Adaptive Regularization of Weight Vectors (AROW)

Crammer and Kulesza and Dredze

2009

AROW の概要

Adaptive Regularization of Weight Vectors (AROW) ▸ PDF

- オンライン線形分類器の学習
  - 学習が高速
- Confidence Weighted (CW) 学習の枠組みで行う
- 割と有名なので日本語の解説もググると見つかります
- 実装が大変に容易
- 更なる改良もある

オンライン線形分類器

時刻 (round) t に事例とラベルをもらう:

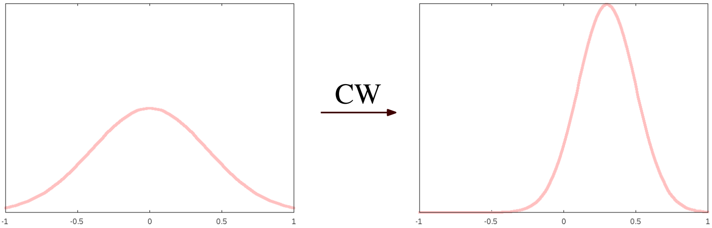
$$(\boldsymbol{x}_t, y_t); \boldsymbol{x}_t \in \mathbb{R}^d, y_t \in \mathcal{Y}$$

2 値分類: $\mathcal{Y} = \{-1, +1\}$
重みベクトル w によって分類: $\hat{y} = h_w(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x})$

Confidence Weighted (CW) learning

次を仮定して $\boldsymbol{w}$ ではなく $\boldsymbol{\mu}, \Sigma$ を学習する.

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



具体的には、$\boldsymbol{\mu} = \mathbf{0}, \Sigma = I$ からスタートして更新していく.

定数 η ($\frac{1}{2} < \eta \leq 1$) を予め設定しておく.
round $t-1$ までに μ_{t-1}, Σ_{t-1} を学習したとき、次のように更新する.

$$(\mu_t, \Sigma_t) = \min_{\mu,\Sigma} D_{KL}(\mathcal{N}(\mu,\Sigma) \| \mathcal{N}(\mu_{t-1}, \Sigma_{t-1}))$$

$$\text{such that } Pr[y_t(w \cdot x_t) \geq 0] \geq \eta$$
$$\text{ただし } w \sim \mathcal{N}(\mu, \Sigma)$$

$Pr[y(w \cdot x) \geq 0] \geq \eta \; (w \sim \mathcal{N}(\mu, \Sigma))$
$\iff \; y(\mu \cdot x) \geq \phi\sqrt{x^\top \Sigma x}$ (ϕ は η に対応する正の定数)
であるらしいので
更新式は結局、

$$(\mu_t, \Sigma_t) = \arg\min_{\mu, \Sigma} D_{KL}(\mathcal{N}(\mu, \Sigma) \| \mathcal{N}(\mu_{t-1}, \Sigma_{t-1}))$$

$$\text{such that } y(\mu \cdot x) \geq \phi\sqrt{x^\top \Sigma x}$$

でよい.
共分散 Σ の固有値の逆数を confidence という. この値は学習のたびに単調増加する (直観的にはそう).

CW 学習の問題点

1. ノイズに弱い

   線形分離可能なデータを仮定している. ノイズのあるデータでも η 以上の確率で予測できるように学習するため"the update is quite aggressive" である.

2. 分類問題にしか適用できない

   先の式だと二値分類しかできない. 例えば回帰といった別な問題に拡張したい.
   (本手法でどう解決したのかは不明)

更新をソフトにする

更新式を次のように変更する

$$(\mu_t, \Sigma_t) = \underset{\mu,\Sigma}{\arg\min}\, D_{KL}(\mathcal{N}(\mu,\Sigma) \| \mathcal{N}(\mu_{t-1},\Sigma_{t-1})) + \lambda_1 \ell_{h^2}(y_t, \mu \cdot x_t) + \lambda_2 x_t^\top \Sigma x_t$$

ヒンジ損失関数: $\ell_{h^2}(y_t, \mu \cdot x_t) := (\max\{0, 1 - y_t(\mu \cdot x_t)\})^2$

第一項 前回の学習結果を引き継ぐ

第二項 新しいデータを低い損失で分類する

第三項 Σ の confidence を高める (制約式の右辺)

こっから更新式を解ける形まで変形します
$D_{KL}(\mathcal{N}\|\mathcal{N})$ を展開
$\lambda_1 = \lambda = 1/(2r)$ と仮定

$$
\begin{aligned}
C(\mu,\Sigma) = \frac{1}{2}\log\left(\frac{det\Sigma_{t-1}}{det\Sigma}\right) + \frac{1}{2}Tr\left(\Sigma_{t-1}^{-1}\Sigma\right) \\
+ \frac{1}{2}(\mu_{t-1}-\mu)^{\top}\Sigma_{t-1}^{-1}(\mu_{t-1}-\mu) - \frac{d}{2} \\
+ \frac{1}{2r}\ell_{h^2}(y_t, \mu\cdot x_t) + \frac{1}{2r}x_t^{\top}\Sigma x_t
\end{aligned}
$$

よく見ると次のような和として書ける.

$$
C(\mu,\Sigma) = C_1(\mu) + C_(\Sigma)
$$

各々独立に最小化すればよい.

μ の更新

$$\mu_t = \arg\min_{\mu} C_1(\mu)$$

$$C_1(\mu, \Sigma) = \frac{1}{2}(\mu_{t-1} - \mu)^\top \Sigma_{t-1}^{-1} (\mu_{t-1} - \mu) + \frac{1}{2r} \ell_{h^2}(y_t, \mu \cdot x_t)$$

こちらはなんとか導ける

$$
\begin{aligned}
C_1(\mu) &= \frac{1}{2}(\mu_{t-1} - \mu)^\top \Sigma_{t-1}^{-1}(\mu_{t-1} - \mu) + \frac{1}{2r}\ell_{h^2}(y_t, \mu \cdot x_t) \\
\frac{\partial}{\partial \mu} C_1(\mu, \Sigma) &= \Sigma_{t-1}^{-1}(\mu_{t-1} - \mu) + \frac{1}{2r}\frac{\partial}{\partial \mu}\ell_{h^2}(y_t, \mu \cdot x_t) \\
&= 0 \text{ っておくと} \\
\mu &= \mu_{t-1} - \frac{1}{2r}\Sigma_{t-1}\frac{\partial \ell(y_t, \mu \cdot x_t)}{\partial \mu} \\
&= \mu_{t-1} - \frac{1}{2r}\frac{d\ell(y_t, z)}{dz}\Sigma_{t-1}x_t
\end{aligned}
$$

最後の式の右辺にはよく見ると μ が混じってる.

$\ell_{h^2}(y_t, \mu \cdot x_t) := (\max\{0, 1 - y_t(\mu \cdot x_t)\})^2$ だったので
$1 - y_t(\mu \cdot x_t) > 0$ を仮定すれば
$\frac{d\ell(y_t,z)}{dz} = -2y_t(1 - y_t z)$ を代入して

$$\mu = \mu_{t-1} + \frac{y_t}{r}(1 - y_t(\mu \cdot x_t))\Sigma_{t-1}x_t$$

両辺に $\cdot x_t$ して (右辺の第二項には右から $x^\top$ を掛けて)
$\mu \cdot x_t$ について解いてそれをまた上の式に入れて (!)

$$\mu = \mu_{t-1} + \frac{\max\{0, 1 - y_t x_t^\top \mu_{t-1}\}}{x_t^\top \Sigma_{t-1} x_t + r}\Sigma_{t-1} y_t x_t$$

"It can be easily verified (this) satisfies our assumption that $1 - y_t(\mu \cdot x_t) > 0$" だそうです.

$\ell_{h^2}(y_t, \mu \cdot x_t) := (\max\{0, 1 - y_t(\mu \cdot x_t)\})^2$ だったので
$1 - y_t(\mu \cdot x_t) > 0$ を仮定すれば
$\frac{d\ell(y_t,z)}{dz} = -2y_t(1 - y_t z)$ を代入して

$$\mu = \mu_{t-1} + \frac{y_t}{r}(1 - y_t(\mu \cdot x_t))\Sigma_{t-1}x_t$$

両辺に $\cdot x_t$ して (右辺の第二項には右から $x^\top$ を掛けて)
$\mu \cdot x_t$ について解いてそれをまた上の式に入れて (!)

$$\mu = \mu_{t-1} + \frac{\max\{0, 1 - y_t x_t^\top \mu_{t-1}\}}{x_t^\top \Sigma_{t-1} x_t + r}\Sigma_{t-1} y_t x_t$$

”It can be easily verified (this) satisfies our assumption that $1 - y_t(\mu \cdot x_t) > 0$” だそうです.

μ の更新式

今のをそのまま、更新式とすればよい.

$$\mu_t = \mu_{t-1} + \frac{\max\{0, 1 - y_t x_t^\top \mu_{t-1}\}}{x_t^\top \Sigma_{t-1} x_t + r} \Sigma_{t-1} y_t x_t$$

Σ の更新

$$\Sigma_t = \arg\min_{\Sigma} C_2(\Sigma)$$

$$C_2(\Sigma) = \frac{1}{2} \log \left(\frac{det\Sigma_{t-1}}{det\Sigma} \right) + \frac{1}{2} Tr \left(\Sigma_{t-1}^{-1} \Sigma \right) + \frac{1}{2r} x_t^\top \Sigma x_t$$

$$C_2(\Sigma) = \frac{1}{2}\log\left(\frac{det\Sigma_{t-1}}{det\Sigma}\right) + \frac{1}{2}Tr\left(\Sigma_{t-1}^{-1}\Sigma\right) + \frac{1}{2r}x_t^\top \Sigma x_t$$

$$\frac{\partial}{\partial\Sigma}C_2(\Sigma) = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma_{t-1}^{-1} + \frac{1}{2r}x_t x_t^\top$$

$$= 0 \text{ っておくと}$$

$$\Sigma^{-1} = \Sigma_{t-1}^{-1} + \frac{x_t x_t^\top}{r}$$

という逆行列についての 更新式が得られる.
Woodbury identity ▸ wikipedia っていうのを用いると逆でない行列に陽に書き直せる.

$$\Sigma = \Sigma_{t-1} - \frac{\Sigma_{t-1}x_t x_t^\top \Sigma_{t-1}}{r + x_t^\top \Sigma_{t-1}x_t}$$

更新式

次の 2 つを用いて、μ, Σ を順に更新してく.

$$\mu_t = \mu_{t-1} + \frac{\max\{0, 1 - y_t x_t^\top \mu_{t-1}\}}{x_t^\top \Sigma_{t-1} x_t + r} \Sigma_{t-1} y_t x_t$$

$$\Sigma_t = \Sigma_{t-1} - \frac{\Sigma_{t-1} x_t x_t^\top \Sigma_{t-1}}{r + x_t^\top \Sigma_{t-1} x_t}$$

特に Σ の更新式は confidence (固有値の逆数) の単調増加を保証してる. あと分母に同じのがあるので共通部分式除去できる.
アルゴリズムとしては、$\mu = 0, \Sigma = I$ から初めて先の更新式を用いてオンライン処理する.

Lemma 1. Representer Theorem

$\mu = 0, \Sigma = I$ から初めて得られる μ, Σ は実は、

- 学習データ ($\{x_t\}$) の線型結合
- 学習データの外積 ($\{x_t x_t^\top\}$) の線形結合

で表現される.
帰納法で証明できる (改めて更新式を見るとそうなってる).

Theorem 2. Mistakes upper bound

最終的な学習結果を用いて、学習データ自体をテストしたときの誤る個数は次で上限が抑えられる.

$$M \leq \sqrt{r\|\boldsymbol{u}\|^2 + \boldsymbol{u}^\top \mathbf{X}_\mathcal{A} \boldsymbol{u}} \sqrt{\log\left(\det\left(I + \frac{1}{r}\mathbf{X}_\mathcal{A}\right)\right) + U + \sum_{t \in \mathcal{M} \cup \mathcal{U}} g_t - U}$$

- $\mathcal{M}$ = 更新直前で誤るデータ集合; $M = |\mathcal{M}|$
- $\mathcal{U}$ = そうでないデータ集合; $U = |\mathcal{U}|$
- $X_A = \sum_{t \in \mathcal{M} \cup \mathcal{U}} x_t x_t^\top$
- $u \in \mathbb{R}^d$ (任意の)
- $g_t = \max\{0, 1 - y_t u^\top x_t\}$

r に調整に用いることができる.

実験

比較手法として他に 3 つ. どれもオンライン線形学習器です.

1. Passive-Aggressive (PA)
2. Second Order Perceptron (SOP)
3. CW learning

PA も CW も AROW も全部同じ人 (Crammer ら)

データセット

人工のデータセット一つと、NLP タスクから沢山.

- 人工: 20 次元のガウス分布. ある 2 次元上に境界面を置いてラベルをつける.
- Amazon: レビューからドメイン (e.g., books or music) を当てる
- 20 Newsgroups
- Reuters (RCV1-v2/LYRL2004)
- Sentiment: Amazon のレビューからポジネガを当てる
- Spam: ECML/PKDD Challenge ▸ web ってのがあって、spam/ham に分類する
- OCR: 手書き文字の認識. MNIST ▸ web ってとことUSPS っていうとこが配ってるデータ

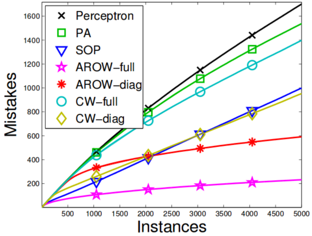
結果 (Table 1)

ノイズレベル (割合?) に応じて事前にデータセット中のラベルにノイズを与えて実験を行う.
ノイズレベル毎の 4 つの順位の平均値:

Algorithm	Noise level 0.0	0.05	0.1	0.15	0.2	0.3
AROW	**1.51**	**1.44**	**1.38**	**1.42**	**1.25**	**1.25**
CW	1.63	1.87	1.95	2.08	2.42	2.76
PA	2.95	2.83	2.78	2.61	2.33	2.08
SOP	3.91	3.87	3.89	3.89	4.00	3.91

結果 (Fig 2(a))

人工データ:



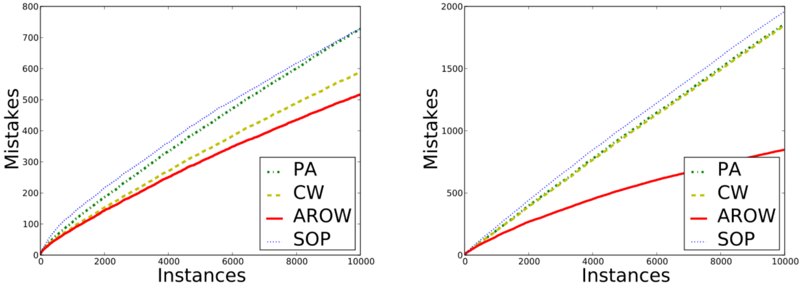
(a) synthetic data

10%のノイズ (ラベルの逆転) を与える

HOGE-diag っていうのは対角化バージョン (共分散行列の対角成分以外をゼロとして計算を省く).

結果 (Fig 2(b))

手書き文字の認識:



(b) MNIST data

左はノイズ 0%, 右はノイズ 10%.

まとめ

強いオンライン線形分類学習

- ノイズに強い
- ノイズが無くても強い

CW 亜種

要は損失関数 ℓ に色々突っ込んだらいい

$$(\mu_t, \Sigma_t) = \arg\min_{\mu,\Sigma} D_{KL}(\mathcal{N}(\mu,\Sigma) \| \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})) + C\ell(y_t, x_t; \mu, \Sigma)$$

"Exact Soft CW learning (2012)" pdf:

SCW-I : $\ell(...) = \max\{0, \phi\sqrt{x_t^\top \Sigma x_t} - y_t(\mu_t \cdot x_t)\}$

SCW-II : $\ell(...) = \max\{0, \phi\sqrt{x_t^\top \Sigma x_t} - y_t(\mu_t \cdot x_t)\}^2$

AROW より良い結果を出してる