Online Linear Classifiers

~PerceptronからCWまで~

機械学習勉強会 2011/06/23

中川研 修士課程2**年 大岩 秀和**

概要

- **オンライン学習による線形分類器の紹介**
  - Perceptron
  - MIRA
  - Passive-Aggressive
  - Confidence-Weighted Algorithms

- **条件設定**
  - **今回の発表は，2値分類に限定**
  - **多クラスへの拡張は容易**

$$f(\mathbf{x}, y)$$

$$\hat{y} = \langle \mathbf{w}, \arg\min_{y} f(\mathbf{x}_i, y) \rangle$$

Notation (Linear Classifier)

- 入力 $\mathbf{x} \in \mathbf{X} \subset \Re^d$
  - 入力ベクトルの各成分は，特徴(feature)と呼ばれる
  - Ex. 文書中の単語出現回数を並べたベクトル
- 出力 $y \in \mathbf{Y} \subset \Re$
  - 構造学習(Structured Learning)の場合はベクトル
- 教師データ $S = \{(\mathbf{x}_i, y_i)\}_{i=1,2,\ldots,T}$
  - 入力と出力のどちらもが既知

Notation (Linear Classifier)

- 重みベクトル $\mathbf{w} \in \mathbf{W} \subset \Re^d$
  - 重みベクトルと入力ベクトルの内積で出力値を予測

例：ニュース記事分類

$$\hat{y} = \langle \mathbf{w}, \mathbf{x} \rangle > 0$$ ：スポーツ記事

$$\hat{y} = \langle \mathbf{w}, \mathbf{x} \rangle < 0$$ ：スポーツ以外の記事

$y = sign(\hat{y})$ としたい

- バイアス項
  - 多くの場合，バイアス項を導入する
  - 全データで1となる特徴を1つ増やせば良い

$$\mathbf{x} = (0, 1, 3, \ldots, 2, 1)$$

Linear Classifierの一般化

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_i \ell(\mathbf{w}; \mathbf{x}_i, y_i) + Cr(\mathbf{w})$$

$\ell(\mathbf{w}; \mathbf{x}_i, y_i)$: 損失関数　　$r(\mathbf{w})$: 正則化項

- **多くのアルゴリズムがこの形式で表せる**
  - Naïve Bayes
  - SVM(Support Vector Machine)
  - Logistic Regression(Maximum Entropy)
  - Conditional Random Field
  - Online Linear Classifiers

Online Learning

- データを一つ受け取るたび，逐次的に$\mathbf{w}$を更新

学習データ

$$\mathbf{x}_{t-1} = \{0, 1, 0, 0, 2\}$$
$$y_{t-1} = 1$$

$$\mathbf{w}_{t-1}$$

学習データ

$$\mathbf{x}_{t} = \{2, 0, 0, 1, 0\}$$
$$y_{t} = -1$$

$$\mathbf{w}_{t}$$

ここの更新則を
上手く設定する

学習データ

$$\mathbf{x}_{t+1} = \{0, 0, 5, 2, 4\}$$
$$y_{t+1} = 1$$

$$\mathbf{w}_{t+1}$$

$$\vdots$$

Online Learningの長所

- 学習の**省メモリ化**
  - 重みベクトルの更新に1データのみ使用
  - 全データを一度に扱えない場合に有用

- **再学習**が容易
  - 再学習：学習器を一度構築した後，新しいデータを用いて学習器を改良
  - 新しいデータのみを用いて，再学習が可能
  - 訓練データが逐次的にやってくる場合，昔のデータを捨てたい場合に有用

- 多くの場合，実装が簡単

Perceptron [Rosenblatt 1958]

- アルゴリズム
  - 誤識別したら，正解ラベル方向へ入力データを重みベクトルに足す

Input $S = \{(\mathbf{x}_i, y_i)\}$, $\mathbf{w}_0 = \mathbf{0}$, $k = 0$.

for $i = 1, 2, \ldots$ **do**
    if $y_i \langle \mathbf{w}_k, \mathbf{x}_i \rangle \leq 0$ **then**
        $\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$
        $k = k + 1$
    end if
end for

$$y_i \neq \hat{y}_i$$

の時，更新

Perceptronの更新の妥当性

- **更新後の重みベクトルは，更新前の重みベクトルよりも，誤識別したデータを上手く識別する**

$$
\begin{aligned}
y_i\langle \mathbf{w}_{k+1}, \mathbf{x}_i\rangle &= y_i\langle \mathbf{w}_k + y_i\mathbf{x}_i, \mathbf{x}_i\rangle \\
&= y_i\langle \mathbf{w}_k, \mathbf{x}_i\rangle + y_i^2\langle \mathbf{x}_i, \mathbf{x}_i\rangle \\
&> y_i\langle \mathbf{w}_k, \mathbf{x}_i\rangle \qquad (\mathbf{x}_i \neq \mathbf{0})
\end{aligned}
$$

同じデータに対して，よりよい識別が可能になっている

線形分離可能

- **以下の条件をみたしつつ，全データを正しく識別する重みベクトル・パラメータが存在するとき，線形分離可能と呼ぶ**

重みベクトル $\|\mathbf{w}^*\|_2 = 1$

パラメータ $\gamma > 0$

$$\forall i \quad y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq \gamma$$

- **このとき，γをマージンと呼ぶ**

パーセプトロンの収束定理

[Block, 1962] [Novikoff, 1962] [Collins,2002]

- データが線形分離可能ならば，以下の定理が成立

$$\text{パーセプトロンによる誤識別回数} \leq \frac{R^2}{\gamma^2}$$

$$R = \max_i \|\mathbf{x}_i\|_2$$

- 重みベクトルのノルムの上限・下限から示す

収束定理の証明 [1/3]

$$\langle \mathbf{w}^*, \mathbf{w}_{k+1} \rangle = \langle \mathbf{w}^*, \mathbf{w}_k \rangle + y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle$$
$$\geq \langle \mathbf{w}^*, \mathbf{w}_k \rangle + \gamma$$

$\mathbf{w}_0 = \mathbf{0}$ より，

$$\langle \mathbf{w}^*, \mathbf{w}_k \rangle \geq k\gamma$$

さらに $\|\mathbf{w}\|_2 = 1$ より，

$$\|\mathbf{w}_k\|_2 \geq k\gamma$$

下限

収束定理の証明 [2/3]

$$\|\mathbf{w}_{k+1}\|_2^2 = \|\mathbf{w}_k\|_2^2 + \|\mathbf{x}_i\|_2^2 + 2y_i\langle \mathbf{w}_k, \mathbf{x}_i\rangle$$
$$\leq \|\mathbf{w}_k\|_2^2 + R^2 + 0$$

- **第2項は，入力ベクトルのノルム上限より**
- **第3項は，パーセプトロンの更新基準より**

$\mathbf{w}_0 = \mathbf{0}$ より，

$$\|\mathbf{w}_k\|_2^2 \leq kR^2$$ 上限

収束定理の証明 [3/3]

下限

$$\|\mathbf{w}_k\|_2 \geq k\gamma$$

上限

$$\|\mathbf{w}_k\|_2^2 \leq kR^2$$

$$k^2\gamma^2 \leq \|\mathbf{w}_k\|_2^2 \leq kR^2$$

$$\Rightarrow k \leq \frac{R^2}{\gamma^2}$$

重みベクトルの更新回数の上限回数が導出できる

Perceptronの亜種

- Voted Perceptron [Freund and Schapire, 1988]
  - **過去の全重みベクトルで識別，多数決を取る**
  - k**が変化しない生存期間に応じて重み付け**
- Averaged Perceptron [Collins+, 2002]
  - **過去の全重みベクトルの平均を取って識別**
- **その他にもたくさん** etc..
  - Second Order Perceptron
  - p-norm Perceptron
  - Margitron

MIRA [Crammer+ 2003]

- Margin Infused Relaxed Algorithm
  - Ultraconservative Online Algorithmsの一種

[Crammer+ 2003]

- **マージン最大化を目指したアルゴリズム**
  - Perceptron**は，マージンを最大化する重みベクトルを導出するアルゴリズムではない**
  - Max-margin Perceptron, Online SVM**と呼ばれることも**

SVM (Support Vector Machine)

$$\min \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$s.t. \quad \forall i \quad y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \geq 1$$

MIRAのアルゴリズム

Input $S = \{(\mathbf{x}_i, y_i)\}$, $\mathbf{w}_0 = \mathbf{0}$, $k = 0$.

1: **for** $i = 1, 2, \ldots$ **do**

2: $\quad \mathbf{w}_{k+1} = \mathrm{argmin}_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}_k\|_2$

3: $\quad\quad \mathrm{s.t.}\ \ y_i \langle \mathbf{x}_i, \mathbf{w}_{k+1} \rangle \geq 1$

4: $\quad k = k + 1$

5: **end for**

二次計画最適化問題に帰着
多クラスの場合は全制約を同時に満たすものを探す

- **構造問題の場合は，マージンをラベル間の編集距離と置くことも**
- **累積損失の上限値が求められる** (Passive-Aggressiveで詳しく説明します)

Online Passive-Aggressive

[Crammer+, 2006]

- Hinge-Lossを定義

$$\ell(\mathbf{w}; (\mathbf{x}, y)) = \begin{cases} 0 & y\langle \mathbf{x}, \mathbf{w} \rangle \geq 1 \\ 1 - y\langle \mathbf{x}, \mathbf{w} \rangle & otherwise \end{cases}$$

- 更新式を以下のように記述する
  - 2値分類の時は，MIRAと同じ

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|^2 \quad s.t. \quad \ell(\mathbf{w}; (\mathbf{x}_t, y_t)) = 0$$

PAの定式化

- 2値の場合，アルゴリズムはMIRAと同じ
- 上の定式化をする意図は？
  - 最適化問題の拡張が容易 (回帰問題,PA-I,PA-II,etc..)

- Ex. 回帰問題への適用

$$\ell_\epsilon(\mathbf{w};(\mathbf{x},y)) = \begin{cases} 0 & |y - \langle \mathbf{x}, \mathbf{w} \rangle| \leq \epsilon \\ |y - \langle \mathbf{x}, \mathbf{w} \rangle| - \epsilon & otherwise \end{cases}$$

PAの閉じた解の導出

- **ラグランジュ乗数法を用いる**

$$L(\mathbf{w}, \tau_t) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + \tau_t \ell(\mathbf{w}; (\mathbf{x}_t, y_t))$$

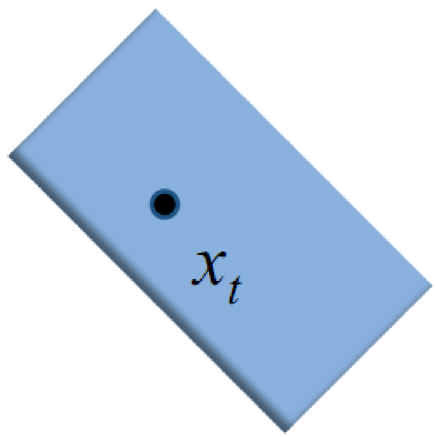
$$\frac{\partial L}{\partial \mathbf{w}} = 0 \qquad \frac{\partial L}{\partial \tau_t} = 0$$ **を計算すれば**…

$$\mathbf{w}_{t+1} = \begin{cases} 0 & y_t\langle \mathbf{x}_t, \mathbf{w}_t \rangle \geq 1 \\ \mathbf{w}_t + \tau_t y_t \mathbf{x}_t & otherwise \end{cases}$$

$$\tau_t = \frac{1 - y_t\langle \mathbf{x}_t, \mathbf{w}_t \rangle}{\|\mathbf{x}\|_2^2}$$

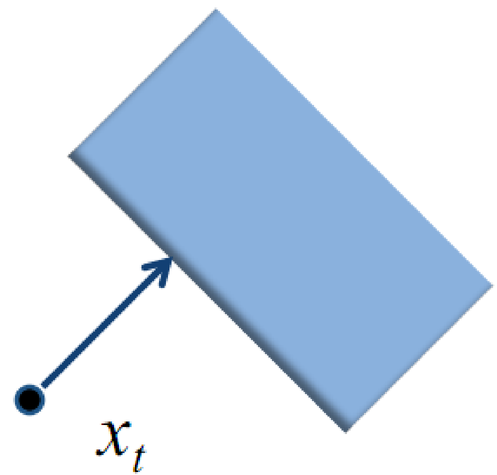
PAの特性

- 今受け取ったデータを正しく判別できるように，重みベクトルを更新する
  - 一方，ノイズに脆弱



Passive

$$\mathbf{w}_{t+1} = \mathbf{w}_t$$



Aggressive

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$$

PA-I, PA-II

- ノイズに頑健な拡張を加える
  - C はAggressiveness parameter

PA-I

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|^2 + C\xi \right\} \quad s.t. \quad \ell(\mathbf{w}; (\mathbf{x}_t, y_t)) \leq \xi, \xi \geq 0$$

PA-II

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|^2 + C\xi^2 \right\} \quad s.t. \quad \ell(\mathbf{w}; (\mathbf{x}_t, y_t)) \leq \xi, \xi \geq 0$$

誤識別を許容

PAの累積損失上限

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_t \ell(\mathbf{w}; (\mathbf{x}_t, y_t))$$ と定義した時,

$$\boxed{\sum_t \tau_t \left(\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)) - \tau_t \|\mathbf{x}_t\|_2^2 - 2\ell(\mathbf{w}^*; (\mathbf{x}_t, y_t))\right) \leq \|\mathbf{w}^*\|_2^2}$$

特に，線形分離可能な時 $(\forall t \quad \ell(\mathbf{w}^*; (\mathbf{x}_t, y_t)) = 0)$

$$\sum_t \ell^2(\mathbf{w}_t; (\mathbf{x}_t, y_t)) \leq \|\mathbf{w}^*\|_2^2 R^2$$

$$s.t. \quad R = \max_i \|\mathbf{x}_i\|_2$$

PA累積損失上限の証明 [1/3]

$\Delta_t = \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2$ **と定義し，**

$\sum_{t=1}^{T} \Delta_t$ **の上限と下限から導く**

$$\sum_{t=1}^{T} \Delta_t = \sum_{t=1}^{T} \left(\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \right)$$

$$= \|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_{T+1} - \mathbf{w}^*\|_2^2$$

$$\leq \|\mathbf{w}^*\|_2^2 - 0$$

上限

PA累積損失上限の証明 [2/3]

$\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)) \geq 0$ **のとき，**

$$\begin{aligned}
\Delta_t &= \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}^* + y_t \tau_t \mathbf{x}_t\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 - \left(\|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + 2 y_t \tau_t \langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{x}_t \rangle + \tau_t^2 \|\mathbf{x}_t\|_2^2\right) \\
&= -2 y_t \tau_t \langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{x}_t \rangle - \tau_t^2 \|\mathbf{x}_t\|_2^2 \\
&\geq 2\tau_t \left((1 - \ell(\mathbf{w}^*; (\mathbf{x}_t, y_t)) - (1 - \ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)))\right) - \tau_t^2 \|\mathbf{x}_t\|_2^2
\end{aligned}$$

下限

最後の不等式は，以下の条件式より

$$\ell(\mathbf{w}^*; (\mathbf{x}_t, y_t)) \geq 1 - y_t \langle \mathbf{w}^*, \mathbf{x}_t \rangle$$

$$\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)) = 1 - y_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle$$

$\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)) < 0$ **のとき，** $\Delta_t = 0$

PA累積損失上限の証明 [3/3]

$$\sum_{t=1}^{T} \Delta_t \leq \|\mathbf{w}^*\|_2^2$$ 上限

$$\Delta_t \geq 2\tau_t \left(\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t) - \ell(\mathbf{w}^*; (\mathbf{x}_t, y_t) - \tau_t \|\mathbf{x}_t\|_2^2 \right)$$ 下限

より，

$$\sum_t \tau_t \left(\ell(\mathbf{w}_t; (\mathbf{x}_\tau, y_\tau)) - \tau_t \|\mathbf{x}_t\|_2^2 - 2\ell(\mathbf{w}^*; (\mathbf{x}_\tau, y_\tau)) \right) \leq \|\mathbf{w}^*\|_2^2$$

が導出される

線形分離時や，PA-I,PA-IIも同様に証明可能

CW以前のアルゴリズムの問題点

- NLP等の分類問題は特徴次元数が**大**
- 多くの特徴は**低頻度**
  - 低頻度の特徴が分類上重要な役割を果たすことも
- 既存手法では，データ中に特徴が出現した時のみ，対応するパラメータが更新される
  - 高頻度の特徴は，パラメータも頻繁に更新
  - 低頻度の特徴は，余り更新されない
- **過去の更新回数**をパラメータ更新に用いていない
  - 非効率的

Ex. Passive-Aggressive

INPUT: aggressiveness parameter $C > 0$

INITIALIZE: $\mathbf{w}_1 = (0, \ldots, 0)$

For $t = 1, 2, \ldots$

- receive instance: $\mathbf{x}_t \in \mathbb{R}^n$
- predict: $\hat{y}_t = \text{sign}(\mathbf{w}_t \cdot \mathbf{x}_t)$
- receive correct label: $y_t \in \{-1, +1\}$
- suffer loss: $\ell_t = \max\{0\ ,\ 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)\}$ ⇐ スカラー
- update:

  1. set:

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2} \qquad \text{(PA)}$$

$$\tau_t = \min\left\{C\ ,\ \frac{\ell_t}{\|\mathbf{x}_t\|^2}\right\} \qquad \text{(PA-I)}$$

⇐ スカラー

$$\tau_t = \frac{\ell_t}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}} \qquad \text{(PA-II)}$$

  2. update: $\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$ ⇐ 特徴ベクトルのスカラー倍

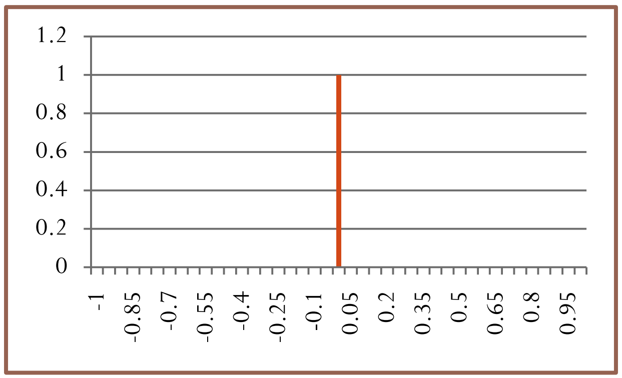
[Crammer+, 2006]

パラメータの更新は，出現頻度と独立

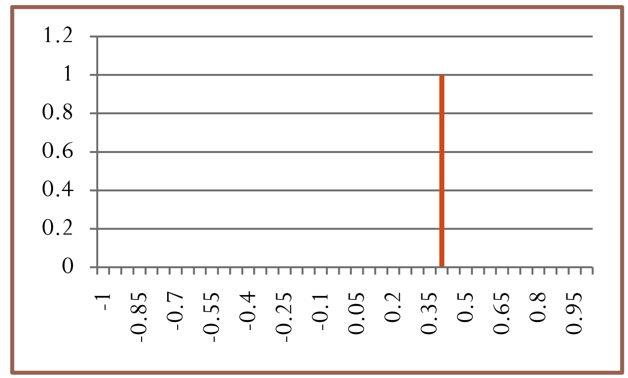
Confidence-Weighted Algorithms(CW)

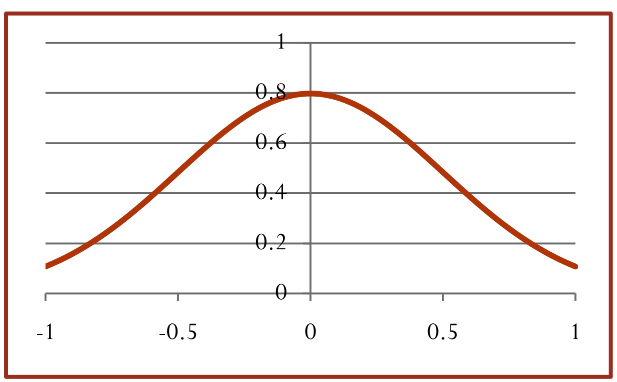
[Clammer+, 2008]

- 重みベクトル上にガウス分布を導入
- 重みベクトルの平均・共分散を逐次的に更新

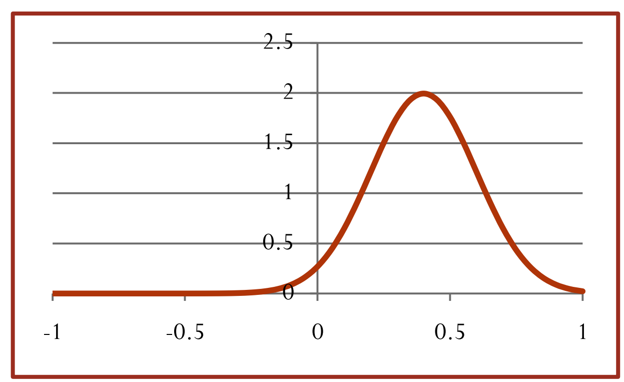


既存手法





CW



CWの特性

- **分散の大きい（自信のない）パラメータは大きく更新，分散の小さい（自信のある）パラメータは小さく更新**
  - **毎回更新するたびに，分散は小さくする**

- **収束速度が高速**
  - **収束に至るまでのデータ数が非常に少ない**
    - **稀な特徴を上手く利用しているため**
  - **一方，稀な特徴を持つデータにラベルノイズが載っていると，性能が急激に悪化する**

CWの重みベクトルを再定義

- **重みベクトル** $\mathbf{w} \sim N(\mu, \Sigma)$
  - **平均** $\mu \in R^d$
  - **分散** $\Sigma \in R^{d \times d}$

- **この時,** $(\mathbf{x}_i, y_i)$**が正しく識別される確率**

$$Pr_{\mathbf{w} \sim N(\mu_i, \Sigma_i)}[y_i \langle \mathbf{w}_i, \mathbf{x}_i \rangle \geq 0]$$

$$= Pr_{m \sim M}[m \geq 0]$$

$$M \sim N\left(y_i \langle \mu_i, \mathbf{x}_i \rangle, \mathbf{x}_i^T \Sigma_i \mathbf{x}_i\right)$$

最適化問題

以前の多変量ガウス分布に
最も近いガウス分布を選択する

$$(\mu_{i+1}, \Sigma_{i+1}) = \arg\min_{(\mu,\Sigma)} D_{KL}\left(N(\mu,\Sigma) \| N(\mu_i,\Sigma_i)\right)$$

$$s.t. \quad Pr_{\mathbf{w}\sim N(\mu,\Sigma)}\left[y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 0\right] \geq \eta$$

誤識別率が1-η以下となるガウス分布の中で

$$0.5 \leq \eta \leq 1.0$$

- Motivationは，PAと同じ
  - i番目の重みベクトルから(KL-divergenceの意味で)一番近い，制約を満たす重みベクトルへ更新
  - 今回受け取ったデータを正確に識別するガウス分布へ移動
    - その制約を外したもの…AROW, NAROW等

最適化問題を展開

$$\min \frac{1}{2}\left\{\log\left(\frac{\det \Sigma_i}{\det \Sigma}\right) + Tr(\Sigma_i^{-1}\Sigma) + (\mu_i - \mu)^T \Sigma_i^{-1}(\mu_i - \mu)\right\}$$

$$s.t. \quad y_i \langle \mu, \mathbf{x}_i \rangle \geq \phi\left(\mathbf{x}_i^T \Sigma \mathbf{x}_i\right)$$

ここで， $\phi = \Phi^{-1}(\eta) \quad \Phi(\cdot)$ **：標準正規分布**

これをラグランジュ乗数法で解くと，

$$\mu_{i+1} = \mu_i + \alpha_i y_i \Sigma_i \mathbf{x}_i$$

$$\Sigma_{i+1}^{-1} = \Sigma_i^{-1} + \alpha_i \phi \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sqrt{\mathbf{x}_i \Sigma_{i+1} \mathbf{x}_i}}$$

まだ，閉じた解にはなっていない

もう少し頑張ると…

Input parameters $a > 0\,;\, \eta \in [0.5, 1]$
Initialize $\boldsymbol{\mu}_1 = \mathbf{0}\,,\ \Sigma_1 = aI \quad , \quad \phi = \Phi^{-1}(\eta)\,,\ \psi = 1 + \phi^2/2\,,\ \xi = 1 + \phi^2\,.$
For $i = 1, \ldots, n$

- Receive a training example $\boldsymbol{x}_i \in \mathbb{R}^d$
- Compute Gaussian margin distribution $M_i \sim \mathcal{N}\left((\boldsymbol{\mu}_i \cdot \boldsymbol{x}_i)\,, (\boldsymbol{x}_i^\top \Sigma_i \boldsymbol{x}_i) \right)$
- Receive true label y_i and compute

$$v_i = \boldsymbol{x}_i^\top \Sigma_i \boldsymbol{x}_i \ \ , \ \ m_i = y_i\,(\boldsymbol{\mu}_i \cdot \boldsymbol{x}_i)\,(11) \quad , \ \ u_i = \frac{1}{4}\left(-\alpha v_i \phi + \sqrt{\alpha^2 v_i^2 \phi^2 + 4 v_i}\right)^2 \qquad (12)$$

$$\alpha_i = \max\left\{0, \frac{1}{v_i \xi}\left(-m_i \psi + \sqrt{m_i^2 \frac{\phi^4}{4} + v_i \phi^2 \xi}\right)\right\}(14) \quad , \ \ \beta_i = \frac{\alpha_i \phi}{\sqrt{u_i} + v_i \alpha_i \phi} \qquad (22)$$

- Update $\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_i + \alpha_i y_i \Sigma_i \boldsymbol{x}_i$

$$\Sigma_{i+1} = \Sigma_i - \beta_i \Sigma_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \Sigma_i \qquad \text{(full)} \qquad (10)$$

$$\Sigma_{i+1} = \left(\Sigma_i^{-1} + \alpha_i \phi u_i^{-\frac{1}{2}} \operatorname{diag}^2(\boldsymbol{x}_i)\right)^{-1} \qquad \text{(diag)} \qquad (15)$$

Output Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}_{n+1}, \Sigma_{n+1}\right)$.

[Clammer+, 2008]

Mistake Bound for CW

- これまでのデータを全て正しく識別できる最適なガウス分布が存在する場合には，更新回数の上限が定められる

Theorem 4 *Let* $(\boldsymbol{x}_1, \boldsymbol{y}_1) \ldots (\boldsymbol{x}_n, \boldsymbol{y}_n)$ *be an input sequence for the algorithm of Fig. 1, initialized with* $(\mathbf{0}, I)$, *with* $\boldsymbol{x}_i \in \mathbb{R}^d$ *and* $\boldsymbol{y}_i \in \{-1, +1\}$. *Assume there exist* $\boldsymbol{\mu}^*$ *and* Σ^* *such that for all* i *for which the algorithm made an update* $(\alpha_i > 0)$,

$$\boldsymbol{\mu}^{*\top} \boldsymbol{x}_i y_i \geq \boldsymbol{\mu}_{i+1}^{\top} \boldsymbol{x}_i y_i \quad \text{and} \quad \boldsymbol{x}_i^{\top} \Sigma^* \boldsymbol{x}_i \leq \boldsymbol{x}_i^{\top} \Sigma_{i+1} \boldsymbol{x}_i \quad . \tag{18}$$

Then the following holds:

$$\text{no. mistakes} \leq \sum_i \alpha_i^2 v_i \leq \frac{1+\phi^2}{\phi^2} \left(-\log\det\Sigma^* + \mathrm{Tr}\left(\Sigma^*\right) + \boldsymbol{\mu}^{*\top} \Sigma_{n+1}^{-1} \boldsymbol{\mu}^* - d \right) \tag{19}$$

[Clammer+, 2008]

証明は略

実験結果 (CW)

CW

	Task	PA	Variance	Variance-Exact	SVM	Maxent	SGD
20 Newsgroups	comp	8.90	†**6.33**	9.63	∗7.67	∗7.62	7.36
	sci	4.22	†**1.78**	3.3	†3.51	†3.55	†4.77
	talk	1.57	1.09	2.21	**0.91**	**0.91**	1.36
Reuters	Business	17.80	17.65	17.70	⋆15.64	⋆**15.10**	⋆15.85
	Insurance	9.76	∗**8.45**	9.49	9.19	8.59	9.05
	Retail	15.41	†**11.05**	14.14	∗12.80	∗12.30	†14.31
Sentiment	books	19.55	∗**17.40**	20.45	†20.45	†19.91	∗19.41
	dvds	19.71	**19.11**	19.91	20.09	19.26	20.20
	electronics	17.40	†**14.10**	17.44	†16.80	†16.21	†16.81
	kitchen	15.64	∗**14.24**	16.35	15.20	14.94	∗15.60
	music	20.05	∗**18.10**	19.66	19.35	19.45	18.81
	videos	19.86	⋆**17.20**	19.85	†20.70	†19.45	⋆19.65

Table 2. Error on test data using batch training. Statistical significance (McNemar) is measured against PA or the batch method against Variance. (∗ p=.05, ⋆ p=.01, † p=.001)

[Dredze+, 2008]

さらなる発展形

- AROW [Crammer+, 2009], NAROW [Orabona+, 2010]
  - PAに対するPA-I等と似たMotivation
  - ノイズに頑健

- Adaptive SubGradient Methods (AdaGrad)

  [Dutch+, 2010]
  - 二次の補正をかけた劣勾配法に拡張
  - CW, AROWと同様の効果を持つ（更新回数を考慮）
  - 以下のブログ記事の考察も興味深いです
    - http://atpassos.posterous.com/the-similarity-between-confidence-weighted-le

参考：Algorithm(AROW)

$$C(\boldsymbol{\mu},\Sigma) = D_{KL}(N(\boldsymbol{\mu},\Sigma) \| N(\boldsymbol{\mu}_{t-1},\Sigma_{t-1})) + \lambda_1 \ell_{h^2}(y_t, \boldsymbol{\mu}\cdot\mathbf{x}_t) + \lambda_2 \mathbf{x}_t^T \Sigma \mathbf{x}_t$$

$$\ell_{h^2}(y_t, \boldsymbol{\mu}\cdot\mathbf{x}_t) = (\max\{0, 1 - y_t(\boldsymbol{\mu}\cdot\mathbf{x}_t)\})^2$$

- **第一項**-- $D_{KL}(N(\boldsymbol{\mu},\Sigma) \| N(\boldsymbol{\mu}_{t-1},\Sigma_{t-1}))$
  - **以前のパラメータから大きく更新しない**
- **第二項**-- $\ell_{h^2}(y_t, \boldsymbol{\mu}\cdot\mathbf{x}_t)$
  - **損失関数を最小にする**
  - Hinge-loss**以外の損失関数でも良い**
- **第三項**-- $\mathbf{x}_t^T \Sigma \mathbf{x}_t$
  - **学習するにつれ、**Σ**を小さくする**

$\mathbf{x} \in \mathbf{R}^d$: 特徴ベクトル

$\boldsymbol{\mu} \in \mathbf{R}^d$: 重みベクトルの平均

$\boldsymbol{\sigma} \in \mathbf{R}^d$: 重みベクトルの分散

$\Sigma \in \mathbf{R}^{d\times d}$: 重みベクトルの共分散

$\mathbf{w} \sim N(\boldsymbol{\mu},\Sigma)$: 重みベクトル

$y \in \{-1,1\}$: 正解ラベル

$\eta \in (0.5,1]$: しきい値

λ_1, λ_2 : *hyperparameters*

まとめ

- Online Linear Classifierについて紹介
  - 特に，CWはSVMとも遜色ない精度
  - BatchのLinearSVM, OnlineのPA,CWは線形識別器におけるベンチマーク
- オンライン学習の特性を最大限利用
  - 高速に収束（特に冗長データに対して）
  - 空間計算量を節約
  - 実装が単純