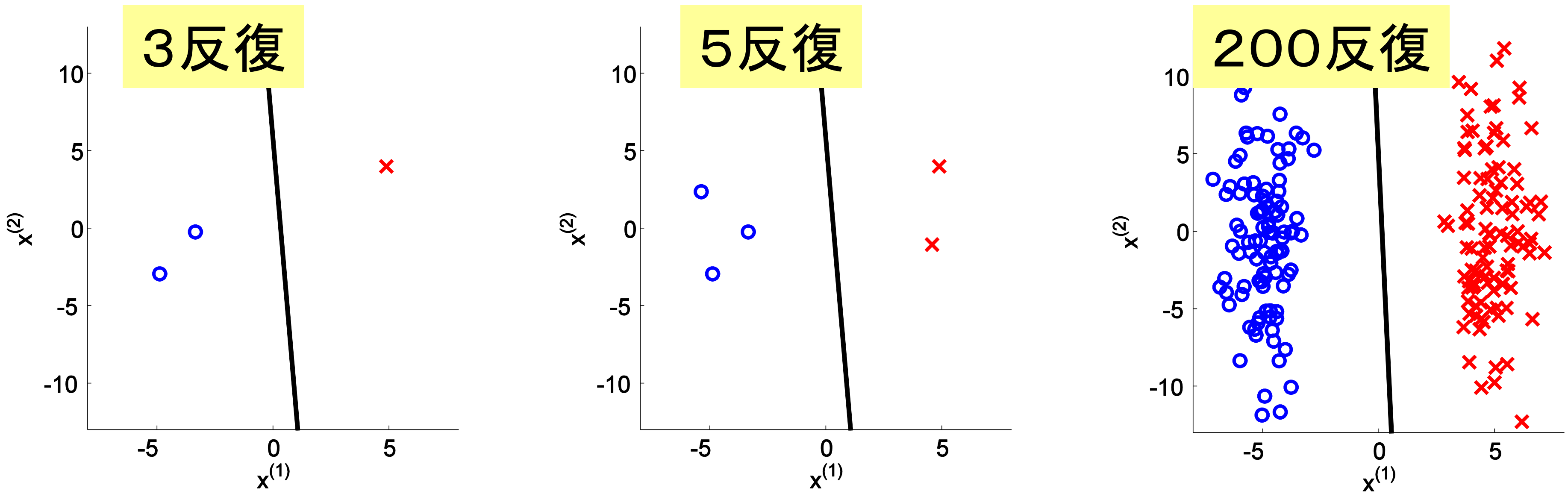
$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \frac{y \max\left(0, 1 - y\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})\right)}{\|\boldsymbol{\phi}(\boldsymbol{x})\|^2 + \gamma} \boldsymbol{\phi}(\boldsymbol{x})$$

導出は後述

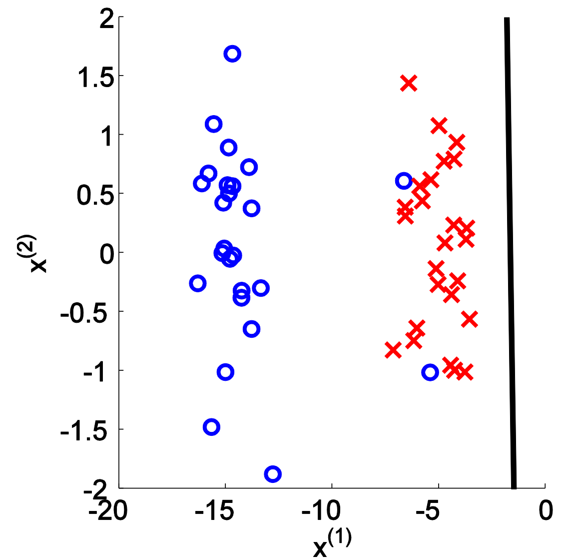


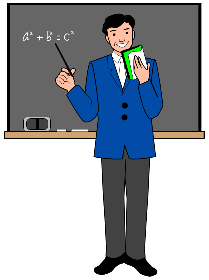
実行例

- うまく分類できている



- ただし，異常値に弱い



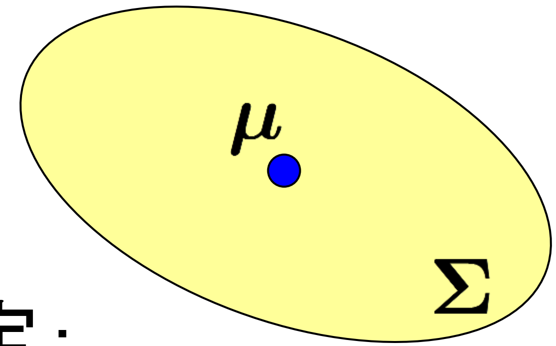


講義の流れ

1. 確率的勾配法
2. PA学習
3. 適応正則化学習

適応正則化学習

- パラメータの確率分布を学習（ベイズ推定とは異なる）
  - 確信度を表現可能



- 簡単のため，ガウス分布を仮定：

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- パラメータの変化量を分布間のカルバック・ライブラー距離で測る：

$$\mathrm{KL}\left[N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \middle\| N(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})\right] \qquad \mathrm{KL}(p\|\widetilde{p}) = \int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}$$

- 予測値の分散 $\phi(\boldsymbol{x})^\top \boldsymbol{\Sigma} \phi(\boldsymbol{x})$ を考慮する

数学演習

- パラメータがガウス分布

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

にしたがうとき，点 $\boldsymbol{x}$ での予測値 $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x})$ の分散は

$$\mathrm{Var}[f_{\boldsymbol{\theta}}(\boldsymbol{x})] = \phi(\boldsymbol{x})^\top \boldsymbol{\Sigma} \phi(\boldsymbol{x})$$

で与えられることを示せ

- ヒント: 分散の定義

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])(X - \mathrm{E}[X])]$$

解答例

$$
\begin{aligned}
\mathrm{Var}[f_{\boldsymbol{\theta}}(\boldsymbol{x})] &= \mathrm{Var}[\boldsymbol{\theta}^\top \phi(\boldsymbol{x})] \\
&= \mathrm{E}[((\boldsymbol{\theta} - \boldsymbol{\mu})^\top \phi(\boldsymbol{x}))^2] \\
&= \mathrm{E}[\phi(\boldsymbol{x})^\top (\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \phi(\boldsymbol{x})] \\
&= \phi(\boldsymbol{x})^\top \mathrm{E}[(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top] \phi(\boldsymbol{x}) \\
&= \phi(\boldsymbol{x})^\top \boldsymbol{\Sigma} \phi(\boldsymbol{x})
\end{aligned}
$$

適応正則化学習(続き)

- パラメータの学習規準:

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \left\{ \mathrm{loss}\Big(f_{\boldsymbol{\mu}}(\boldsymbol{x}), y \Big) \right.$$

訓練出力に対する
適合の良さ

$$+2\gamma \mathrm{KL}\left[N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \middle\| N(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}) \right]$$

パラメータの
変化量を抑制

$\gamma > 0$

$$\left. +\boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x}) \right\}$$

適応正則化:
予測値の分散を抑制

(最後の項について正則化係数 γ' を導入してもよい)

数学演習

$$\mathrm{KL}(p\|\widetilde{p}) = \int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}$$

■ガウス分布間のカルバック･ライブラー距離が

$$\mathrm{KL}\left[N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \middle\| N(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}) \right]$$

$$= \frac{1}{2}\left\{ \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) - d \right\}$$

で与えられることを示せ.

d: $\boldsymbol{x}$の次元数

ヒント

■ガウス分布 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$の確率密度関数 $p(\boldsymbol{\theta})$は

$$p(\boldsymbol{\theta}) = (2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right)$$

■ $(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \right)$

■分散共分散行列: $\int p(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \mathrm{d}\boldsymbol{\theta} = \boldsymbol{\Sigma}$

■期待値ベクトル: $\int p(\boldsymbol{\theta}) \boldsymbol{\theta} \mathrm{d}\boldsymbol{\theta} = \boldsymbol{\mu}$

数学演習(再掲)

■ 以下の式を示せ:

$$\mathrm{KL}\left[N(\boldsymbol{\mu},\boldsymbol{\Sigma})\middle\|N(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})\right]$$

$$\mathrm{KL}(p\|\widetilde{p}) = \int p(\boldsymbol{\theta})\log\frac{p(\boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}$$

$$= \frac{1}{2}\left\{\log\frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}})^{\top}\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}) - d\right\}$$

$$p(\boldsymbol{\theta}) = (2\pi)^{-d/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right)$$

$$(\boldsymbol{\theta}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^{\top}\right)$$

$$\int p(\boldsymbol{\theta})\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta} = \boldsymbol{\mu}$$

$$\int p(\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^{\top}\mathrm{d}\boldsymbol{\theta} = \boldsymbol{\Sigma}$$

解答例

$$\int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \frac{1}{2} \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})}$$

$$-\frac{1}{2} \int p(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \mathrm{d}\boldsymbol{\theta} + \frac{1}{2} \int p(\boldsymbol{\theta})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\mu}}) \mathrm{d}\boldsymbol{\theta}$$

$$= \frac{1}{2} \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \frac{1}{2} \mathrm{tr} \Bigg(-\boldsymbol{\Sigma}^{-1} \int p(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \mathrm{d}\boldsymbol{\theta}$$

$$+ \widetilde{\boldsymbol{\Sigma}}^{-1} \int p(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\mu} + \boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})(\boldsymbol{\theta} - \boldsymbol{\mu} + \boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \mathrm{d}\boldsymbol{\theta} \Bigg)$$

$$= \frac{1}{2} \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \frac{1}{2} \mathrm{tr} \left(-\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} + \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \right) + \frac{1}{2} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})$$

$$= \frac{1}{2} \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} - \frac{d}{2} + \frac{1}{2} \mathrm{tr} \left(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \right) + \frac{1}{2} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})$$

PA学習との関係

- Σ を単位行列に固定すれば,
  μ の学習規準はPA学習と一致:

適応正則化学習		PA学習
$\mathrm{loss}\Big(f_{\boldsymbol{\mu}}(\boldsymbol{x}), y\Big)$	$\Rightarrow$	$\mathrm{loss}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y\Big)$
$\mathrm{KL}\Big[N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\Big\Vert N(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})\Big]$	$\Rightarrow$	$\Vert\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\Vert^2$
$\phi(\boldsymbol{x})^\top \boldsymbol{\Sigma} \phi(\boldsymbol{x})$	$\Rightarrow$	定数

解の求め方：共分散行列

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \mathop{\mathrm{argmin}}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{loss}\Big(f_{\boldsymbol{\mu}}(\boldsymbol{x}), y\Big) + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x})$$

$$+\gamma \left\{ \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) - d \right\}$$

- 数学演習：適応正則化学習における$\boldsymbol{\Sigma}$の解が次式で与えられることを示せ

$$\widehat{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma}$$

ヒント

- 行列での微分の公式を利用する:
  - $$\frac{\partial}{\partial \boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top$$
  - $$\frac{\partial}{\partial \boldsymbol{\Sigma}} \log \det(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}$$
  - $$\frac{\partial}{\partial \boldsymbol{\Sigma}} \mathrm{tr}\left(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}\right) = \widetilde{\boldsymbol{\Sigma}}^{-1}$$

- 逆行列の公式を利用する:

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \widetilde{\boldsymbol{\Sigma}}^{-1} + \frac{1}{\gamma} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top$$

$$\Longrightarrow \quad \widehat{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma}$$

解の求め方：共分散行列（再掲）

$$J(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathrm{loss}\Big(f_{\boldsymbol{\mu}}(\boldsymbol{x}), y\Big) + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})$$
$$+\gamma\left\{\log\frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}})^\top\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}) - d\right\}$$

- 数学演習：以下の式を示せ

$$\widehat{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})+\gamma}$$

$$(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}) = \mathop{\mathrm{argmin}}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} J(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\log\det(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathrm{tr}\left(\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\right) = \widetilde{\boldsymbol{\Sigma}}^{-1}$$

$$\left(\widetilde{\boldsymbol{\Sigma}}^{-1} + \frac{1}{\gamma}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top\right)^{-1} = \widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})+\gamma}$$

解答例

$$J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y\right)^2 + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x})$$

$$+\gamma \left\{ \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) - d \right\}$$

- 目的関数をΣに関して微分してゼロとおくと

$$\frac{\partial J(\boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}})}{\partial \boldsymbol{\Sigma}} = \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top - \gamma \widehat{\boldsymbol{\Sigma}}^{-1} + \gamma \widetilde{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{O}$$

$$\Longrightarrow \widehat{\boldsymbol{\Sigma}}^{-1} = \widetilde{\boldsymbol{\Sigma}}^{-1} + \frac{1}{\gamma} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top$$

$$\Longrightarrow \widehat{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma}$$

（逆行列の公式）

解の求め方：期待値

- 数学演習：二乗損失に対する適応正則化回帰の μ の解は次式で与えられることを示せ

$$\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}} - \frac{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\mu}} - y}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x})$$

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \operatorname*{argmin}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y\right)^2 + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x})$$
$$+\gamma \left\{ \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) - d \right\}$$

- ヒント：スカラー a と縦ベクトル $\boldsymbol{b}$ に対して $a \cdot \boldsymbol{b} = \boldsymbol{b} a$

また
$$\left(\widetilde{\boldsymbol{\Sigma}}^{-1} + \frac{1}{\gamma} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top\right)^{-1} = \widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma}$$

解答例

$$J(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \left(\boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y\right)^2 + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})$$

$$+\gamma\left\{\log\frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}) - d\right\}$$

- 目的関数を $\boldsymbol{\mu}$ に関して微分してゼロとおくと

$$\frac{\partial J(\widehat{\boldsymbol{\mu}},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = 2\boldsymbol{\phi}(\boldsymbol{x})\left(\widehat{\boldsymbol{\mu}}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y\right) + 2\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}) = \mathbf{0} \quad \text{(A)}$$

$$\Longrightarrow \left(\widetilde{\boldsymbol{\Sigma}}^{-1} + \frac{1}{\gamma}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top\right)\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\mu}} + \frac{\boldsymbol{\phi}(\boldsymbol{x})y}{\gamma} \quad \text{(B)}$$

- 逆行列の公式より

$$\widehat{\boldsymbol{\mu}} = \left(\widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma}\right)\left(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\mu}} + \frac{\boldsymbol{\phi}(\boldsymbol{x})y}{\gamma}\right) \quad \text{(C)}$$

解答例(続き)

$$\widehat{\boldsymbol{\mu}} = \left(\widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \right) \left(\widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{\mu}} + \frac{\boldsymbol{\phi}(\boldsymbol{x}) y}{\gamma} \right) \quad \text{(C)}$$

- (C)の各項を計算すると

$$\left(\widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \right) \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{\mu}}$$

$$= \widetilde{\boldsymbol{\mu}} - \frac{\widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \overbrace{\boxed{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\mu}}}}^{\text{スカラー}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma}$$

$$= \widetilde{\boldsymbol{\mu}} - \frac{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\mu}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) \quad \text{(D)}$$

解答例(続き)

$$\widehat{\boldsymbol{\mu}} = \left(\widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \right) \left(\widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{\mu}} + \frac{\boldsymbol{\phi}(\boldsymbol{x})y}{\gamma} \right) \quad \text{(C)}$$

- (C)の各項を計算すると

$$\left(\widetilde{\boldsymbol{\Sigma}} - \frac{\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \right) \frac{\boldsymbol{\phi}(\boldsymbol{x})y}{\gamma}$$

$$= \left(\frac{(\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma) \cdot \widetilde{\Sigma}\boldsymbol{\phi}(\boldsymbol{x}) - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \right) \frac{y}{\gamma}$$

$$= \left(\frac{\gamma \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x})}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \right) \frac{y}{\gamma} \quad \text{(E)}$$

($a \cdot \widetilde{\Sigma}\boldsymbol{\phi}(\boldsymbol{x}) = \widetilde{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top a$ を用いた)

適応正則化回帰

- パラメータの更新式:

$$\boldsymbol{\mu} \longleftarrow \boldsymbol{\mu} - (\boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y)\boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})/\beta$$

$$\boldsymbol{\Sigma} \longleftarrow \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma}/\beta$$

$$\beta = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma$$

- 不確かさ$\boldsymbol{\Sigma}$が小さくなるにつれて$\boldsymbol{\mu}$の更新量が抑えられる
- 不確かさ$\boldsymbol{\Sigma}$(の固有値)は単調に減少

適応正則化回帰

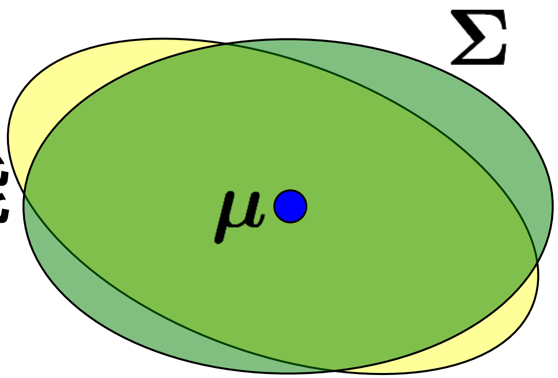
- パラメータの更新式:

$$\boldsymbol{\mu} \longleftarrow \boldsymbol{\mu} - (\boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y)\boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})/\beta$$

$$\boldsymbol{\Sigma} \longleftarrow \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma}/\beta$$

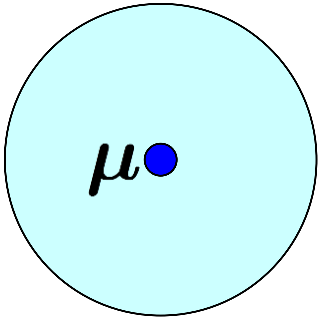
$$\beta = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma}\boldsymbol{\phi}(\boldsymbol{x}) + \gamma$$

- $\boldsymbol{\Sigma}$ を対角行列で近似すれば, 計算時間とメモリ使用量を削減可能



- $\boldsymbol{\Sigma}$ を単位行列に固定すれば, $\boldsymbol{\mu}$ の更新式はPA学習と一致:

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) - y}{\|\boldsymbol{\phi}(\boldsymbol{x})\|^2 + \gamma}\boldsymbol{\phi}(\boldsymbol{x})$$



二乗ヒンジ損失に対する期待値の解

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \mathop{\mathrm{argmin}}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\max\left(0, 1 - \boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) y\right)\right)^2 + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x})$$

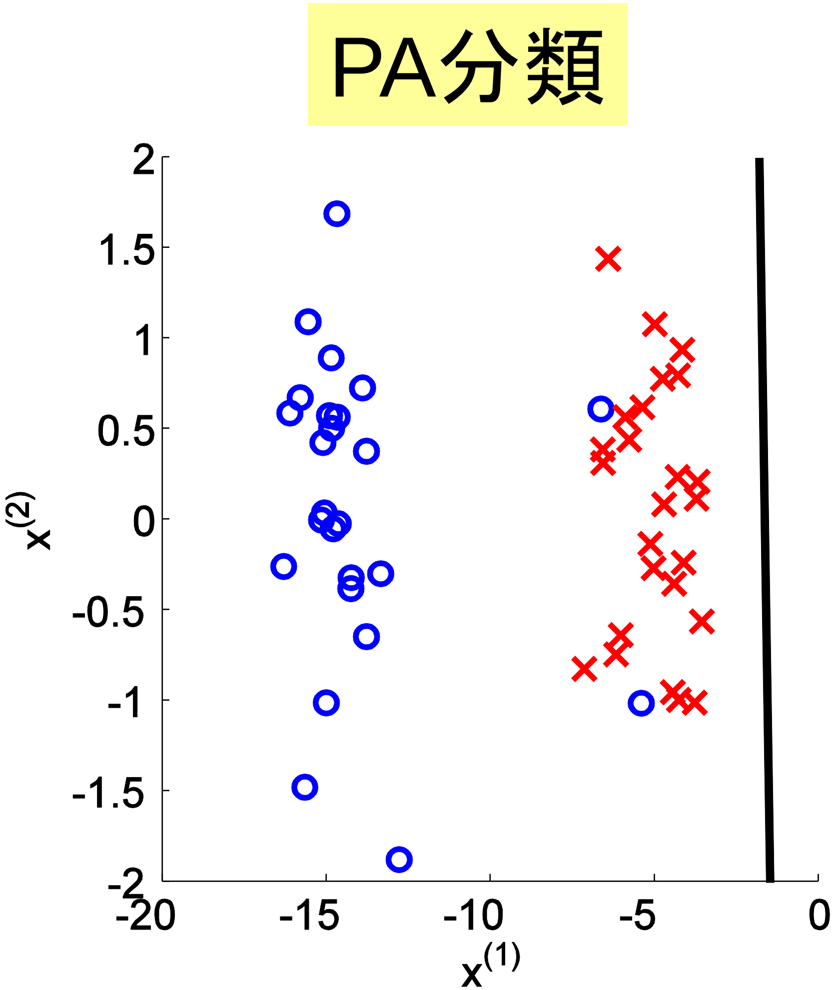
$$+\gamma \left\{ \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) - d \right\}$$

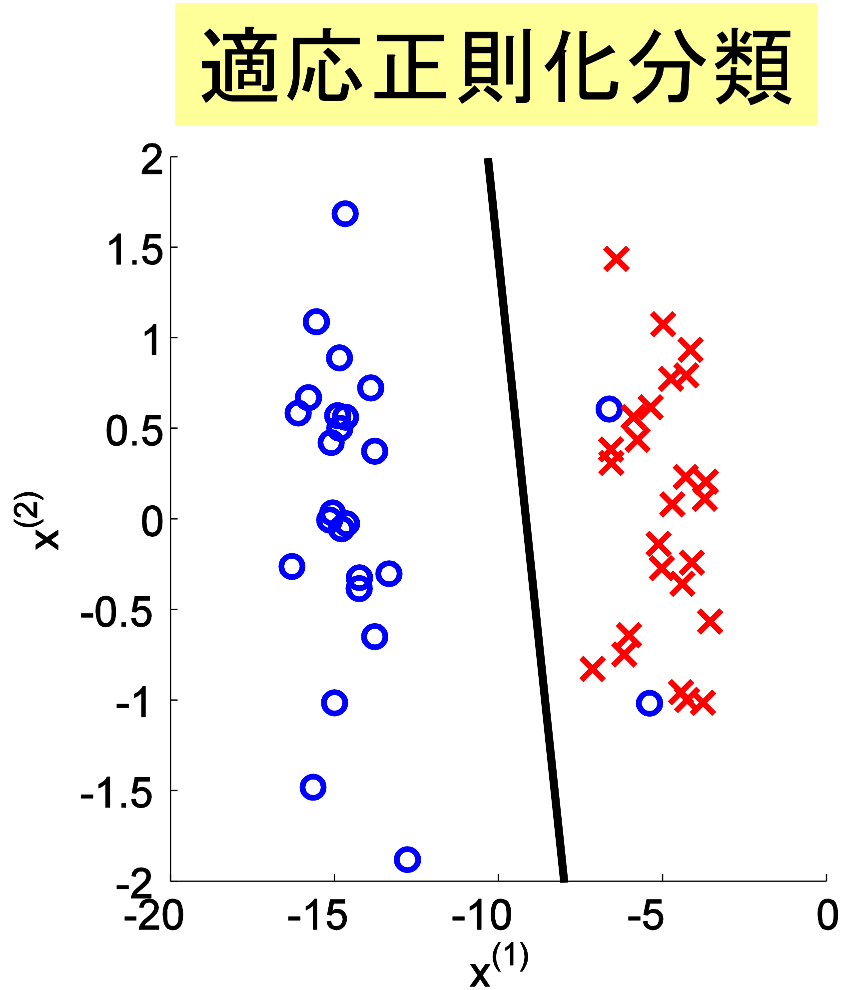
- 二乗ヒンジ損失に対する適応正則化分類の $\boldsymbol{\mu}$ の解は次式で与えられる:

$$\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}} + \frac{y \max(0, 1 - \widetilde{\boldsymbol{\mu}}^\top \boldsymbol{\phi}(\boldsymbol{x}) y)}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x})$$

証明は宿題

実行例

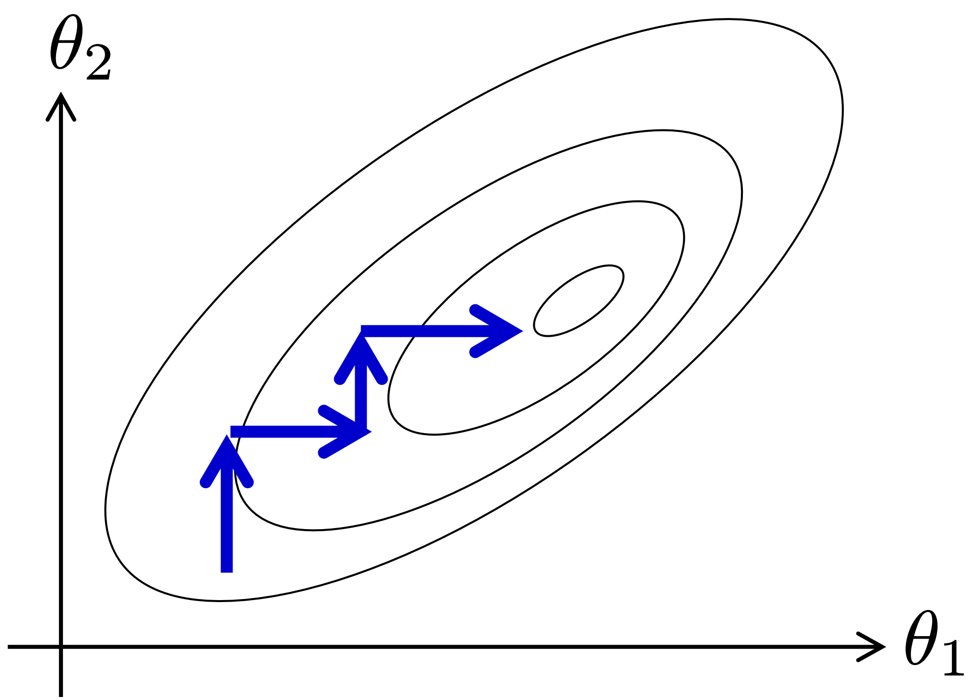


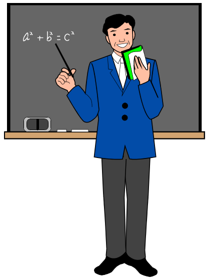


- 異常値に対してロバストな解が得られている

オンライン学習：まとめ

- オンライン学習では，一部の標本のみを使うことにより，勾配の計算を高速化
- パラメータを一部分ごとに学習する座標降下法も実用上有用



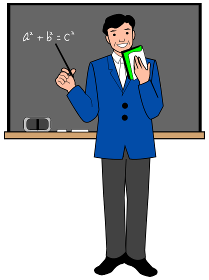


講義の流れ

1. 確率的勾配法
2. PA学習
3. 適応正則化学習

まとめ

- オンライン学習：データを一つ一つ学習していく
  - 確率的勾配法：学習係数の決定が困難
  - PA学習：パラメータの変化量を抑制
  - 適応正則化学習：パラメータの分布を考える



次回の予告

- 6/11は休講，次回は6/18

- 半教師付き学習（16章）
- 転移学習（18章）

宿題1

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \mathop{\mathrm{argmin}}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\max\left(0, 1 - \boldsymbol{\mu}^\top \boldsymbol{\phi}(\boldsymbol{x}) y\right)\right)^2 + \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\phi}(\boldsymbol{x})$$

$$+\gamma \left\{ \log \frac{\det(\widetilde{\boldsymbol{\Sigma}})}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) - d \right\}$$

- 二乗ヒンジ損失に対する適応正則化分類の $\boldsymbol{\mu}$ の解は次式で与えられることを示せ

$$\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}} + \frac{y \max(0, 1 - \widetilde{\boldsymbol{\mu}}^\top \boldsymbol{\phi}(\boldsymbol{x}) y)}{\boldsymbol{\phi}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x}) + \gamma} \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\phi}(\boldsymbol{x})$$

$$y \in \{-1, +1\}$$

宿題2

データ

```
clear all; rand('state',0); randn('state',0); n=50;
x=[randn(1,n/2)-15 randn(1,n/2)-5; randn(1,n)]';
y=[ones(n/2,1); -ones(n/2,1)]; x(1:2,1)=x(1:2,1)+10;
x(:,3)=1; p=randperm(n); x=x(p,:); y=y(p);
```

を用いて，二乗ヒンジ損失に基づく適応正則化分類を線形モデル

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}}(x^{(1)}, x^{(2)})$$

$$= (x^{(1)} \ x^{(2)} \ 1)\boldsymbol{\theta}$$

に対して実装せよ