オンライン学習

定式化
評価法：Regretなど
パーセプトロン
Passive Aggressive Algorithm
(アルゴリズムと損失の限界の評価）
Confidence Weighted Algorithm
Pegasos
Coordinate Descent
バッチ、オンライン、ストリームの比較
ビッグデータへの対応

オンライン（あるいは逐次）学習とは

- データを1つづつ読み込んで、それまでの学習結果を更新する。
- 2つの利用局面
  1. データ全体は保持しているが、学習を1データ毎に行う
  2. データが1こずつ時系列としてやってくる
     - この場合はストリームという。
- データ全体をメモリの乗せなくてよいのでマシンに必要なメモリ少、あるいはメモリに乗りきらないほど大きなデータから学習可能
- 1個のデータからの学習（これを1roundという）だけなら高速

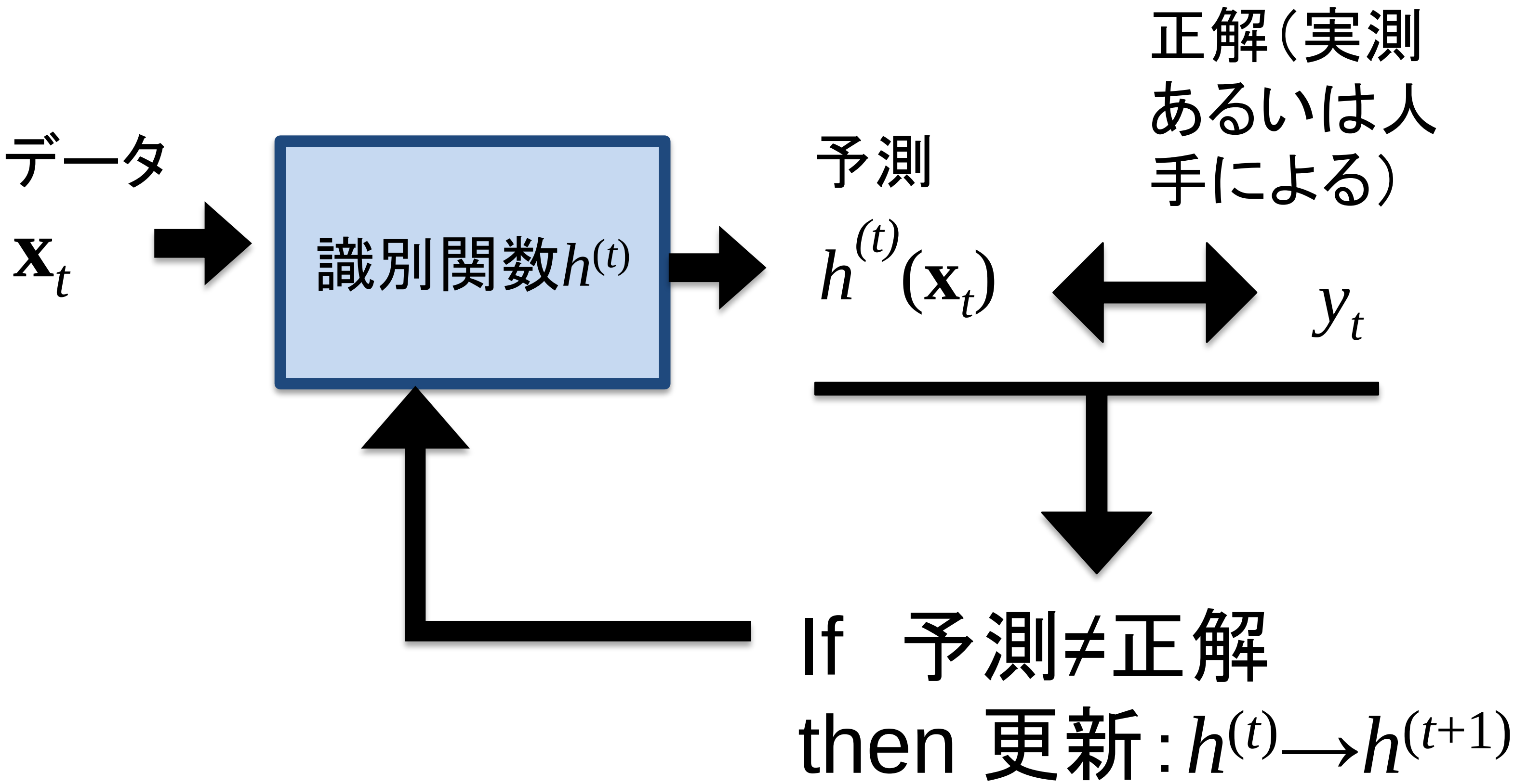
オンライン学習の概観

以下1,2,3を時刻 t=1,2,…,Tで繰り返す

1. 時刻tにおいて、仮説h_t、入力データ$\mathbf{x}_t$、正しい結果データ$y_t \in \mathcal{Y}$が与えられる。
2. 仮説h_tによる結果$h^{(t)}(\mathbf{x}_t)$を計算し、その後でy_tとの比較を損失関数ℓによって行う。つまり$\ell(h^{(t)}, (\mathbf{x}_t, y_t))$を計算
   - ℓとしては2乗損失やヒンジ損失など
3. 損失関数ℓの値に応じて$h^{(t)}$を更新し、新しい仮説$h^{(t+1)}$を求める

- 最終的な目的の累積損失$\sum_{t=1}^{T} \ell\left(h^{(t)}, (\mathbf{x}_t, y_t)\right)$などを最小化すること
- 簡単(線形)な仮説として重みベクトル$\mathbf{w}$とxの内積$\langle \mathbf{w}, x \rangle$を使う場合はhを$\mathbf{w}$と書き、$f_t(\mathbf{w}) = \ell\left(\mathbf{w}, (\mathbf{x}_t, y_t)\right)$　と定義

オンライン学習のイメージ



オンライン学習の評価法

- 仮説hのなす空間を$\mathcal{H}$, tラウンドの予測値を$h^{(t)}(\mathbf{x}_t)$
- 累積損失 $\sum_{t=1}^{T} \ell\left(h^{(t)}(\mathbf{x}_t), (\mathbf{x}_t, y_t)\right)$（最小化したい）:

$$\mathrm{Regret}_{\mathrm{T}}\left(h^*\right) = \sum_{t=1}^{T} \ell\left(h^{(t)}(\mathbf{x}_t), (x_t, y_t)\right) - \sum_{t=1}^{T} \ell\left(h^*, (x_t, y_t)\right): \quad h^* \in \mathcal{H}$$

$$\mathrm{Regret}_{\mathrm{T}}\left(\mathcal{H}\right) = \max_{h^* \in \mathcal{H}} \mathrm{Regret}_{\mathrm{T}}\left(h^*\right)$$

$$= \sum_{t=1}^{T} \ell\left(h^{(t)}(\mathbf{x}_t), (x_t, y_t)\right) - \min_{h^* \in \mathcal{H}} \sum_{t=1}^{T} \ell\left(h^*, (x_t, y_t)\right)$$

- Mistake(失敗回数）のupper bound
  - 以後は識別に失敗しなくなるまでの学習回数＝学習データ数

オンライン学習をオンライン凸最適化の観点から定式化

By Shai Shalev-Shwartz

- 以下では $\mathcal{L}(\mathbf{w}, (x_i, y_i))$ を $f_i(\mathbf{w})$ と略記することに留意。
- 最も簡単なオンライン学習は、過去の全roundの損失を最小化するようなwを選ぶ方法：Follow-The-Leader(FTL)

Follow – The – Leader (FTL)

$$\forall t \quad \mathbf{w}_t = \arg\min_{w \in S} \sum_{i=1}^{t-1} f_i(\mathbf{w}) \qquad S\text{は}\mathbf{w}\text{の取り得る範囲で凸}$$

Lemma 10 $\mathbf{w}_1, \mathbf{w}_2 \cdots$ を FTL で生成された重みベクトル

$\forall \mathbf{u} \in S$

$$\mathrm{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \right) \le \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \right)$$

Proof

$\sum_t f_t(\mathbf{w}_t)$をLemma 10の不等式の両辺から引き移項すると

$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \le \sum_{t=1}^{T} f_t(\mathbf{u})$　　そこのこの不等式を帰納法で導く。

base case : $T=1$の場合は$\mathbf{w}_{t+1}$の定義$\mathbf{w}_2 = \arg\min_{\mathbf{w}} f_1(\mathbf{w})$より、式の成立する$\mathbf{u}$を選べる。

induction step : $t = T-1$で不等式が成立すると仮定する。つまり

$\forall \mathbf{u} \in S \ \ \sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1}) \le \sum_{t=1}^{T-1} f_t(\mathbf{u})$　両辺に$f_T(\mathbf{w}_{T+1})$を加えると

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \le f_T(\mathbf{w}_{T+1}) + \sum_{t=1}^{T-1} f_t(\mathbf{u})$$

この式は$\forall \mathbf{u} \in S$で成立し、$\mathbf{u} = \mathbf{w}_{T+1}$でも成り立つ

$$\Rightarrow \sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \le \sum_{t=1}^{T} f_t(\mathbf{w}_{T+1}) = \min_{\mathbf{u} \in S} \sum_{t=1}^{T} f_t(\mathbf{u})$$

最後の等式は$\mathbf{w}_{T+1}$の定義$\mathbf{w}_{T+1} = \arg\min_{\mathbf{w} \in S} \sum_{t=1}^{T} f_t(\mathbf{w})$より。　□

Follow-The-Regularized-Leader (FoReL)

- FTLではwに制約がないので、過学習が危ぶまれる。そこで、正則化項(Regularizer)を加えたものを最適化（FoReL)

FoReL

$$\forall t \quad \mathbf{w}_t = \arg\min_{w \in S} \sum_{i=1}^{t-1} f(\mathbf{w})_i + R(\mathbf{w}) \qquad S\text{は}w\text{の取り得る範囲で凸}$$

Lemma 20 $\mathbf{w}_1, \mathbf{w}_2 \cdots$をFoReLで生成された重みベクトル

$$\forall \mathbf{u} \in S \quad \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \right) \le R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \right)$$

Proof　Lemma10で$t = 0..T$とし、$f_0 = R$

とおけばよい　□

Example of FoReL: $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$

$f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle$かつ$S = R^d$で正則化項$R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2 \quad \text{where } \eta > 0$

この場合はFoReLは $\quad \mathbf{w}_{t+1} = \arg\min \sum_{i=1}^{t} \langle \mathbf{w}, \mathbf{z}_i \rangle + \frac{1}{2\eta}\|\mathbf{w}\|_2^2$

より、$0 = \frac{\partial}{\partial \mathbf{w}} \sum_{i=1}^{t} \langle \mathbf{w}, \mathbf{z}_i \rangle + \frac{1}{2\eta}\|\mathbf{w}\|_2^2 = \sum_{i=1}^{t} \mathbf{z}_i + \frac{\mathbf{w}}{\eta}$を使えば

$$\Rightarrow \quad \mathbf{w}_{t+1} = -\eta \sum_{i=1}^{t} \mathbf{z}_i = \mathbf{w}_t - \eta \mathbf{z}_t$$

つまり $\quad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t) \qquad (30)$

Online Gradient Descent: OGD

FoReLのRegretのUpper Bound

➤ Theorem 30

$$f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle, \quad S = R^d, \qquad R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2 \quad \text{とする}$$

$$\forall \mathbf{u} \quad \mathrm{Regret}_T(\mathbf{u}) \le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2$$

$$\text{if } U = \{\mathbf{u} : \|\mathbf{u}\| \le B\} \text{ and } \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2 \le L^2$$

$$\text{then } \eta = \frac{B}{L\sqrt{2T}} \quad \text{とおくと} \qquad \mathrm{Regret}_T(U) = \inf_{\mathbf{u} \in U} \mathrm{Regret}_T(\mathbf{u}) \le BL\sqrt{2T}$$

Proof Lemma 20と式(30)より

$$\mathrm{Regret}_T(\mathbf{u}) \le R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \right)$$

$$= \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t \rangle \quad = \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|_2^2$$

Regretの上限が$\sqrt{T}$に比例していることに注目！

損失fが連続でない場合 Sub-gradient(劣勾配）のFoReL

➢ fの凸性が重要

Lemma 50 Sが凸集合、fがS上の凸関数

$$\text{iff} \quad \forall \mathbf{w} \in S, \exists \mathbf{z} \forall \mathbf{u} \in S, \ f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{z} \rangle \qquad (50)$$

(50)を満たす$\mathbf{z}$の集合をfのsub-gradient　と呼び　$\partial f(\mathbf{w})$　と書く。
連続なら$\nabla f(\mathbf{w})$と同じ。
これを使った Online Gradient Descent が以下。

$$\eta > 0, \ \mathbf{w}_1 = 0,$$
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t \quad \text{where} \ \ \mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$$

Sub-gradient の場合のFoReLのRegret Bound

再掲：Lemma 20 and Theorem30：$\mathbf{w}_1, \mathbf{w}_2 \cdots$をFoReLで生成された重みベクトル

$$\forall \mathbf{u} \in S \quad \sum_{t=1}^{T}\left(f_t(\mathbf{w}_t) - f_t(\mathbf{u})\right) = \mathrm{Regret}_T(\mathbf{u}) \le R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T}\left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})\right)$$

$$= \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t \rangle \quad = \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta\sum_{t=1}^{T}\|\mathbf{z}_t\|_2^2 \qquad (60)$$

Lemma 50を少し変形して再掲：Sが凸集合、fがS上の凸関数

$$\text{iff} \quad \forall \mathbf{w}_t \in S, \exists \mathbf{z}_t \forall \mathbf{w}_{t+1} \in S, \quad f(\mathbf{w}_{t+1}) \ge f(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \mathbf{z}_t \rangle \qquad (50)$$

凸だと各round tに対して$f_t(\mathbf{w}_{t+1}) \ge f_t(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \mathbf{z}_t \rangle$　だから

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \le \langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t \rangle \Rightarrow \sum_{t=1}^{T}\left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})\right) \le \sum_{t=1}^{T}\left(\langle \mathbf{w}_t, \mathbf{z}_t \rangle - \langle \mathbf{w}_{t+1}, \mathbf{z}_t \rangle\right)$$

これをLemma 20 and Theorem30にplug inするとsub-gradient OGDでも

$$\mathrm{Regret}_T(\mathbf{u}) \le R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T}\left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})\right)$$

$$\le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{z}_t \rangle \quad = \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta\sum_{t=1}^{T}\|\mathbf{z}_t\|_2^2$$

問題はこの部分

$$\mathrm{Regret}_T(\mathbf{u}) \le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta\sum_{t=1}^{T}\|\mathbf{z}_t\|_2^2 \le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta T L^2$$ にしたいが

そのためには$\|\mathbf{z}_t\|$が上から押さえられることを示す必要がある⇒

sub-gradientの定義から

$$f(\mathbf{w}_t) - f(\mathbf{w}_{t+1}) \ge \langle \mathbf{z}_t, \mathbf{w}_t - \mathbf{w}_{t+1} \rangle \ \text{where } \mathbf{z}_t \in \partial f(\mathbf{w}_t)$$

fが$L-$Lipfshitzだとすると $L\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \ge f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})$ where $L < \infty$

上の２式を合わせると　$L\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \ge \langle \mathbf{z}_t, \mathbf{w}_t - \mathbf{w}_{t+1} \rangle$

$\eta\mathbf{z}_t = \mathbf{w}_t - \mathbf{w}_{t+1}$だったので$L\|\eta\mathbf{z}_t\| \ge \eta\|\mathbf{z}_t\|_2^2 \Rightarrow \infty > L \ge \|\mathbf{z}_t\|$

$$\Rightarrow \mathrm{Regret}_T(\mathbf{u}) \le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta T L^2 \le BL\sqrt{2T}$$ same as Theorem 30

このboundは$\frac{1}{\sqrt{2}}$にできる。
付録参照

FoReLの上界を厳しくする

- まず、FoReLの別形式を導入する

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{i=1}^{t} \langle \mathbf{w}, \mathbf{z}_i \rangle$$

$$= \arg\max_{\mathbf{w}} \left\langle \mathbf{w}, -\sum_{i=1}^{t} \mathbf{z}_i \right\rangle - R(\mathbf{w})$$

ここで $g(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - R(\mathbf{w})$ とおくとFoReLは次式で書ける

$$\begin{aligned}
&\text{initialize } \boldsymbol{\theta}_1 = 0 \\
&\text{for } t = 1, .., T \\
&\quad \mathbf{w}_t = g(\boldsymbol{\theta}_t); \\
&\quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t; \quad (\mathbf{z}_t \in \partial f_t(\mathbf{w}_t))
\end{aligned}$$

Online Mirror Descent (OMD)という

数学的ツールの準備

- Fenchel-Young 不等式

$f^*(\boldsymbol{\theta}) = \max_{\mathbf{u}} \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u})$とすると

$$\forall \mathbf{u}, \qquad f^*(\boldsymbol{\theta}) \geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u})$$

$\mathbf{u} \in \partial f^*(\boldsymbol{\theta})$　　あるいは$\boldsymbol{\theta}$で微分可能なら　$\mathbf{u} = \nabla f^*(\boldsymbol{\theta})$

$$\Rightarrow \qquad f^*(\boldsymbol{\theta}) = \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u})$$

$$R(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2\eta} \;\Rightarrow R^*(\boldsymbol{\theta}) = \max_{\mathbf{w}} \left(\langle \mathbf{w}, \boldsymbol{\theta} \rangle - \frac{\|\mathbf{w}\|^2}{2\eta} \right) = \frac{\eta \|\boldsymbol{\theta}\|^2}{2}$$

数学的ツールの準備

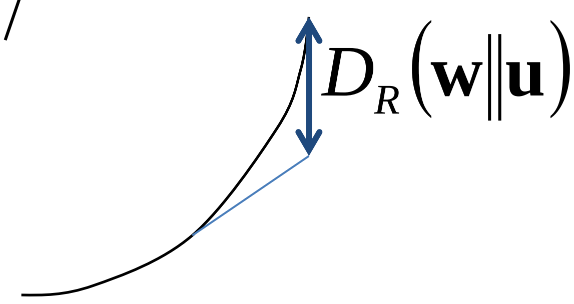
- Bregman Divergence: D_R

$$D_R(\mathbf{w}\|\mathbf{u}) = R(\mathbf{w}) - \left(R(\mathbf{u}) + \left\langle \nabla R(\mathbf{u}), \mathbf{w} - \mathbf{u} \right\rangle\right)$$

$$R^*(\mathbf{z}) = \frac{\eta\|\mathbf{z}\|^2}{2} \text{だと} \nabla R^*(\mathbf{z}) = \eta\|\mathbf{z}\|$$

$$D_{R^*}(\mathbf{z}_1\|\mathbf{z}_2) = \frac{\eta}{2}\|\mathbf{z}_1\|^2 - \frac{\eta}{2}\|\mathbf{z}_2\|^2 - \left\langle \eta\|\mathbf{z}_2\|, \mathbf{z}_1 - \mathbf{z}_2 \right\rangle$$

$$= \frac{\eta}{2}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 \qquad (DR1)$$



補題　OML1:　Online　Mirror　Descent で　$g = \nabla R^*$であるなら

$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle \le R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} D_{R^*}\left(-\sum_{i=1}^{t} \mathbf{z}_i \,\|\, -\sum_{i=1}^{t-1} \mathbf{z}_i \right)$$

Proof

Fenchel - Young 不等式から

$$R(\mathbf{u}) + \sum_{t=1}^{T} \langle \mathbf{u}, \mathbf{z}_t \rangle = R(\mathbf{u}) - \left\langle \mathbf{u}, -\sum_{t=1}^{T} \mathbf{z}_t \right\rangle \ge -R^*\left(-\sum_{t=1}^{T} \mathbf{z}_t \right) \qquad (1)$$

$$\mathbf{w}_t = g(\boldsymbol{\theta}_t) = \nabla R^*\left(-\sum_{i=1}^{t-1} \mathbf{z}_i \right) \qquad (2)$$

Bregman Divergenceの定義より

$$-R^*\left(-\sum_{t=1}^{T} \mathbf{z}_t \right) = -R^*(0) - \sum_{t=1}^{T} \left(R^*\left(-\sum_{i=1}^{t} \mathbf{z}_i \right) - R^*\left(-\sum_{i=1}^{t-1} \mathbf{z}_i \right) \right)$$

$$= -R^*(0) + \sum_{t=1}^{T} \left(\langle \mathbf{w}_t, \mathbf{z}_t \rangle - D_{R^*}\left(-\sum_{i=1}^{t} \mathbf{z}_i \,\|\, -\sum_{i=1}^{t-1} \mathbf{z}_i \right) \right) \qquad (3)$$

なお、$R^*(0) = \max_{\mathbf{w}} \langle 0, \mathbf{w} \rangle - R(\mathbf{w}) = -\min_{\mathbf{w}} R(\mathbf{w}) = -R(\mathbf{w}_1)$　(4) これらを合わせると

$$-\sum_{t=1}^{T} \langle \mathbf{u}, \mathbf{z}_t \rangle \le R(\mathbf{u}) + R^*\left(-\sum_{t=1}^{T} \mathbf{z}_t \right) = R(\mathbf{u}) - R^*(0) + \sum_{t=1}^{T} \left(\langle \mathbf{w}_t, \mathbf{z}_t \rangle - D_{R^*}\left(-\sum_{i=1}^{t} \mathbf{z}_i \,\|\, -\sum_{i=1}^{t-1} \mathbf{z}_i \right) \right)$$

$$= R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} \left(-\langle \mathbf{w}_t, \mathbf{z}_t \rangle + D_{R^*}\left(-\sum_{i=1}^{t} \mathbf{z}_i \,\|\, -\sum_{i=1}^{t-1} \mathbf{z}_i \right) \right)$$

定理 OMD2

Online Mirror Descentにおいて

$$R(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2\eta}, \left(\text{つまり} R(\boldsymbol{\theta})^* = \frac{\eta\|\boldsymbol{\theta}\|^2}{2} \text{である} \right) \text{とする。}$$

$$\Rightarrow \sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle \le \frac{\|\mathbf{u}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{z}_t\|^2 \qquad (OMD10)$$

Proof 補題OMD1と$(DR1)$より明らか ■

$$L^2 = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{z}_t\|^2 \quad \text{かつ} \quad \eta = \frac{B}{L\sqrt{T}} \text{ただし} \|\mathbf{u}\| \le B \quad \text{とすると}$$

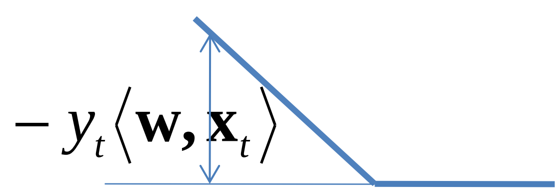
$$\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle \le BL\sqrt{T} \qquad (OMD20)$$

パーセプトロン(Perceptron)

- FoReLから導出されたOnline Gradient Descentの例としてパーセプトロンを紹介する。
- パーセプトロンはF. Rosenblattが1956年に提案した線形識別の繰り返しアルゴリズム
- 効率がよく、現在でもその価値が高い

- 入力x_tが目的のクラスに
  - 属する場合に$y_t=1$, 属さない場合に$y_t=-1$

- $f_t(\mathbf{w}) = \left[-y_t\langle \mathbf{w},\mathbf{x}_t\rangle\right]_+$ 右図
- $f_t(0)=0$ これより失敗側では正



- $f_t(w)$のsub-gradientを計算すると

$$\mathbf{z}_t \in \partial f_t(\mathbf{w}_t) \Rightarrow \text{ if } y_t\langle \mathbf{w}_t, \mathbf{x}_t \rangle > 0 \text{ then } \nabla f_t(\mathbf{w}_t) = \mathbf{z}_t = 0$$

$$\text{otherwise } \nabla f_t(\mathbf{w}_t) = \mathbf{z}_t = -y_t \mathbf{x}_t \in \partial f_t(\mathbf{w}_t)$$

Online Gradient Descentの形式：$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t$ に当てはめると

$$\mathbf{w}_{t+1} = \begin{cases} \mathbf{w}_t & \text{if } y_t\langle \mathbf{w}_t, \mathbf{x}_t \rangle > 0 \\ \mathbf{w}_t + \eta y_t \mathbf{x}_t & \text{otherwise} \end{cases}$$

- ηは正
- 次にPerceptronのアルゴリズムが得られる。
- FoReLの別形式として導入したOnline Mirror Descentとみれば、(OMD20)の上界が使える

Perceptron アルゴリズム

初期化：$\mathbf{w}_1 = 0$

for $t = 1,2,\ldots,T$

入力= $\mathbf{x}_t$

if $y_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle \le 0$

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

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

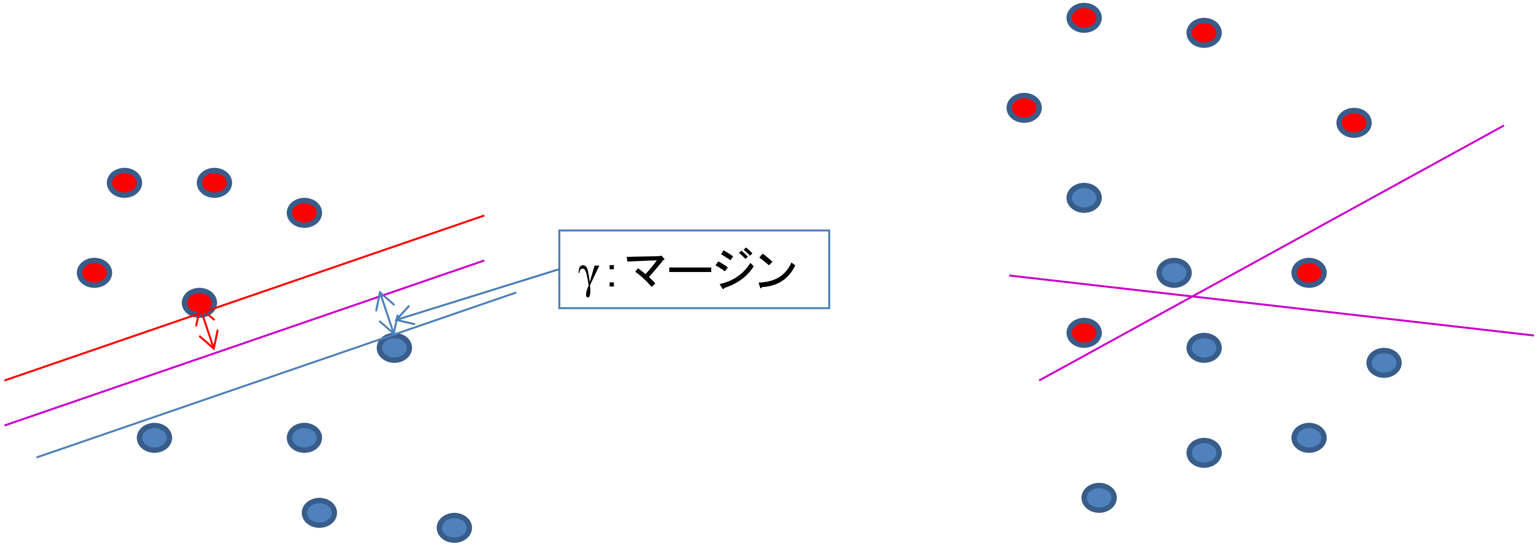
分類に失敗したときだけ

そのデータを分類器Wに足し込むという至って単純な更新

線形分離可能性

- 線形分離可能:クラスを識別するする超平面が存在する場合
  - そうでない場合を線形分離不可能という。
  - 下図参照

線形分離可能　　　　線形分離不可能



Perceptronアルゴリズムの分析

FoReLの別形式として導入したOnline Mirror Descentとみれば、(OMD20)の上界が使える

$$\mathrm{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{z}_t\|_2^2 \qquad (60)$$

ここで解析の容易さのため$f_t = [1 - y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle]_+$で近似する。

（もとのf_tより必ず大きいので上界は甘くなる。）

sub - gradientは　$\mathbf{z}_t = -\mathbf{1}_{[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]} y_t \mathbf{x}_t$

Wt,Xtをスケール変換して、最も0に近い正例で$\langle w_t, x_t\rangle$ =1となったと見なしたと考えてもよい

Perceptronアルゴリズムの分析

判定の失敗 : mistakeを起こした$\mathbf{x}_t$の集合を$\mathcal{M}$、mistake回数$=|\mathcal{M}|$

$\Rightarrow$ f_tの形より明らかに$\sum_{t=1}^{T} f_t(\mathbf{w}_t) \geq |\mathcal{M}|$　$R = \max_t \|\mathbf{x}_t\|$とおくと

(60)の右辺の最小値は$\eta = \dfrac{\|\mathbf{u}\|}{\sqrt{|\mathcal{M}|}R}$のときなので

$$|\mathcal{M}| - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \sqrt{|\mathcal{M}|} R \|\mathbf{u}\| \leq \frac{1}{2\eta}\|\mathbf{u}\|^2 + \frac{\eta}{2}|\mathcal{M}|R^2 \qquad (70)$$

if $\exists \mathbf{u}\ \ y_t \langle \mathbf{u}, \mathbf{x}_t \rangle \geq 1$ then $\sum_{t=1}^{T} f_t(\mathbf{u}) = 0$: データが線形分離可能

$$\Rightarrow |\mathcal{M}| \leq \sqrt{|\mathcal{M}|} R \|\mathbf{u}\| \Rightarrow |\mathcal{M}| \leq R^2 \|\mathbf{u}\|^2 = \frac{R^2}{\gamma^2}$$

$\Rightarrow$ 線形分離できる場合$\dfrac{1}{\|\mathbf{u}\|}$は境界面に最も近いデータと境界面の距離$\gamma$

Passive Aggressive Algorithm

- K.Crammer et.al. Online Passive-Aggressive Algorithms Journal of Machine Learning Research 7 (2006) 551–585

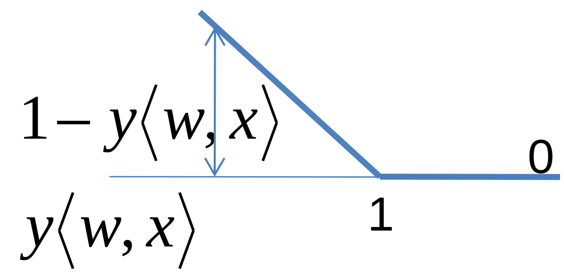
- 識別（あるいは分類）の問題設定
  - round t でn次元データ　$x_t \in R^n$　が到着
  - x_tの正負は $y_t = \{+1:\text{正}, -1:\text{負}\}$のように与えられる
  - 重みベクトル:$w \in R^n \Rightarrow sign\langle w, x\rangle$: 正負を表すので
  - 正しい（誤った）判定: $y_t\langle w_t, x_t\rangle > 0$　、(< 0)

  - w_tはデータが到着するたびに更新されている

損失関数による定式化

- 境界面そのもので判定はきわどいのでマージンを持たせる。マージンを1とした場合の損失関数(hinge-loss function)は以下の通り

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



以下では$\ell_t = \ell(\mathbf{w}_t;(\mathbf{x}_t, y_t))$　と書く

- この設定で、round tの更新は次の条件付き最適化問題となる。

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

FoReLとして見ると

FoReL

$$\forall t \ \ \mathbf{w}_t = \arg\min_{w \in S} \sum_{i=1}^{t-1} f(\mathbf{w})_i + R(\mathbf{w}) \qquad S\text{は}w\text{の取り得る範囲で凸}$$

$$= \arg\min_{w \in S} \sum_{i=1}^{t-1} \ell(\mathbf{w}_i;(\mathbf{x}_i, y_i)) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2$$

次ページのようにFoReLの定式化ではηに相当するτは個別の$y_t, \mathbf{x}_t, \mathbf{w}_t$に依存するため

(60)のような簡単な解析ができない。

最適化問題(PA-1)を解く

- If ℓ_t=0 then w_t minimizes $\frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|^2$

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

Passive

- If $\ell_t \neq 0$ then Lagrange 未定乗数法で解く。

$$L(\mathbf{w},\tau) = \frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|^2 + \tau\left(1 - y_t\langle \underline{\mathbf{w}}, \mathbf{x}_t\rangle\right)$$

$$\frac{\partial L(\mathbf{w},\tau)}{\partial \mathbf{w}} = \mathbf{w} - \mathbf{w}_t - \tau y_t \mathbf{x}_t = 0$$

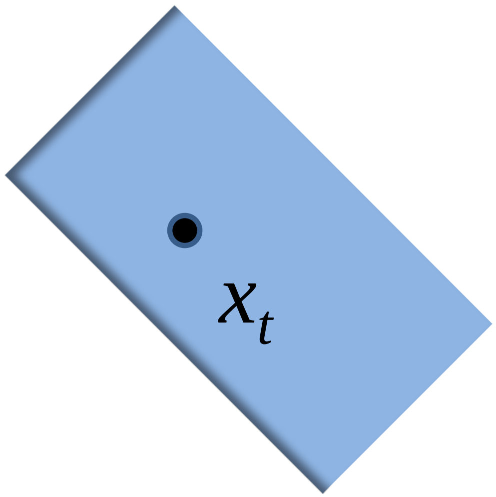
$$\Rightarrow \quad \mathbf{w} = \mathbf{w}_t + \tau y_t \mathbf{x}_t \qquad \Rightarrow \quad L(\tau) = -\frac{1}{2}\tau^2\|\mathbf{x}_t\|^2 + \tau\left(1 - y_t\langle \underline{\mathbf{w}_t}, \mathbf{x}_t\rangle\right)$$

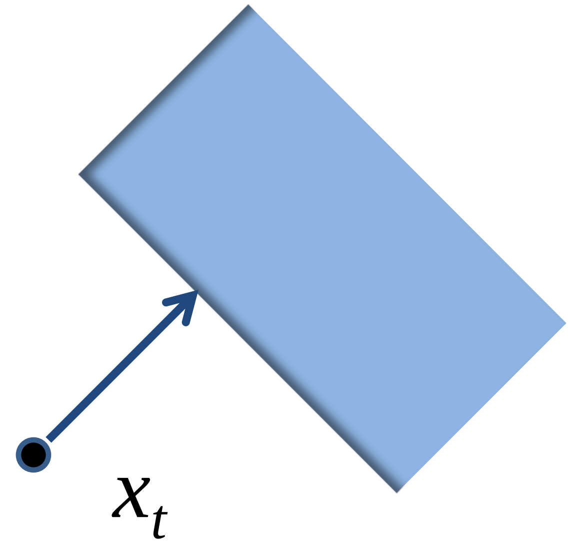
$$\Rightarrow \quad \frac{\partial L(\tau)}{\partial \tau} = -\tau\|\mathbf{x}_t\|^2 + \left(1 - y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle\right) = 0$$

$$\Rightarrow \quad \boxed{\tau_t = \frac{1 - y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle}{\|\mathbf{x}_t\|^2} \qquad \mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t}$$

Aggressive

Passive Aggressive





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

Passive

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

Aggressive

$$\tau_t = \frac{\ell_t}{\left\|\mathbf{x}_t\right\|^2}$$

soft marginの学習法 PA-I, PA-II

$$\mathbf{w}_{t+1} = \arg\min_{w \in R^n} \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}_t\right\|^2 + C\xi \quad s.t. \quad \ell(\mathbf{w};(\mathbf{x}_t, y_t)) \le \xi, \quad \xi \ge 0 \quad (PA-I)$$

$$\mathbf{w}_{t+1} = \arg\min_{w \in R^n} \frac{1}{2}\left\|\mathbf{w} - \mathbf{w}_t\right\|^2 + C\xi^2 \quad s.t. \quad \ell(\mathbf{w};(\mathbf{x}_t, y_t)) \le \xi \quad (PA-II)$$

$\Rightarrow$

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

$$\tau_t = \min\left\{C, \frac{\ell_t}{\left\|\mathbf{x}_t\right\|^2}\right\} \quad (PA-I) \qquad \tau_t = \frac{\ell_t}{\left\|\mathbf{x}_t\right\|^2 + \frac{1}{2C}} \quad (PA-II)$$

Passive Aggressive Algorithm

INPUT: aggressiveness parameter $C > 0$

INITIALIZE: $\mathbf{w_1 = (0, . . . ,0)}$

For $t = 1,2, . . .$

- receive instance: $\mathbf{x}_{\boldsymbol{t}} \in \mathrm{R}^{\boldsymbol{n}}$
- predict: $\hat{y}_t = \mathrm{sign}\langle \mathbf{w}_t, \mathbf{x}_t \rangle$
- receive correct label: $y_t \in \{-1,+1\}$
- suffer loss: $\ell_t = \max\{0\ ,\ 1 - y_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle\}$
- update:

1. set:

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

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

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

2. update: $\mathbf{w}_{\boldsymbol{t+1}} = \boldsymbol{w_t} + \tau_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle$

付録：PA-Iの導出

$PA-I$の$Lagrangian$は以下の通り

$$L(w,\xi,\tau,\lambda)=\frac{1}{2}\|w-w_t\|^2+C\xi+\tau(1-\xi-y(w\cdot x_t))-\lambda\xi$$

$$=\frac{1}{2}\|w-w_t\|^2+\xi(C-\tau-\lambda)+\tau(1-y(w\cdot x_t)) \qquad \tau\geq 0, \qquad \lambda\geq 0 \qquad (pa1-1)$$

$$\frac{\partial L}{\partial w}=0\Rightarrow w=w_t+\tau y_t x_t$$

$\xi(C-\tau-\lambda)$の最小値は0で　　$C-\tau-\lambda=0 \quad (pa1-2)$のとき。

$C-\tau-\lambda\geq 0$である。そうでないとすると、　$\xi(C-\tau-\lambda)$はいくらでも小さくなれるので、Lの最小化ができない。

KKT条件より$\lambda\geq 0$なので、$(pa1-2)$より$C-\tau\geq 0\Rightarrow C\geq\tau \quad (pa1-3)$

以下では$C\geq \ell_t/\|x_t\|^2$ (case 1)と$C<\ell_t/\|x_t\|^2$ (case 2)に分けて考える。

case 1

$(pa1-2)$を$(pa1-1)$に代入し

$L(w,\xi,\tau,\lambda)=\frac{1}{2}\|w-w_t\|^2+\tau(1-y(w\cdot x_t))$となるので、これを$w$について最適化すると

元々のPAと同じく　$\tau_t=\ell_t/\|x_t\|^2 \quad (pa1-4)$

$(\text{case } 2) \quad \ell_t / \|x_t\|^2 > C \quad \Rightarrow \quad C\|x_t\|^2 < 1 - y_t \langle w_t, x_t \rangle \quad (pa1-6)$

元々の$optimazation$

$$w_{t+1} = \arg\min_{w} \frac{1}{2}\|w - w_t\|^2 + C\xi \quad s.t \quad 1 - y_t \langle w\ , x_t \rangle \le \xi \quad (pa1-7) \quad \text{and } \xi \ge 0$$

と　$w = w_t + \tau y_t x_t$　により$1 - y_t \langle w_t, x_t \rangle - \tau \|x_t\|^2 \le \xi \quad (pa1-8)$

$(pa1-8)$と$(pa1-6)$を組み合わせると　$C\|x_t\|^2 - \tau\|x_t\|^2 < \xi$

$(pa1-3)$で$C \ge \tau$だったから、$0 < \xi$

KKT条件から$\lambda\xi = 0$なので、$\lambda = 0 \quad \Rightarrow (pa1-2)$より　$\tau = C$

(case 1) (case 2)を合せると　$\tau_t = \min\left\{ C, \dfrac{\ell_t}{\|x_t\|^2} \right\}$

付録：PA-IIの導出

$\ell_t = 0 \Rightarrow \tau_t = 0 \Rightarrow \ell_t > 0$の場合について考えればよい

$$L(w,\xi,\tau) = \frac{1}{2}\|w - w_t\|^2 + C\xi^2 + \tau\left(1 - \xi - y\left(\underline{w} \cdot x_t\right)\right) \quad \tau \geq 0 \qquad (pa2-1)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = w_t + \tau y_t x_t$$

$$\frac{\partial L}{\partial \xi} = 2C\xi - \tau = 0 \Rightarrow \xi = \tau/2C \Rightarrow$$ このξとwをLに代入すると

$$L(w,\tau) = -\frac{\tau^2}{2}\left(\|x_t\|^2 + \frac{1}{2C}\right) + \tau\left(1 - y_t\left(\underline{w_t} \cdot x_t\right)\right)$$

$$\frac{\partial L}{\partial \tau} = 0 \Rightarrow \tau = \frac{1 - y_t\left(w_t \cdot x_t\right)}{\|x_t\|^2 + \frac{1}{2C}} = \frac{\ell_t}{\|x_t\|^2 + \frac{1}{2C}} \qquad \blacksquare$$

損失の限界の評価

任意の固定された重みベクトルuに対する損失をl_t^*とする。

$l_t = l(w_t;(x_t, y_t))$ $\qquad$ $l_t^* = l(u;(x_t, y_t))$

Lemma 1

$(x_1, y_1), ..., (x_T, y_T)$ はデータ列。ただし $x_t \in R^n, y_t \in \{+1, -1\}$

τ_tはPA, PA - I, PA - IIのAlgorithmにおける更新式のparameter

上記の $u \in R^n$に対して

$$\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x_t\|^2 - 2\ell_t^*\right) \le \|u\|^2$$

Proof

$$\Delta_t \stackrel{def}{=} \|w_t - u\|^2 - \|w_{t+1} - u\|^2$$

$$\sum_{t=1}^{T} \Delta_t = \sum_{t=1}^{T} \left(\|w_t - u\|^2 - \|w_{t+1} - u\|^2 \right) = \|w_1 - u\|^2 - \|w_{T+1} - u\|^2$$

$w_1 = 0$　　$\|w_{T+1} - u\|^2 \geq 0$　　なので

$$\sum_{t=1}^{T} \Delta_t \leq \|u\|^2$$

minimum marginをviolateしない場合は、$\ell_t = 0, \tau_t = 0 \quad \Rightarrow \quad \Delta_t = 0$

よって、$\ell_t > 0$の場合にだけ注目：$w_{t+1} = w_t + y_t \tau_t x_t$

$$\Delta_t = \|w_t - u\|^2 - \|w_{t+1} - u\|^2 = \|w_t - u\|^2 - \|w_t - u + y_t \tau_t x_t\|^2$$

$$= \|w_t - u\|^2 - \left(\|w_t - u\|^2 + 2\tau_t y_t \langle (w_t - u), x_t \rangle + \tau_t^2 \|x_t\|^2 \right)$$

$$= -2\tau_t y_t \langle (w_t - u), x_t \rangle \cdot x_t - \tau_t^2 \|x_t\|^2$$

$\ell_t > 0$の場合　$\ell_t = 1 - y_t \langle w_t, x_t \rangle \quad \Rightarrow \quad y_t \langle w_t, x_t \rangle = 1 - \ell_t$

同じく　$y_t \langle u, x_t \rangle = 1 - \ell_t^*$

$$\Rightarrow \quad \Delta_t = -2\tau_t y_t \langle (w_t - u), x_t \rangle - \tau_t^2 \|x_t\|^2$$

$$\geq 2\tau_t \left(\left(1 - \ell_t^*\right) - \left(1 - \ell_t\right) \right) - \tau_t^2 \|x_t\|^2$$

$$= \tau_t \left(2\ell_t - \tau_t \|x_t\|^2 - 2\ell_t^* \right) \qquad \blacksquare$$

Theorem 2

Lemma1 と同じ設定。$\exists u \quad s.t. \quad \forall t\left[\ell_t^* = 0\right],\ \max_t \|x_t\| \le R$

$$\Rightarrow \quad \sum_{t=1}^{T} \ell_t^2 \le \|u\|^2 R^2$$

Proof

$$\forall t\left[\ell_t^* = 0\right] \text{and Lemma1} \quad \Rightarrow \quad \sum_{t=1}^{T} \tau_t\left(2\ell_t - \tau_t \|x_t\|^2\right) \le \|u\|^2$$

PAでは $\quad \tau_t = \ell_t / \|x_t\|^2 \quad \Rightarrow \quad \sum_{t=1}^{T} \ell_t^2 / \|x_t\|^2 \le \|u\|^2$

$$\forall t\left[\|x_t\|^2 \le R^2\right] \quad \Rightarrow \quad \sum_{t=1}^{T} \ell_t^2 / R^2 \le \|u\|^2 \quad \Rightarrow \quad \sum_{t=1}^{T} \ell_t^2 \le \|u\|^2 R^2 \quad \blacksquare$$

- Theorem 2では次の制約が厳しい。 $\exists u \quad s.t. \quad \forall t\left[l_t^* = 0\right]$
- この制約は、uで完全な識別ができること。
- この制約を外す定理を考える

Theorem 3

Lemma1　と同じ設定。

$\forall t \quad \|x_t\|^2 = 1$ 　このとき$\ell_t^* = \ell(u;(x_t, y_t))$であるような$u \in R^n$に対して

$$\sum_{t=1}^{T} \ell_t^2 \leq \left(\|u\| + 2\sqrt{\sum_{t=1}^{T} \left(\ell_t^*\right)^2} \right)^2$$

Proofは次ページ

Proof

$$\|x_t\|^2 = 1 \quad \Rightarrow \quad \tau_t = l_t$$

$$\therefore \sum_{t=1}^{T} \tau_t \left(2l_t - \tau_t \|x_t\|^2 - 2l_t^*\right) \le \|u\|^2 \quad \Rightarrow \quad \sum_{t=1}^{T} l_t^2 \le 2\sum_{t=1}^{T} l_t l_t^* + \|u\|^2$$

Cauchy - Schwartzより $\displaystyle\sum_{t=1}^{T} l_t l_t^* \le \sqrt{\sum_{t=1}^{T} l_t^2} \cdot \sqrt{\sum_{t=1}^{T} \left(l_t^*\right)^2}$ だから

$$\sum_{t=1}^{T} l_t^2 \le 2\sqrt{\sum_{t=1}^{T} l_t^2} \cdot \sqrt{\sum_{t=1}^{T} \left(l_t^*\right)^2} + \|u\|^2$$

ここで $\displaystyle\sqrt{\sum_{t=1}^{T} l_t^2} = LT$ とおくと

$$LT^2 - 2LT \cdot \sqrt{\sum_{t=1}^{T} \left(l_t^*\right)^2} - \|u\|^2 \le 0$$

LTの最大値$\max LT$は上式で$\le$を$=$とした２次式の大きなほうの解。

$$\max LT = \sqrt{\sum_{t=1}^{T} \left(\ell_t^*\right)^2} + \sqrt{\sum_{t=1}^{T} \left(\ell_t^*\right)^2 + \|u\|^2}$$

$\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$を使うと $\displaystyle\max LT = \max \sqrt{\sum_{t=1}^{T} \ell_t^2} \le 2\sqrt{\sum_{t=1}^{T} \left(\ell_t^*\right)^2} + \|u\|$

$$\Rightarrow \quad \max \sum_{t=1}^{T} \ell_t^2 \le \left(2\sqrt{\sum_{t=1}^{T} \left(\ell_t^*\right)^2} + \|u\|\right)^2 \qquad \blacksquare$$

PA-Iにおける入力データ識別の失敗回数の上限

$$\|x_t\|^2 \leq R^2 \qquad \tau_t = \min\left\{\frac{\ell_t}{\|x_t\|^2}, C\right\}$$

$$\Rightarrow \# mistakes \leq \max\{R^2, 1/C\}\left(\|u\|^2 + 2C\sum_{t=1}^{T}\ell_t^*\right)$$

Proofは次のページ

t回目の繰り返しで*mistake*が起きたとすると$\ell_t \geq 1$

$\|x_t\|^2 \leq R^2$と $\tau_t = \min\left\{\ell_t / \|x_t\|^2, C\right\}$より $\min\left\{1/R^2, C\right\} \leq \tau_t \ell_t$

M を繰り返し全体における *mistake*回数とする。

$0 \leq \tau_t \ell_t$ なので $\min\left\{1/R^2, C\right\} M \leq \sum_{t=1}^{T} \tau_t \ell_t$ (pa1-10)

$\ell_t^* = \ell(u;(x_t, y_t)) \Rightarrow \tau_t \ell_t^* \leq C \ell_t^*$かつ$\tau_t \|x_t\|^2 \leq \ell_t$

これをLemma1 $\sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x_t\|^2 - 2\ell_t^*\right) \leq \|u\|^2$に代入すると

$$\sum_{t=1}^{T} \tau_t \ell_t \leq \|u\|^2 + 2C \sum_{t=1}^{T} \ell_t^* \quad \text{(pa1-20)}$$

(pa1-10)(pa1-20) より $\min\left\{1/R^2, C\right\} M \leq \|u\|^2 + 2C \sum_{t=1}^{T} \ell_t^*$ (pa1-30)

(pa1-30)の両辺に$\max\left\{R^2, 1/C\right\}$を掛けると

$$M \leq \max\left\{R^2, 1/C\right\} \left(\|u\|^2 + 2C \sum_{t=1}^{T} \ell_t^* \right) \qquad \blacksquare$$

PA-IIにおける累積損失の上限

$$\|x_t\|^2 \le R^2 \qquad \tau_t = \frac{\ell_t}{\|x_t\|^2 + \frac{1}{2C}}$$

$$\Rightarrow \quad \sum_{t=1}^{T} \ell_t^2 \le \left(R^2 + \frac{1}{2C} \right) \left(\|u\|^2 + 2C \sum_{t=1}^{T} \ell_t^{*} \right)$$

Proofは次のページ

Lemma1: $\|u\|^2 \geq \sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x_t\|^2 - 2\ell_t^*\right)$で右辺のΣ内で$\left(\alpha\tau_t - \ell_t^*/\alpha\right)^2$を差し引く

ただし　　$\alpha = 1/\sqrt{2C}$

$$\|u\|^2 \geq \sum_{t=1}^{T} \tau_t \left(2\ell_t - \tau_t \|x_t\|^2 - 2\ell_t^* - \left(\alpha\tau_t - \ell_t^*/\alpha\right)^2\right)$$

$$= \sum_{t=1}^{T} \left(2\tau_t\ell_t - \tau_t^{\,2}\|x_t\|^2 - 2\tau_t\ell_t^* - \alpha^2\tau_t^2 + 2\tau_t\ell_t^* - \ell_t^{*2}/\alpha^2\right)$$

$$= \sum_{t=1}^{T} \left(2\tau_t\ell_t - \tau_t^{\,2}\left(\|x_t\|^2 + \alpha^2\right) - \ell_t^{*2}/\alpha^2\right)$$

$\alpha = 1/\sqrt{2C}$を代入すると

$$\|u\|^2 \geq \sum_{t=1}^{T} \left(2\tau_t\ell_t - \tau_t^{\,2}\left(\|x_t\|^2 + \frac{1}{2C} \right) - 2C\ell_t^{*2} \right)$$

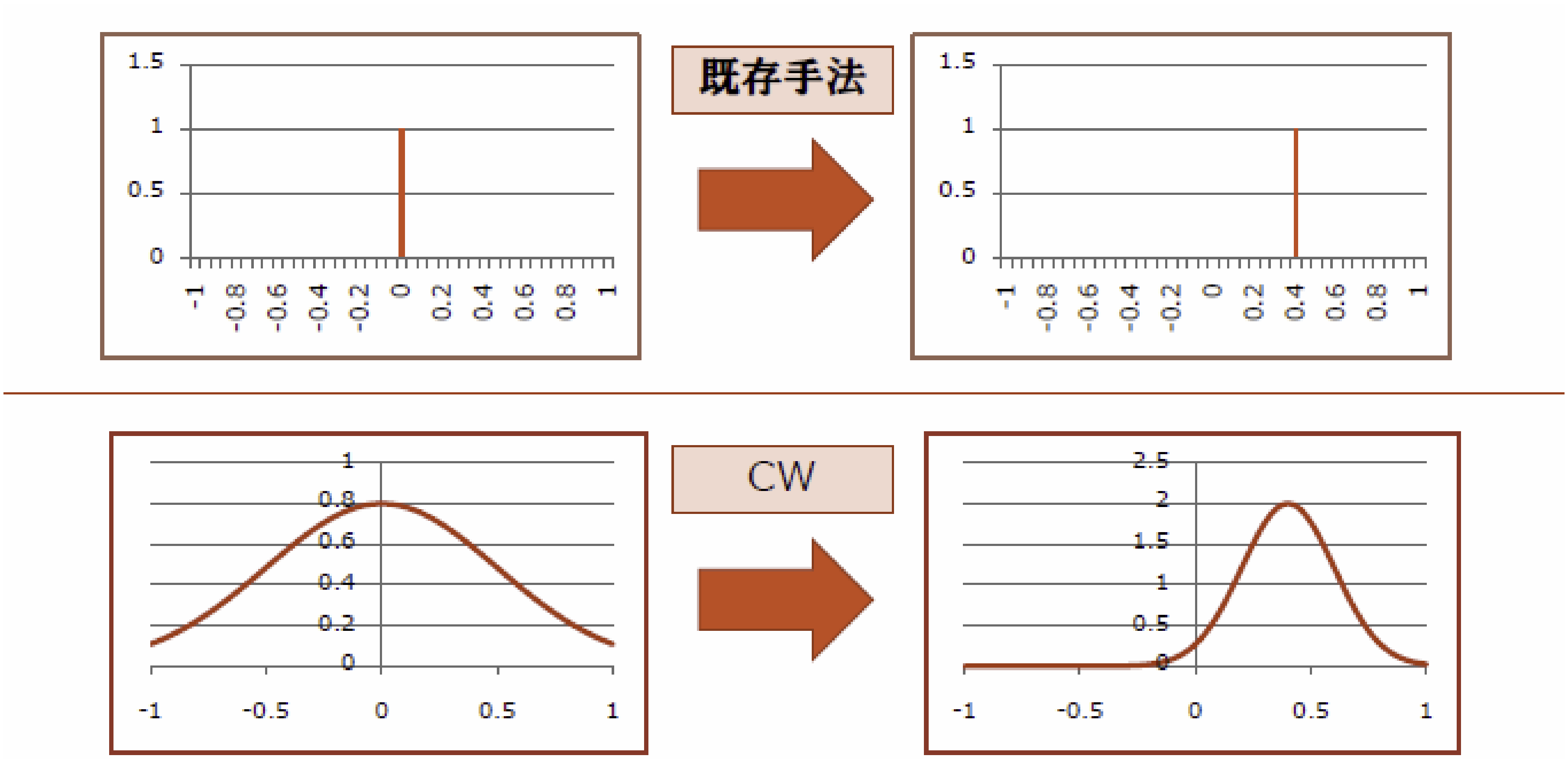
$\tau_t = \ell_t / \left(\|x_t\|^2 + 1/(2C)\right)$を代入すると$\|u\|^2 \geq \sum_{t=1}^{T} \left(\frac{\ell_t^{\,2}}{\|x_t\|^2 + \frac{1}{2C}} - 2C\ell_t^{*2} \right)$

$\|x_t\|^2 \leq R^2$を使えば、$\sum_{t=1}^{T} \ell_t^2 \leq \left(R^2 + \frac{1}{2C} \right) \left(\|u\|^2 + 2C \sum_{t=1}^{T} \ell_t^* \right)$ ■

Confidence Weighted Algorithm

Crammer et al. 2008

学習する重みベクトルWを点ではなく分布（正規分布）にする
→ Wの期待値と分散を更新する



Pegasos：
Primal Estimated sub-**G**r**A**dient**SO**lver for **S**VM

- L2正則化＋L1損失のSVM

$$\min_{\mathbf{w}} f(\mathbf{w};B) \equiv \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{k}\sum_{i\in B}\max\left(0,1-y_i\mathbf{w}^T\mathbf{x}\right) \quad A\text{は学習に使うデータ、} k=|A|$$

- Pegasosの更新は次式による

$$\mathbf{w}_{t+1/2} \leftarrow \mathbf{w}_t - \eta\nabla f(\mathbf{w}_t;A) \qquad \nabla f(\mathbf{w}_t;A)\text{は}f\text{の劣微分}\partial f(\mathbf{w}_t;A)\text{の要素}$$

$$where \quad \nabla f(\mathbf{w};A) = \lambda\mathbf{w}_t - \frac{1}{|A|}\sum_{i\in A^+} y_i\mathbf{x}_i \qquad A^+ \equiv \left\{i \mid i\in A, 1-y_i\mathbf{w}^T\mathbf{x}_i > 0\right\},$$

$$\eta = \frac{1}{\lambda t}, \quad l\text{はベクトル}\mathbf{w}\text{，}\mathbf{x}_i\text{の次元、}t\text{は繰り返し回数}$$

- 更新の後、wを半径 $\min\left(1,\sqrt{1/\lambda}\right)$ の球にproject

$$\boxed{\mathbf{w} \leftarrow \min\left(1, \sqrt{1/\lambda}\big/\|\mathbf{w}\|_2\right)\mathbf{w}}$$

- 以上を収束するまで繰り返す。データ集合Aごとなので、onlineというよりはmini-batch

Pegasos：**Primal Estimated sub-GrAdient SOlver for SVM** のアルゴリズム

初期化：$\mathbf{w}_1 = 0$

For $\mathrm{t} = 1,2,......,\mathrm{T}$

全データDからA_tを選ぶ。$|A_t| = k$

$$A_t^+ = \{ i \in A_t \mid 1 - y_i \langle \mathbf{w}_t, \mathbf{x}_i \rangle > 0 \}$$

$$\eta^{(t)} = \frac{1}{\lambda t}$$

$$\mathbf{w}_{t+1/2} = (1 - \eta^{(t)} \lambda) \mathbf{w}_t + \frac{\eta^{(t)}}{|A_t|} \sum_{i \in A_t^+} y_i \mathbf{x}_i$$

$$\mathbf{w}_{t+1} = \min \left\{ 1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{t+1/2}\|} \right\} \mathbf{w}_{t+1/2}$$

Output：$\mathbf{w}_{T+1}$

$f(\mathrm{w})$の評価

まず　$\mathbf{w}^{*} = \arg\min_{f(\mathbf{w})} f(\mathbf{w})$　とする。

また、関数fは、$f(\mathbf{w}) - \frac{\lambda}{2}\|\mathbf{w}\|^{2}$が凸のとき、$\lambda-$strongly convex　という。

Lemma1. $f_1,\ldots,f_T$をλ-strongly convexとし、Bを閉凸集合とする。
$\mathbf{w}_1,\ldots,\mathbf{w}_{T+1}$を以下のようなベクトルの列とする：$\nabla_t$は$f_t$の劣微分の要素

$\mathbf{w}_1 \in B,\quad \forall t \geq 1\ \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}' \in B} \|\mathbf{w}_t - \eta_t \nabla_t(\mathbf{w}_t) - \mathbf{w}'\|$ただし$\eta_t = 1/(\lambda t)$

$\|\nabla_t(\mathbf{w}_t)\| \leq G$のとき、$\forall \mathbf{u} \in B$に対して次式が成り立つ

$$\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{w}_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{u}) + \frac{G^2(1+\ln(T))}{2\lambda T}$$

Proof： f_t はstrongly-convexで、∇_tは劣微分の要素なので

内積$(\mathbf{w}_t - \mathbf{u} \cdot \nabla_t) \geq f_t(\mathbf{w}_t) - f_t(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{w}_t - \mathbf{u}\|^2 \qquad (10)$

$\mathbf{w}'_t = \mathbf{w}_t - \eta_t \nabla_t$ 　とおき、$\mathbf{w}_{t+1}$は$\mathbf{w}'_t$のBへの$projection$であり

$\mathbf{u} \in B$ 　なので　$\|\mathbf{w}'_t - \mathbf{u}\|^2 \geq \|\mathbf{w}_{t+1} - \mathbf{u}\|^2$ 　である。よって

$$\|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 \geq \|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}'_t - \mathbf{u}\|^2 = 2\eta_t(\mathbf{w}_t - \mathbf{u} \cdot \nabla_t) - \eta_t^2 \|\nabla_t\|^2 \qquad (15)$$

$\|\nabla_t\| \leq G$かつ$\eta_t = 1/(\lambda t)$なので(15)により

$$\langle \mathbf{w}_t - \mathbf{u}, \nabla_t \rangle \leq \frac{\|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{u}\|^2}{2\eta_t} + \frac{\eta_t}{2} G^2 \qquad (20)$$

(10)と(20)を組み合わせると $f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leq \frac{\|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{u}\|^2}{2\eta_t} + \frac{\eta_t}{2} G^2 - \frac{\lambda}{2}\|\mathbf{w}_t - \mathbf{u}\|^2 \qquad (30)$

(30)を$t = 1, T$まで総和をとると

$$\sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T} \left(\frac{\|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{u}\|^2}{2\eta_t} - \frac{\lambda}{2}\|\mathbf{w}_t - \mathbf{u}\|^2 \right) + \sum_{t=1}^{T} \frac{\eta_t}{2} G^2 \qquad (40)$$

$\eta_t = 1/(\lambda t)$を代入すると$(40) = \sum_{t=1}^{T} \left(\lambda t \left(\frac{\|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}_{t+1} - \mathbf{u}\|^2}{2} \right) - \frac{\lambda}{2}\|\mathbf{w}_t - \mathbf{u}\|^2 \right) + \frac{G^2}{2} \sum_{t=1}^{T} \frac{1}{\lambda t}$

$$= \sum_{t=1}^{T} \lambda \left((t-1)\|\mathbf{w}_t - \mathbf{u}\|^2 - t\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 \right) + \frac{G^2}{2\lambda} \sum_{t=1}^{T} \frac{1}{t} = -\frac{\lambda T}{2}\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 + \frac{G^2}{2\lambda} \sum_{t=1}^{T} \frac{1}{t} \leq \frac{G^2}{2\lambda}(1 + \ln(T))$$

Lemma1を拡張するとさらに強力な次の定理が得られる。
詳細は:Mathematical Programming 127(1), 2011, pp.3-30
Pegasos:primal estimated sub-gradient solver for SVM. Shai Shalev-Schwarts, et.al.

Theorem 1: $\forall \mathbf{x}$(入力データ): $\|\mathbf{x}\| \le R,\quad \mathbf{w}^* = \arg\min_{f(\mathbf{w})} f(\mathbf{w})$

かつ　$projection$ したときは　$c = \left(\sqrt{\lambda} + R\right)^2$

$projection$ しないときは　$c = 4R^2$　とすると

$T \ge 3$ に対して

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t; A_t) \le \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}^*; A_t) + \frac{c(1+\ln(T))}{2\lambda T}$$

A_t は全データ D から選ばれた部分集合。

Proof :

if projectionが起こらない

then Bは半径$=1/\sqrt{\lambda}$の球で、$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}' \in B} \left\| \mathbf{w}_t - \eta_t \nabla f\left(\mathbf{w}_t; A_t\right) - \mathbf{w}' \right\|$

$\Rightarrow$ Theorem1を証明するにはLemma1の前提条件が成立することを証明する。

(1) $f\left(\mathbf{w}, A_t\right)$は、$\lambda$ - strongly convex : $\frac{\lambda}{2}\|\mathbf{w}\|^2$と$A_t$のヒンジ損失の平均値の和なので

λ - strongly convex

Proof : cont'd

(2)次に$\|\nabla_t\|$の上界を求める。

if projection stepが実行された

$$\text{then }\ \|\mathbf{w}_t\| \le 1/\sqrt{\lambda}\ \text{かつ}\ \|\mathbf{x}\| \le R \Rightarrow \nabla f\left(\mathbf{w}_t; A_t\right) = \lambda \mathbf{w}^{(t)} - \frac{1}{|A|}\sum_{i \in A^+} y_i \mathbf{x}_i \le \sqrt{\lambda} + R$$

if projection stepが実行されなかった

$$\text{then }\ \mathbf{w}_{t+1} = (1 - \eta_t \lambda)\mathbf{w}_t + \frac{\eta_t}{|A_t|}\sum_{i \in A_t^+} y_i \mathbf{x}_i = \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \frac{1}{\lambda t |A_t|}\sum_{i \in A_t^+} y_i \mathbf{x}_i$$

$$\text{where } A_t^+ = \left\{ i \in A_t \mid 1 - y_i\left(\mathbf{w}^{(t)} \cdot \mathbf{x}_i\right) > 0 \right\}$$

$$\text{ここで}\ \mathbf{w}_1 = 0 \Rightarrow \mathbf{w}_2 = \frac{1}{\lambda |A_1|}\sum_{i \in A_1^+} y_i \mathbf{x}_i \Rightarrow \mathbf{w}_3 = \left(\frac{1}{2}\right)\frac{1}{\lambda |A_1|}\sum_{i \in A_1^+} y_i \mathbf{x}_i + \frac{1}{2\lambda |A_2|}\sum_{i \in A_2^+} y_i \mathbf{x}_i$$

$$\Rightarrow \mathbf{w}_4 = \left(\frac{\cancel{2}}{3}\right)\left(\frac{1}{\cancel{2}}\right)\left(\frac{1}{\lambda |A_1|}\sum_{i \in A_1^+} y_i \mathbf{x}_i + \frac{1}{\lambda |A_2|}\sum_{i \in A_2^+} y_i \mathbf{x}_i\right) + \frac{1}{3\lambda |A_3|}\sum_{i \in A_3^+} y_i \mathbf{x}_i$$

$$\Rightarrow \mathbf{w}_t = \frac{1}{\lambda t}\sum_{j=1}^{t} \frac{1}{|A_j|}\sum_{i \in A_j^+} y_i \mathbf{x}_i$$

$$\Rightarrow \|\mathbf{w}_{t+1}\| \le \frac{R}{\lambda}$$

ここの導出は初等的が
だちょっとした計算

Proof : cont'd

(3)最後に$\mathbf{w}^* \in B$を示す。

projectionしない場合は、$\mathbf{w}_{t+1} = \underset{\mathbf{w}' \in B}{\arg\min} \left\| \mathbf{w}_t - \eta_t \nabla_t(\mathbf{w}_t) - \mathbf{w}' \right\|$により

$\mathbf{w}^* \in B$と言える。

projectionした場合は、$\left\| \mathbf{w}^* \right\| \leq 1/\sqrt{\lambda}$を示す。

ここで対象にしているSVMの主問題は以下の形式

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i \ \text{ s.t. } \forall \mathrm{i} \in [1, \mathrm{m}] : \xi_i \geq 0, \xi_i \geq 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i)$$

$C = 1/(\lambda m)$とおき、以下の双対問題を導く。

ここで双対問題を思いつくところがいかにもSVM的

$$\min \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 \text{ s.t. } \forall \mathrm{i} \in [1, \mathrm{m}] : 0 \leq \alpha_i \leq C \qquad (50)$$

主問題の解を$(\mathbf{w}^*, \xi^*)$、双対問題の解をα^*とすると

Proof : cont'd

主問題の解を$(\mathbf{w}^*, \xi^*)$、双対問題の解をα^*とすると

$\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \mathbf{x}_i$であり(50)は次のように書き直せる。$\|\alpha^*\|_1 - \frac{1}{2}\|\mathbf{w}^*\|^2$

SVMの問題では強双対定理が成り立つので

強双対定理の実に賢い使い方だ

$$\frac{1}{2}\|\mathbf{w}^*\|^2 + C\|\xi^*\|_1 = \|\alpha^*\|_1 - \frac{1}{2}\|\mathbf{w}^*\|^2$$

$$\|\alpha^*\|_\infty = \max\{\alpha_i \in \alpha^*\} \le C = 1/(\lambda m) \Rightarrow \|\alpha^*\|_1 = \sum_{i=1}^{m} |\alpha_i| \le Cm = 1/\lambda$$

$$\Rightarrow \frac{1}{2}\|\mathbf{w}^*\|^2 \le \frac{1}{2}\|\mathbf{w}^*\|^2 + C\|\xi^*\|_1 = \|\alpha^*\|_1 - \frac{1}{2}\|\mathbf{w}^*\|^2 \le \frac{1}{\lambda} - \frac{1}{2}\|\mathbf{w}^*\|^2$$

$$\Rightarrow \|\mathbf{w}^*\| \le 1/\sqrt{\lambda}$$

$\Rightarrow$ 以上の結果(1)(2)(3)をLemma1に適用すれば定理が得られる。 $\square$

Coordinate Descent

C.-J. Hsieh, et.al. ICML2008

- Target: $L1$損失- $L2$正則化のSVM　双対化して解く。下に定義

$$\min_{\boldsymbol{\alpha}} f^D(\boldsymbol{\alpha}) \equiv \frac{1}{2}\boldsymbol{\alpha}^T Q\,\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}$$

$$\text{subject to} \quad 0 \le \alpha_i \le C \quad \forall i = 1,\dots,l$$

$$\text{where} \quad Q_{ij} \equiv y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

Coordinate Descent

- Coordinate Descentは順番に1変数づつ選び、他の変数は固定して最適化。

$$\min_d f^D(\boldsymbol{\alpha}+de_i)-f^D(\boldsymbol{\alpha})\equiv\frac{1}{2}Q_{ii}d^2+\nabla_i f^D(\boldsymbol{\alpha})d \qquad f^D\text{は}f\text{の双対}$$

$$\text{subject to}\quad 0\le\alpha_i+d\le C \quad \text{where}\ \ e_i=\Bigl[\underbrace{0,..0}_{i-1},1,0,...0\Bigr]^T$$

dを計算することによるα_iの更新式は下式

$$\alpha_i\leftarrow\min\left(\max\left(\alpha_i-\frac{\nabla_i f^D(\boldsymbol{\alpha})}{Q_{ii}},0\right),C\right) \qquad (CD10)$$

Coordinate Descent　つづき

- $(CD10)$のQ_{ii}はα_iの最適化の中で1回計算すればよいが
- $\nabla_i f^D(\boldsymbol{\alpha}) = (Q\boldsymbol{\alpha})_i - 1 = \sum_{t=1}^{l} \left(y_i y_t \langle \mathbf{x}_i, \mathbf{x}_t \rangle \right) \alpha_t - 1$ は　$\langle \mathbf{x}_i, \mathbf{x}_t \rangle \forall t = 1,..,l$

  の計算コストが$O(nl)$でうれしくない。そこで
- $\mathbf{u} \equiv \sum_{t=1}^{l} y_t \alpha_t \mathbf{x}_t$ を保持しておけば
- $\nabla_i f^D(\boldsymbol{\alpha}) = (Q\boldsymbol{\alpha})_i - 1 = y_i \langle \mathbf{u}, \mathbf{x}_i \rangle - 1$　(CD20)

  となり計算コストは 以下の計算のための$O(n)$でうれしい。

  $\mathbf{u} \leftarrow \mathbf{u} + y_i (\alpha_i - \overline{\alpha}_i) \mathbf{x}_i$　　　(CD30)

  $\overline{\alpha}_i$ (CD10)の更新前、α_i (CD10)の更新後

L1損失-L2正則化のSVMの Coordinate Descent アルゴリズム

$\boldsymbol{\alpha}$の初期化、および$\mathbf{u} = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i$

$Q_{ii} \ \forall i = 1,\ldots,l$の計算

while $\boldsymbol{\alpha}$ is not optimal

For $i = 1,..l$

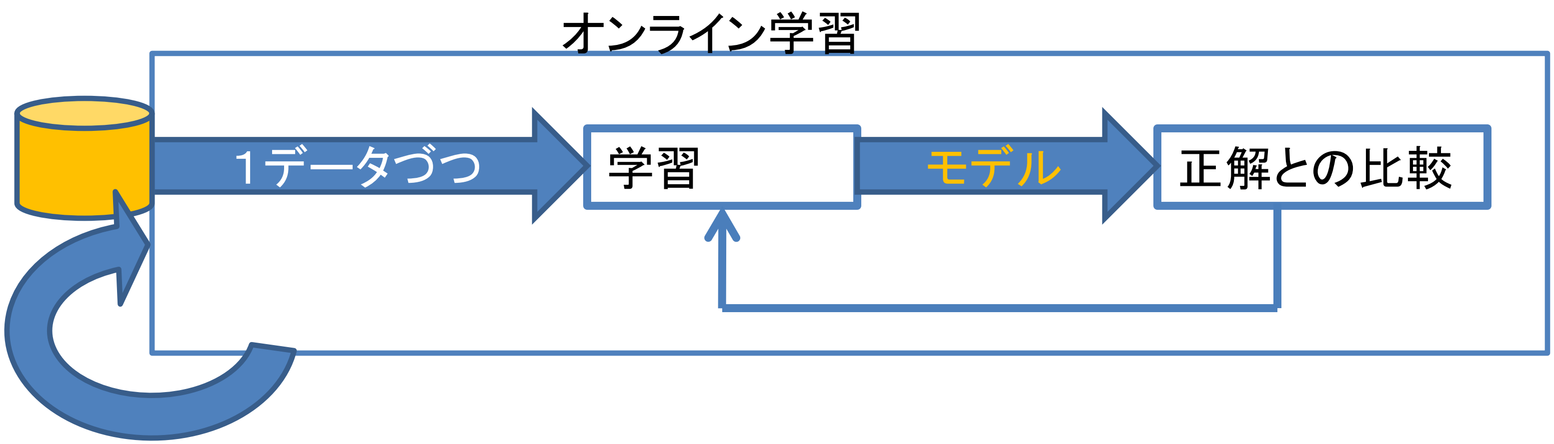
(CD20)により$G = y_i \langle \mathbf{u}, \mathbf{x}_i \rangle - 1$の計算

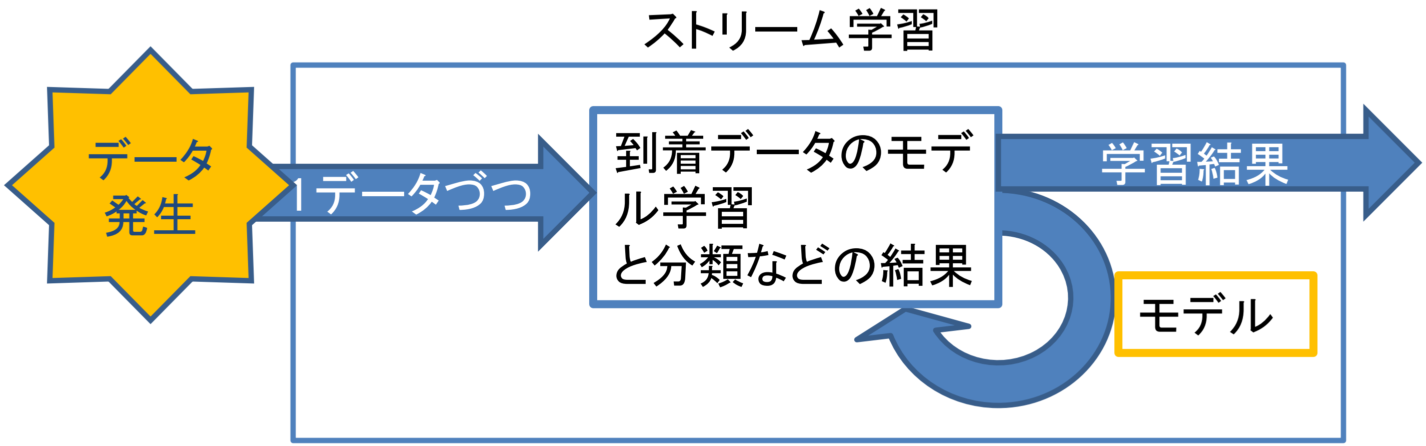
$$\overline{\alpha}_i \leftarrow \alpha_i$$

$$\alpha_i \leftarrow \min\left(\max\left(\alpha_i - \frac{G}{Q_{ii}}, 0 \right), C \right)$$

$$\mathbf{u} \leftarrow \mathbf{u} + y_i (\alpha_i - \overline{\alpha}_i) \mathbf{x}_i$$

オンライン学習とストリーム学習

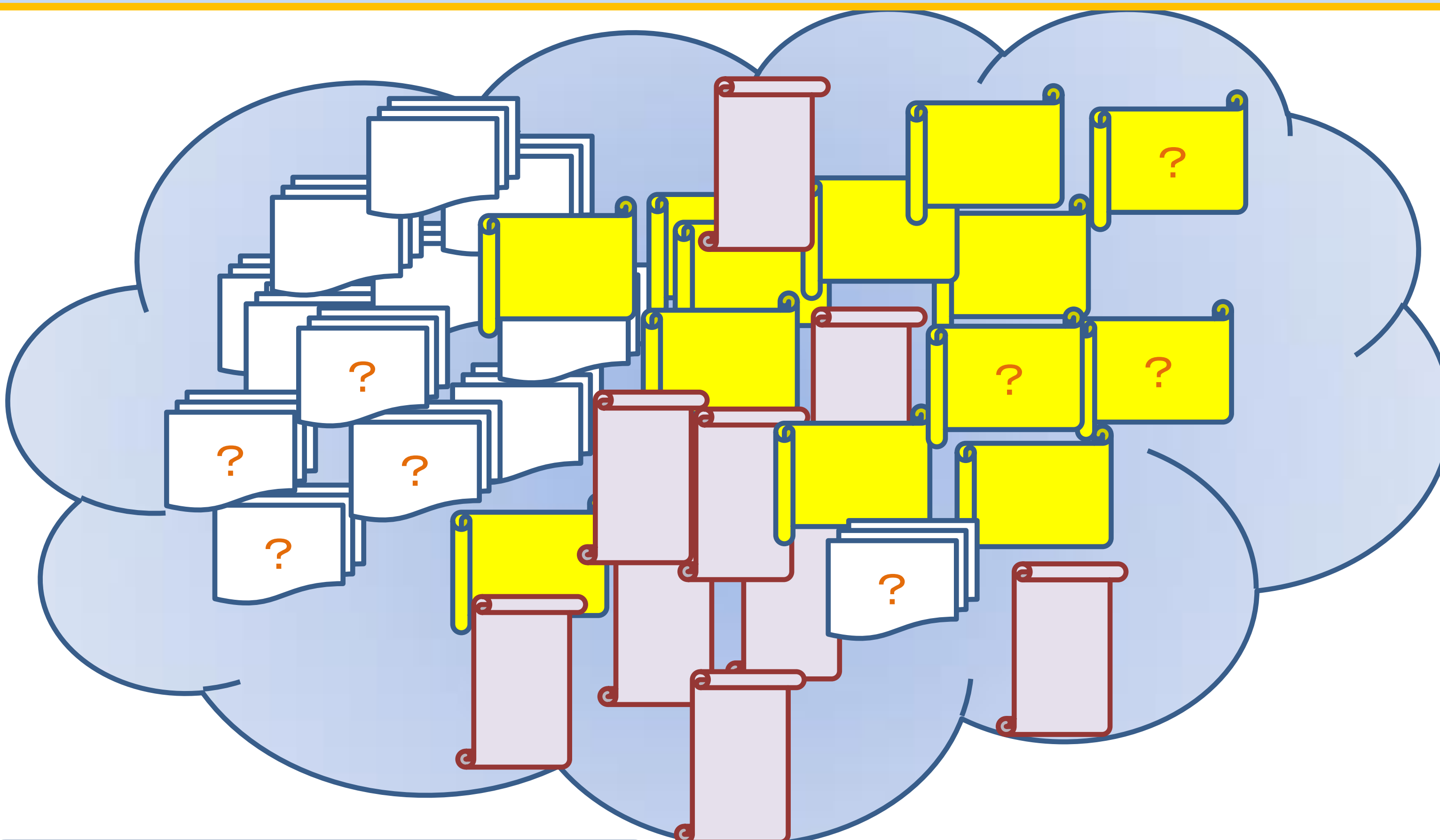


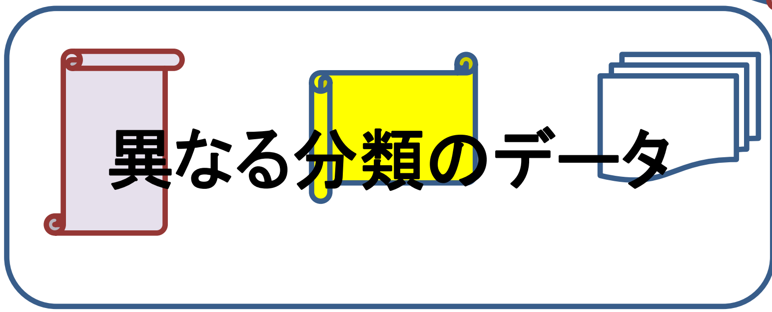


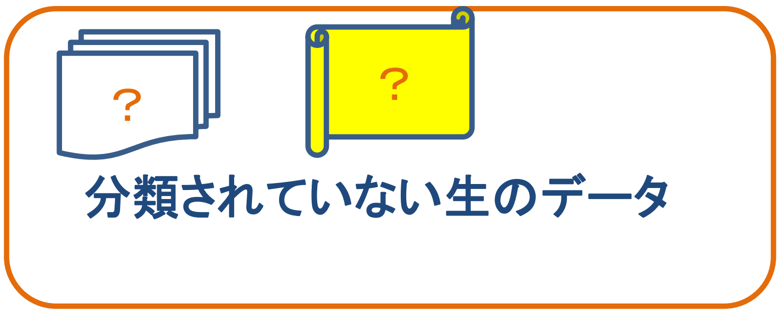
バッチ、オンライン、ストリームの比較

	バッチ学習	オンライン学習	ストリーム学習
メモリに乗せるデータ	同時に全部	同時には1データ	同時には1データ
メモリ量	大	小でも可能	小
データの到来	全データが揃ってから処理	全データが揃ってから処理	1データ到着ごとに処理
データの消去	消去せず	消去せず	データは処理後に消去
同一データの処理回数	収束するまで繰り返し	収束するまで繰り返し	1回
メモリに保持するモノ	全データと途中に内部状態	内部状態のみでも可能	内部状態のみ
性能	精度高	バッチより劣る。ただし、最近はバッチに肉迫	劣る
可能な処理	何でもあり	やや制限あり	限定的

捕捉：世の中、ビッグデータがホットだと言うけれど







パーセプトロンの別のアルゴリズム

データx_iはN個ある。

$w(0)=0; k=0; R=\max_{1\le i\le N}\|x_i\|;$

repeat

　for $i=1$ *to* N

　　$\boxed{\{\text{if } y_i\langle w(k),x_i\rangle\le 0 \text{ then}}$

　　　$\{w(k+1)=w(k)+\eta y_i x_i;\quad k=k+1\quad;\}$

　　}

until $\left\{\begin{array}{l}\text{forループ内で失敗しない} \\ \text{ように}w(k)\text{を最適化の結果とする。} \\ \text{（すなわち}y_i\langle w(k),x_i\rangle\le 0\quad\text{の場合なし）}\end{array}\right.$

この部分が線形識別

$y_i\langle w(k),x_i\rangle\le 0$　という識別に失敗したデータに、その値を重み（学習率と呼ぶ）ηでwに足しこんで是正を図るアルゴリズム

パーセプトロンは有限回で収束

→ mistakeのupper bound

Novikoffの定理(バイアスのない場合）

$$R = \max_{1 \le i \le N} \|x_i\| \qquad (0)$$

$$y_i \langle w_{opt}, x_i \rangle \ge \gamma : \text{マージン} \qquad (1)$$

であるw_{opt}が存在するなら、パーセプトロン

アルゴリズムが失敗する回数はたかだか

$\left(\frac{R}{\gamma}\right)^2 \|w_{opt}\|^2$回である

証明

t回目の失敗　に先立つ重みをw_{t-1}

更新は、$y_i\langle w_{t-1}, x_i\rangle < 0$　のとき起こる。このとき$w_t = w_{t-1} + \eta y_i x_i$ (2)

(1)より　$\langle w_t, w_{opt}\rangle = \langle w_{t-1}, w_{opt}\rangle + \eta y_i \langle x_i, w_{opt}\rangle \geq \langle w_{t-1}, w_{opt}\rangle + \eta\gamma$　　(3)

$w_0 = 0$とすれば(3)を繰り返し用いて$\langle w_t, w_{opt}\rangle \geq t\eta\gamma$　　(4)

(2)より$\| w_t \|^2 = \| w_{t-1} \|^2 + 2\eta y_i \langle w_{t-1}, x_i\rangle + \eta^2 \| x_i \|^2$

$\leftarrow x_i$は負例なので第2項は負

$\leq \| w_{t-1} \|^2 + \eta^2 \| x_i \|^2 \leq \| \hat{w}_{t-1} \|^2 + \eta^2 R^2$

$\Rightarrow \quad \| w_t \|^2 \leq t\eta^2 R^2 \quad \Rightarrow \quad \| w_t \| \leq \sqrt{t}\eta R$　　(5)

(4)(5)より$\| w_{opt} \| \sqrt{t}\eta R \geq \| w_{opt} \| \| w_t \| \geq \langle w_t, w_{opt}\rangle \geq t\eta\gamma$

$\Rightarrow \quad t \leq \left(\frac{R}{\gamma}\right)^2 \| w_{opt} \|^2$　　■

メモリ容量より大きなデータのSVM
Hsiang-Fu Yu　et.al KDD2010

主問題の場合はPegasos, 双対問題の場合はCoordinate Descent（CD）をブロックごとに適用

1. 全データインデクス{1,..,l}をmブロック $B_1,\ldots B_m$に分割
2. 主問題なら**w**, 双対問題なら**α**を初期化
3. For k=1,2,… (外側の繰り返し)
4. 　For j=1,…,m (内側の繰り返し)
5. 　　read x_r ∀r∈Bj from Disk
6. 　　{x_r| r∈Bj}に関して最適化(PegasosかCD)を行う
7. 　　**w**あるいは**α**を更新

ここが重要

主問題をPegasosで解く場合の6.の部分

Pegasosでは次の最適化をする

$$\min_{\mathbf{w}} \frac{1}{2lC}\|\mathbf{w}\|^2 + \frac{1}{l}\sum_{i=1}^{l}\max\left(1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle, 0\right)$$

$\mathbf{B}_j$を$B_j^1,\ldots.,B_j^{\bar{r}}$に分割

For $r = 1,\ldots,\bar{r}$

$$\bar{\mathbf{w}} = \mathbf{w} - \eta^t \nabla^t$$

where $\eta^t = \dfrac{lC}{t}$ $\qquad \nabla^t = \dfrac{1}{lC}\mathbf{w} + \dfrac{1}{|B|}\sum_{i \in B^+}^{l} y_i \mathbf{x}_i,\ \ B^+ = \left\{ i \in B \mid 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle > 0 \right\}$

$$\mathbf{w} \leftarrow \min\left(1, \frac{\sqrt{lC}}{\|\bar{\mathbf{w}}\|}\right)\bar{\mathbf{w}}$$

$$t \leftarrow t + 1$$

end For

双対問題でCoordinate Descent（CD）を使う場合

$$\min_{\mathbf{dBj}} f^D(\boldsymbol{\alpha}+\mathbf{d}_{\mathbf{Bj}}) \qquad f^D\text{は主問題}f\text{の双対問題} \quad (\text{CD10})$$

subject to $\mathbf{d}_{\overline{\mathbf{B}}j} = 0$ and $0 \le \alpha_i + d_i \le C \quad \forall i \in B_j$

where $Q_{ij} \equiv y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ $\overline{\mathbf{B}}j = \{1,\dots,l\} \setminus \mathbf{B}j$

ここで、Qのうちメモリにいるブロック$\mathbf{B}j$に関する部分だけを使い、下の最適化を行う（$i=1,..,l$に注意）

$$\min_{\mathbf{d}_{\mathbf{Bj}}} f^D(\boldsymbol{\alpha}+\mathbf{d}_{\mathbf{Bj}}) \equiv \frac{1}{2}\mathbf{d}_{\mathbf{Bj}}{}^T Q_{\mathbf{BjBj}} \mathbf{d}_{\mathbf{Bj}} + \boldsymbol{\alpha}^T Q_{\mathbf{Bj},i} \mathbf{d}_{\mathbf{Bj}} - \mathbf{e}_{\mathbf{Bj}}{}^T \mathbf{d}_{\mathbf{Bj}} - f^D(\boldsymbol{\alpha}) \quad (\text{CD20})$$

$$\Rightarrow 6.\text{の}\boldsymbol{\alpha}\text{更新部分：}\boldsymbol{\alpha}_{\mathbf{Bj}} \leftarrow \boldsymbol{\alpha}_{\mathbf{Bj}} + \arg\min_{\mathbf{d}_{\mathbf{Bj}}} f^D(\boldsymbol{\alpha}+\mathbf{d}_{\mathbf{Bj}})$$

さて、$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$ をメモリ中に保持しておけば

$\boldsymbol{\alpha}^T Q_{r,i} \mathbf{d}_r = y_r \langle \mathbf{w}, \mathbf{x}_r \rangle, \forall r \in \mathbf{B}j$ $\Rightarrow$ 最適化に必要なのはブロック$\mathbf{B}j$だけ。

なお、$\boxed{\mathbf{w} \leftarrow \mathbf{w} + \sum_{r \in \mathbf{B}j} d_r y_r \mathbf{x}_r \ (\text{CD30})}$

という更新で6.の更新部分で$\mathbf{w}$は更新すればよい。

双対化の御利益：
教師データアクセスの観点から

- 主問題と双対問題は最適化するパラメター数が違う。
  - 主問題パラメター数 >>双対問題パラメター数 なら双対問題を解くほうが楽　→教科書的
- SVMの場合：
  - 主問題のパラメターは重みベクトル:w
  - 双対問題にパラメターは個別データ:x_i
  - →必ずしも教科書的なお得感ではない。

双対化の御利益

- SVMの場合：
  - 主問題のパラメターは重みベクトル:w
  - 下の定式化なので、全教師データ$\{t_n, \mathbf{x}_n\}$が同時に必要

$$\arg\min_{\mathbf{w},b} \frac{1}{2} ||\mathbf{w}||^2$$

$$\text{subject to} \quad t_n(\mathbf{w}^T \phi(\boxed{\mathbf{x}_n}) + b) \geq 1 \quad n = 1, \cdots, N \quad \cdots(SVM\,30)$$

- データ量が大きくメモリにロード仕切れない場合に困ったことになる。
  - データ量は最近、増加傾向

双対化の御利益

- →必ずしも教科書的なお得感ではない。
- 一方、双対問題では入力データ$\mathbf{x}_i t_i$のと最適化するa_iが対応する形で最適化式に現れるので、どのデータを学習で使うか制御しやすい。（下の式参照）
  - 例えば、$a_i(i\neq j)$を固定して、a_jを最適化する操作をjを動かして繰り返すなど。そのときには $k(\mathbf{x}_i,\mathbf{x}_j)\ j=1,\dots,N$ だけしか使わない。

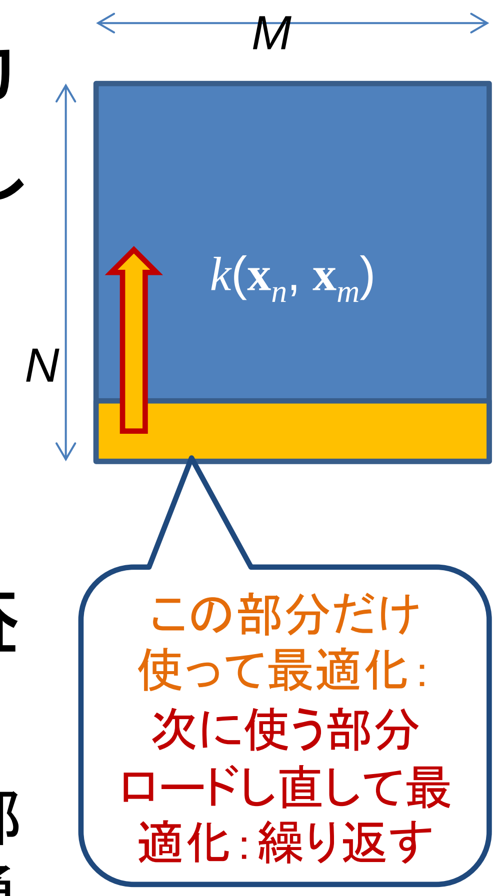
$$\max \tilde{L}(\mathbf{a}) = \max\left[\sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n,\mathbf{x}_m)\right] \quad \cdots (SVM\,70)$$

$$\text{subject to} \quad a_n \geq 0 \quad n=1,..,N \quad \cdots (SVM\,80)$$

$$\sum_{n=1}^{N} a_n t_n = 0 \quad \cdots (SVM\,90) \ \ where \quad k(\mathbf{x}_n,\mathbf{x}_m) = \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$$

双対化の御利益

- 入力データ、あるいはカーネル行列全体がメモリに乗り切らないビッグデータを扱うために、入力（すなわちカーネル$k(\mathbf{x}_n, \mathbf{x}_m)$)の一部を取捨選択してメモリにロードして使う方法が、この双対化で可能になっている。

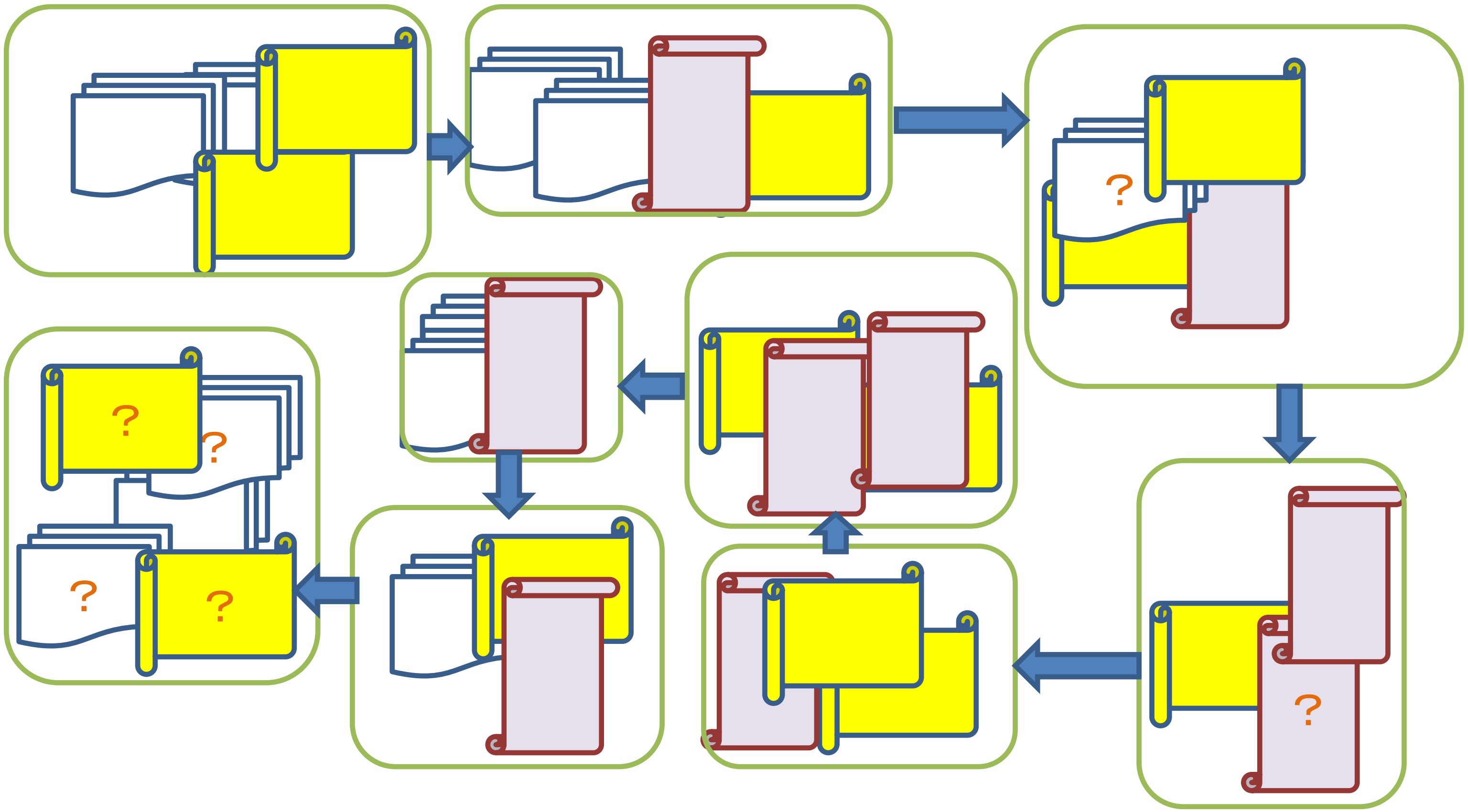


- →ビッグデータ時代における御利益
  - cf. 台湾大学のLIBSVM（SVMの実装）
  - 全データからどのようにメモリにロードする部分を切り出すかについてはここで紹介した通り。

内外のバランス　など

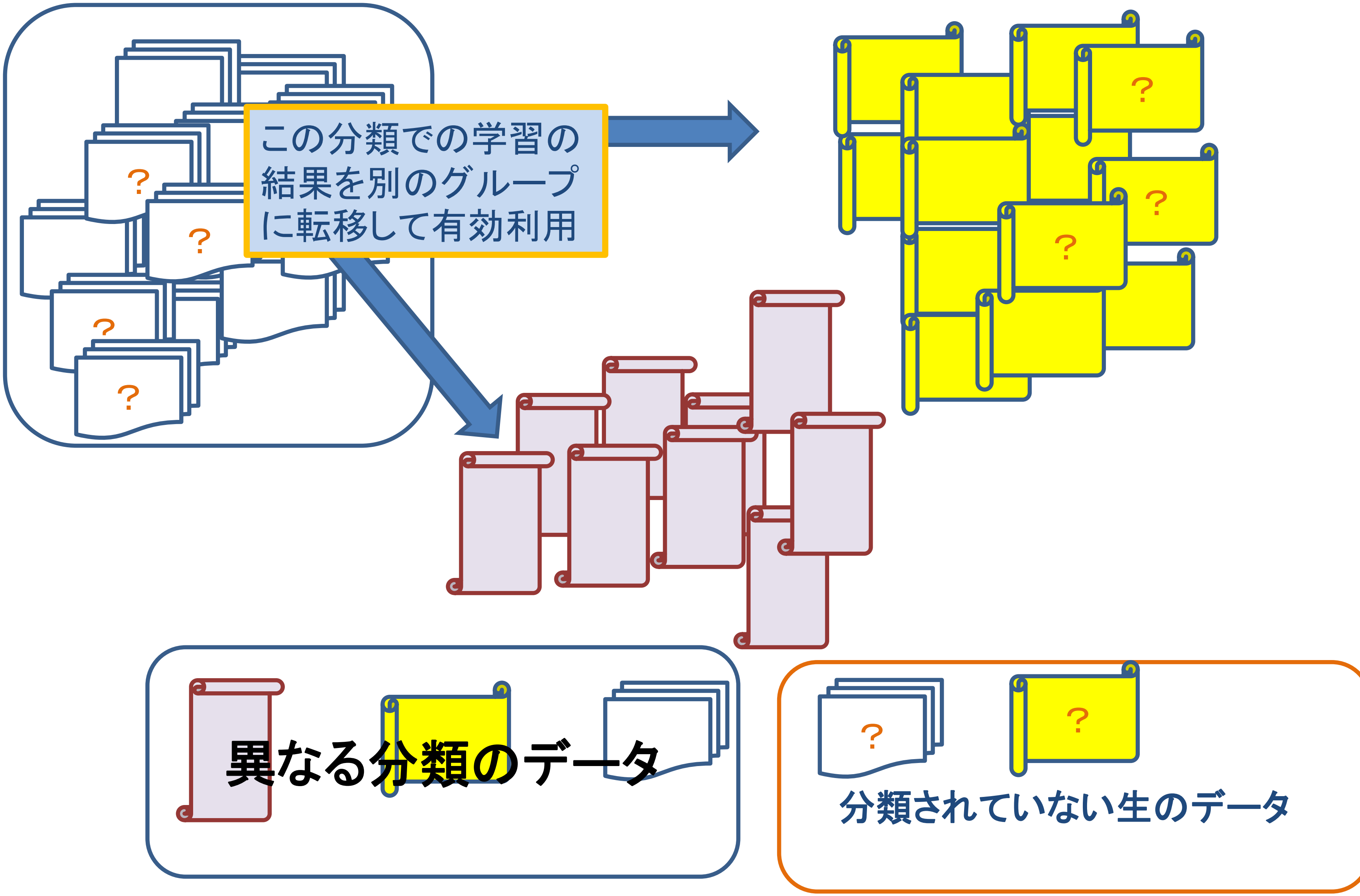
- 内側の繰り返しで最適化でCDにおいてαの更新を1回にし、looseな解を求めると、外側の繰り返しが多数回必要
- 内側の繰り返しで精密な最適化を行えば、外側の繰り返しは少なくてよい。
- Bj{j=1,..,m}内の要素の最適化処理における順番は外側の繰り返し毎にランダムに変更した方が収束が早い

小さなブロックに分けてデータマイニング、機械学習

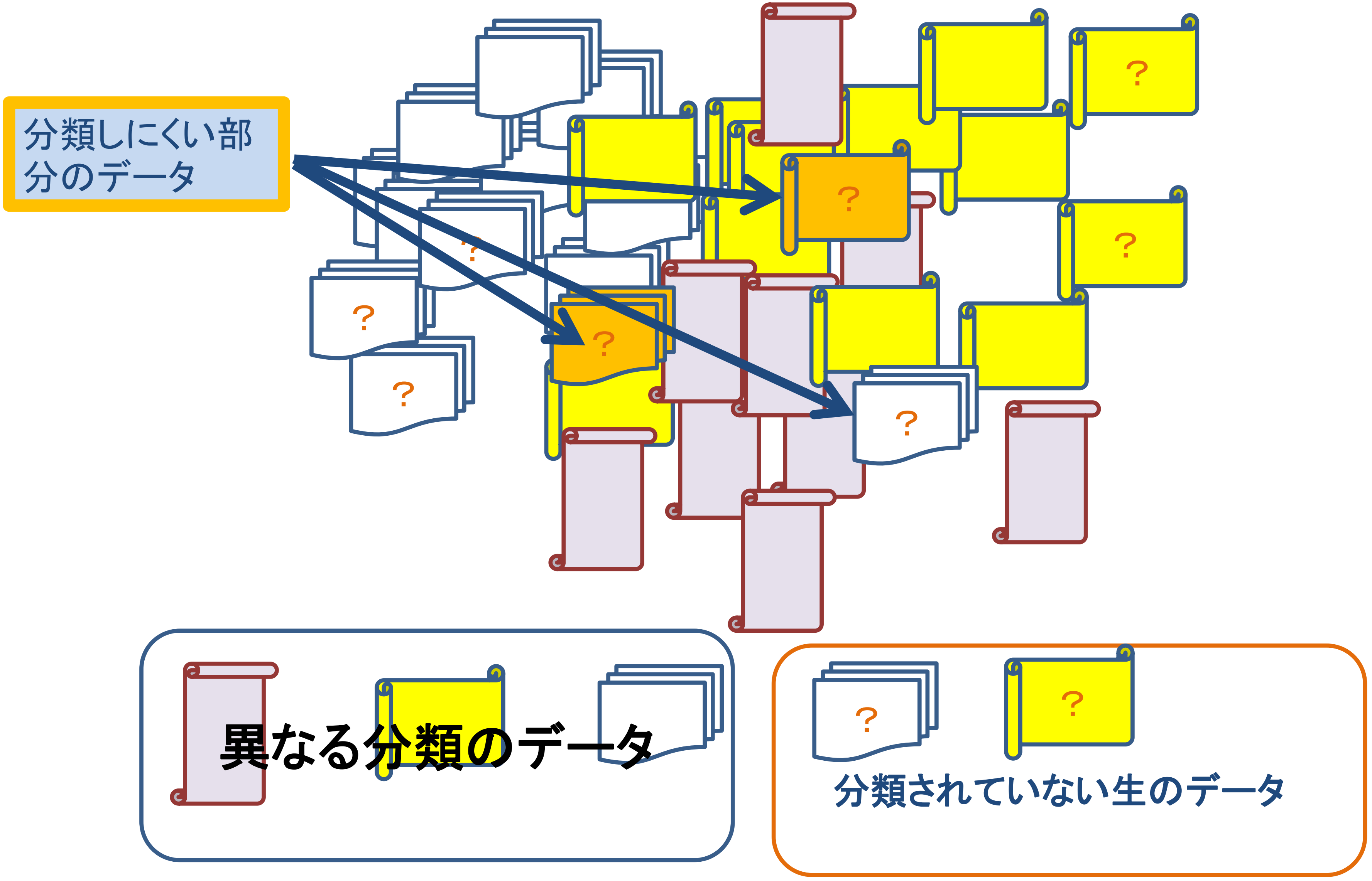


ブロックをメモリに順次ロードして学習し、その結果を活用して次のブロックへ　と繰り返す：
例えば Stream SVM

転移学習



人間に正解をつけてもらうデータを絞り込む：Active学習



付録：DualityによるFoReLの定式化

Fenchel双対 : $f^*(\boldsymbol{\theta}) = \max_{\mathbf{u}} \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u})$

明らかに次式が成立 : Fenchel - Young Equality :

$\forall \mathbf{u}, \quad f^*(\boldsymbol{\theta}) \geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u})$

$\mathbf{w}_t = \arg\max_{\mathbf{u}} \langle \mathbf{u}, \mathbf{z}_t \rangle - f(\mathbf{u})$だとすると

$\langle \mathbf{w}_t, \mathbf{z}_t \rangle - f(\mathbf{w}_t) \geq \langle \mathbf{u}, \mathbf{z}_t \rangle - f(\mathbf{u})$

$\Rightarrow f(\mathbf{u}) - f(\mathbf{w}_t) \geq \langle \mathbf{u} - \mathbf{w}_t, \mathbf{z}_t \rangle$

$\mathbf{z}_t$はfの$\mathbf{w}_t$におけるsub - gradient

fが微分可能なら、等式が成立するときはf^*の定義より$\mathbf{z}_t = \nabla f(\mathbf{w}_t)$

Online Mirror Descent:OMD

FoReLで$f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}_t \rangle + R(\mathbf{w})$ $R(\mathbf{w})$は正則化関数とする

なお、$\mathbf{w} \notin S$だと$R(\mathbf{w}) = \infty$とする

$\mathbf{z}_{1:t} = \sum_{i=1}^{t} \mathbf{z}_i$と略記するとFoReLは

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} R(\mathbf{w}) + \sum_{i=1}^{t} \langle \mathbf{w}, \mathbf{z}_t \rangle = \arg\min_{\mathbf{w}} R(\mathbf{w}) + \langle \mathbf{w}, \mathbf{z}_{1:t} \rangle$$

$$= \arg\max_{\mathbf{w}} \langle \mathbf{w}, -\mathbf{z}_{1:t} \rangle - R(\mathbf{w})$$

$g(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - R(\mathbf{w})$とおくとFoReLは

1. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t$
2. $\mathbf{w}_{t+1} = g(\boldsymbol{\theta}_{t+1})$

Online Mirror Descent (OMD)

$g : R^d \to S$

$\boldsymbol{\theta}_1 = 0$

for $t = 1,2,\ldots$

$$\mathbf{w}_t = g(\boldsymbol{\theta}_t) \left(= \arg\max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - R(\mathbf{w})\right)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{z}_t \quad \text{where } \mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$$

ここで以下が言える

$$\mathrm{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \right) \le R(\mathbf{u}) - \min_{\mathbf{v} \in S} R(\mathbf{v}) + \eta \sum_{t=1}^{T} \left\| \mathbf{z}_t \right\|_*^2 \quad \text{(OMD100)}$$

$$R(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 \Rightarrow R^* = \frac{1}{2} \|\boldsymbol{\theta}\|^2$$

パーセプトロンの別のアルゴリズム

データx_iはN個ある。

$w(0)=0; b(0)=0; k=0; R=\max_{1\le i\le N}\|x_i\|;$

repeat

$\quad for \quad i=1 \quad to \quad N$

$\quad\quad \{\text{if } y_i \boxed{\left(w(k)^T x_i + b(k)\right)} \le 0 \text{ then}$

この部分が線形識別

$\quad\quad\quad \{w(k+1)=w(k)+\eta y_i x_i; \quad b(k+1)=b(k)+\eta y_i R^2; \quad k=k+1 \quad ;\}$

$\quad\quad \}$

until
- forループ内で失敗しない
- (すなわち$y_i\left(w(k)^T x_i + b(k)\right)\le 0$　の場合なし)
- ように$w(k), b(k)$を最適化の結果とする。

$y_i\left(w(k)^T x_i + b(k)\right)\le 0$ という識別に失敗したデータに、その値を重み(学習率と呼ぶ)ηでwに足しこんで是正を図るアルゴリズム

更新後の判定式を見てみる

$$y_i\left(w(k+1)^T x_i + b(k+1)\right)$$

$$= y_i\left(w(k)^T x_i + b(k)\right) + y_i\left(\eta y_i x_i + \eta y_i R^2\right)$$

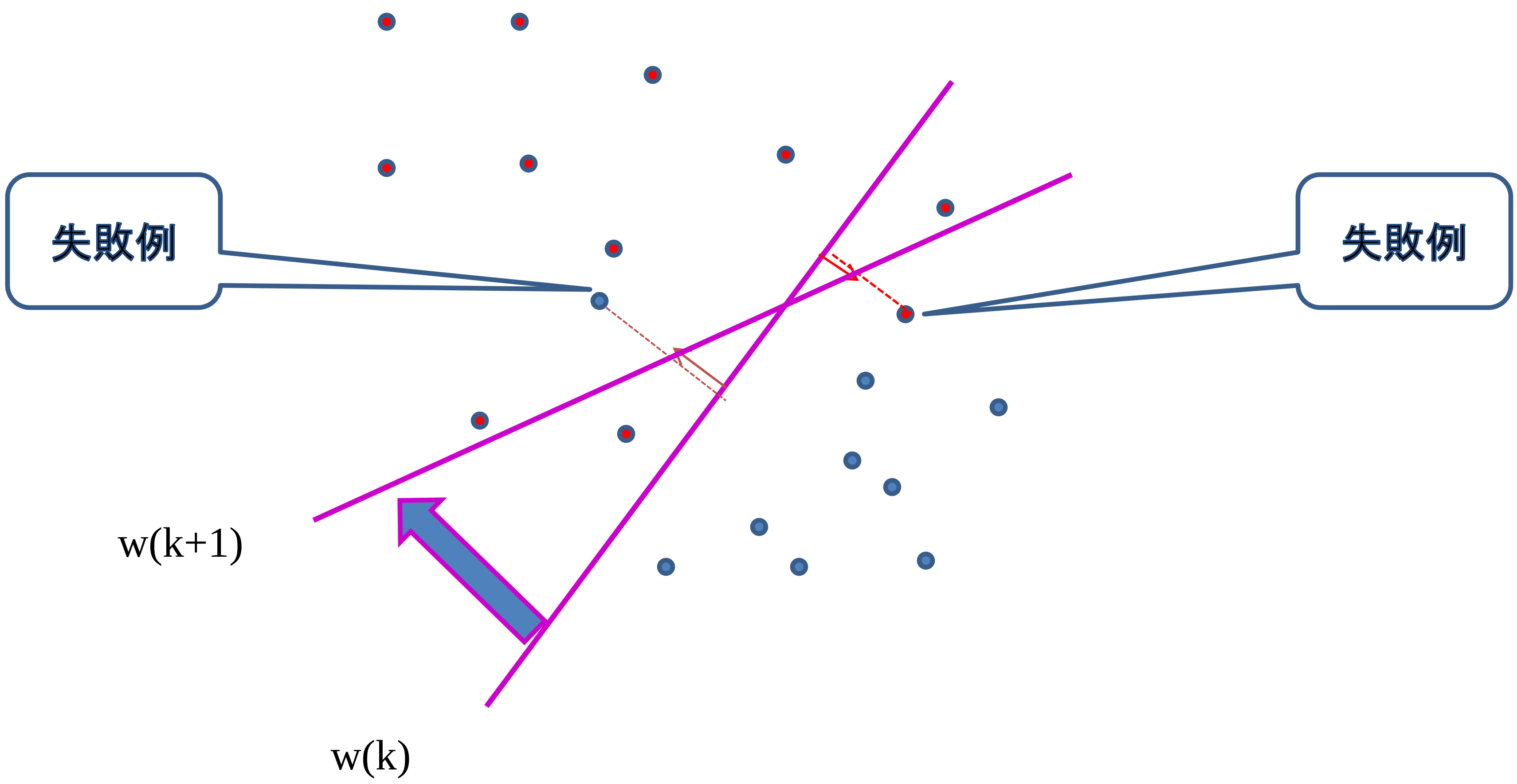
$$= y_i\left(w(k)^T x_i + b(k)\right) + \eta y_i^{\,2}\left(x_i^{\,2} + R^2\right)$$

$$= y_i\left(w(k)^T x_i + b(k)\right) + \eta y_i^{\,2}\left(x_i^{\,2} + \max\|x_i\|^2\right)$$

第2項は必ず正

よって、 $y_i\left(w(k+1)^T x_i + b(k+1)\right) > y_i\left(w(k)^T x_i + b(k)\right)$

したがって、失敗しない方向にwは改善していく。



失敗例
失敗例
w(k+1)
w(k)

パーセプトロンは有限回で収束

→ mistakeのupper bound

Novikoffの定理（バイアスのある場合）

$$R = \max_{1 \le i \le N} \|x_i\| \qquad (0)$$

$\|w_{opt}\| = 1$　かつ　$y_i\left(w_{opt}{}^T x_i + b_{opt}\right) \ge \gamma$: マージン　(1)

であるw_{opt}が存在するなら、パーセプトロン

アルゴリズムが失敗する回数はたかだか

$\left(\dfrac{2R}{\gamma}\right)^2$回である

証明

入力ベクトルに値Rとなる1次元加え、$\hat{x}_i = \left(x_i^T, R\right)^T$とする。

$\hat{w}_i = \left(w_i^T, \frac{b_i}{R}\right)^T$とする。

t回目の失敗　に先立つ重みを$\hat{w}_{t-1}$

更新は、$y_i\left(\hat{w}_{t-1} \cdot \hat{x}_i\right) = y_i\left(w_{t-1} \cdot x_i + b_{t-1}\right) \leq 0$

のとき起こる。　$x \cdot y$は内積

このとき$\hat{w}_t = \left(w_t^T, \frac{b_t}{R}\right)^T = \left(w_{t-1}^T, \frac{b_{t-1}}{R}\right)^T + \eta y_i \left(x_i^T, R\right)^T = \hat{w}_{t-1} + \eta y_i \hat{x}_i \qquad (2)$

なぜなら、$b_t = b_{t-1} + \eta y_i R^2 \quad \Rightarrow \quad \frac{b_t}{R} = \frac{b_{t-1}}{R} + \eta y_i R$

証明　つづき

$$\hat{w}_t = \left(w_t^T, \frac{b_t}{R} \right)^T = \left(w_{t-1}^T, \frac{b_{t-1}}{R} \right)^T + \eta y_i \left(x_i^T, R \right)^T = \hat{w}_{t-1} + \eta y_i \hat{x}_i \qquad (2)$$

(1)より　$\hat{w}_t \cdot \hat{w}_{opt} = \hat{w}_{t-1} \cdot \hat{w}_{opt} + \eta y_i \left(\hat{x}_i \cdot \hat{w}_{opt} \right) \geq \hat{w}_{t-1} \cdot \hat{w}_{opt} + \eta\gamma \qquad (3)$

$\hat{w}_0 = 0$とすれば(3)を繰り返し用いて$\hat{w}_t \cdot \hat{w}_{opt} \geq t\eta\gamma \qquad (4)$

(2)より$\| \hat{w}_t \|^2 = \| \hat{w}_{t-1} \|^2 + 2\eta y_i \left(\hat{w}_{t-1} \cdot \hat{x}_i \right) + \eta^2 \| \hat{x}_i \|^2$

$\leftarrow \hat{x}_i$は負例なので第2項は負

$$\leq \| \hat{w}_{t-1} \|^2 + \eta^2 \| \hat{x}_i \|^2 = \| \hat{w}_{t-1} \|^2 + \eta^2 \left(\| x_i \|^2 + R^2 \right) \leq \| \hat{w}_{t-1} \|^2 + 2\eta^2 R^2$$

$$\Rightarrow \quad \| \hat{w}_t \|^2 \leq 2t\eta^2 R^2 \quad \Rightarrow \quad \| \hat{w}_t \| \leq \sqrt{2t}\eta R \qquad (5)$$

(4)(5)より$\| \hat{w}_{opt} \| \sqrt{2t}\eta R \geq \| \hat{w}_{opt} \| \| \hat{w}_t \| \geq \hat{w}_t \cdot \hat{w}_{opt} \geq t\eta\gamma$

$$\Rightarrow \quad t \leq 2\left(\frac{R}{\gamma} \right)^2 \| \hat{w}_{opt} \|^2 \leq 2\left(\frac{R}{\gamma} \right)^2 \left(\| \hat{w}_{opt} \|^2 + 1 \right) = \left(\frac{2R}{\gamma} \right)^2 \qquad \blacksquare$$