Online Learning by Ellipsoid Method

Liu Yang, Rong Jin and Jieping Ye

Carnegie Mellon University

Michigan State University and Arizona State University

Outline

- **Motivation**
  - It is important to represent version space, why?
  - How to represent the **version space** ?
  - This work is the **first** attempt to **explicitly** represent **version space** for online learning
  - Use ellipsoid as outer approximation of set of hypotheses that is consistent with hindsight
  - Our mistake bound is same with that of percetron up to a constant factor
- Algorithm
  - Introduction to the Ellipsoid Method
  - Online Learning by Ellipsoid Methods
- Evaluation

Bayesian viewpoint : Representing the Version Space

- Why version space ?
  - Most online learners only maintain a single classifier (like point estimation), **insufficient**
  - We want to compute not only the most likely solution but also the distribution of all possible solutions
- Why important to represent version space **explicitly**?
  - Online Learning can benefit from having an explicit repressentation of the version space
  - In selective sampling (request label if only in the region of disagreement), such a reprsentation helps interchangability between model space and data region

Bayesian viewpoint : Representing the Version Space

- How to represent the **version space** ?
  - Use ellipsoid as outer approximation of set of hypotheses that is consistent with hindsight
  - **Nice properties** of the Ellipsoid Method
    - Simple updating formula for $\mathcal{E}(k+1)$
    - $\mathcal{E}(k+1)$ can be larger than $\mathcal{E}(k)$ in max semi-axis length, but always smaller in volume
    - $vol(\mathcal{E}(k+1)) < e^{-\frac{1}{2n}} vol(\mathcal{E}(k))$
      (volume reduction factor degrades rapidly with n)
- Information viewpoint : centroid and the positive definite shape matrix of ellipsoid maintain more information of training data than most existing online learners

Related Work in Online Learning

- Most are **Additive** : given a misclassified (x_i, y_i), update w by shifting along the direction of $y_i x_i$, $w + \alpha_i y_i x_i \rightarrow w$
- An quasi-additive framework unifying Perception and Winnow (Grove et al., 01)
- Extend online learning to multilabel cases (Fink et al., 06; Crammer & Singer, 03; Crammer et al., 06)
- Extend graph-based approaches for online learning (Herbster et al., 05)
- Exploited dual formation of optimization for online learning (Shalev-Shwartz & Singer, 06; Amit et al., 07)

Outline

- Motivation
- **Algorithm**
  - Introduction to Ellipsoid Method for Convex Programming
  - The Classical Ellipsoid Method for Online Learning (CELLIP)
  - Improved Ellipsoid Method for Online Learning (IELLIP)
  - Ellipsoid Methods for Multiple-Label Online Learning
- Evaluation

Ellipsoid Method for Convex Programming (Shor, 1977)

$x^* = \arg\min\{f(x) : x \in G\}$ where $f(x)$ is convex

- Starts with $\mathcal{E}_1 \supseteq G$.
- Repeat until ϵ-suboptimal
  - $\mathcal{E}_k = \{x | (x - x_k)^\top P_k^{-1}(x - x_k) \leq 1\}$ containing x^*, $x_k \in \mathbb{R}^d$ and $P_k \in S_{++}^{d\times d}$
  - Compute gradient h_k of $f(x)$ at x^k
  - Construct half-plane $\mathcal{P}_k = \{x | h_k^\top(x - x_k) \leq 0\}$. $x^* \in \mathcal{P}_k \cap \mathcal{E}_k$ proved by convexity of $f(x)$
  - $\mathcal{E}_{k+1} = \{x | (x - x_{k+1})P_{k+1}^{-1}(x - x_{k+1}) \leq 1\}$ as minimum volume ellipsoid covering $\mathcal{P}_k \cap \mathcal{E}_k$

$$x_{k+1} = x_k - \frac{P_k h_k}{(d+1)\sqrt{h_k^\top P_k h_k}},$$

$$P_{k+1} = \frac{d^2}{d^2 - 1}\left(P_k - \frac{2P_k h_k h_k^\top P_k}{(d+1)h_k^\top P_k h_k}\right)$$

Classical Ellipsoid Method for Online Learning (CELLIP)

- A feasibility problem – find a solution that is close to the γ-margin classifier u given sequentially received training examples
- $\mathcal{A}_t = \{z \in \mathbb{R}^d | y_i x_i^\top z \geq a\gamma, i = 1, \ldots, t\}$ includes all the classifiers that are able to classify with margin $a\gamma$ training examples received so far
- To efficiently represent $\mathcal{A}_t$, we construct an ellipsoid

$$\mathcal{E}_t = \{z \in \mathbb{R}^d | (z - w_t)^\top P_t^{-1} (z - w_t) \leq 1\}$$

  such that $\mathcal{E}_t \supseteq \mathcal{A}_t$
- Now our goal is to efficiently reduce $vol(\mathcal{E}_t)$, since $\mathcal{E}_t \supseteq \mathcal{A}_t \supseteq \mathcal{B}$

Classical Ellipsoid Method for Online Learning (CELLIP)

Efficiently update the ellipsoid $\mathcal{E}_t$ given a misclassified example

- Assume $x_t \in \mathbb{R}^d$ is misclassified by w_t : $y_t w_t^\top x_i \leq 0$
- $\mathcal{C}_t = \{z \in \mathbb{R}^d | y_t x_t^\top z \geq a\gamma\}$: the half plane generated by x_t, ($u \in \mathcal{C}_t \cap \mathcal{E}_t$ since $y_t u^\top x_t \geq \gamma$)
- Rewrite the set $\mathcal{C}_t$ as

$$\mathcal{C}_t = \{z \in \mathbb{R}^d | \alpha_t - g_t^\top (z - w_t) \leq 0\}$$

$$\alpha_t = \frac{a\gamma - y_t w_t^\top x_t}{\sqrt{x_t^\top P_t x_t}}, \quad g_t = \frac{y_t x_t}{\sqrt{x_t^\top P_t x_t}}$$

Note that $\alpha_t \geq 0$ since $y_t w_t^\top x_t \leq 0$ and $g_t^\top P_t g_t = 1$.

Classical Ellipsoid Method for Online Learning (CELLIP)

A family of updating equations for w_t and P_t that ensures $\mathcal{E}_{t+1} \supseteq \mathcal{E}_t \cap \mathcal{C}_t$

Theorem 1 *Given a misclassified instance* (x_t, y_t)*, the following updating equations for* w_{t+1} *and* P_{t+1} *will guarantee that the resulting new ellipsoid* $\mathcal{E}_{t+1}$ *covers the intersection* $\mathcal{E}_t \cap \mathcal{C}_t$*:*

$$\begin{aligned} w_{t+1} &= w_t + (\alpha_t + \rho) P_t g_t \\ P_{t+1} &= \mu^2 P_t + ([1 - \alpha_t - \rho]^2 - \mu^2) P_t g_t g_t^\top P_t \end{aligned}$$

if parameter $\rho > 0$ *and* $\mu > 0$ *satisfy the following constraint*

$$\frac{1 - \alpha_t^2}{\mu^2} + \frac{\rho^2}{(1 - \alpha_t - \rho)^2} \leq 1$$

Classical Ellipsoid Method for Online Learning (CELLIP)

INPUT:

- $\gamma \geq 0$: the desired classification margin
- $a \in [0, 1]$: a tradeoff parameter

INITIALIZE: $w_1 = \mathbf{0}$ and $P_1 = (1 + (1 - a)\gamma)I_d$

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

  receive an instance x_t

  predict its class label: $\hat{y}_t = \text{sign}(w_t^\top x_t)$

  receive correct class label y_t

  if $y_t \neq \hat{y}_t$ **then**

    compute w_{t+1} and P_{t+1} ($\rho = 0$ and $\mu = \sqrt{1 - \alpha_t^2}$)

$$
\begin{aligned}
w_{t+1} &= w_t + \alpha_t P_t g_t \\
P_{t+1} &= (1 - \alpha_t^2)P_t - 2\alpha_t(1 - \alpha)P_t g_t g_t^\top P_t
\end{aligned}
$$

  else

    $w_{t+1} \leftarrow w_t$ and $P_{t+1} \leftarrow P_t$

  end if

end for

Mistake Bound for CELLIP

Theorem 2 *Let $\mathcal{D} = \{(x_i, y_i), i = 1, 2, \ldots, T\}$ be the set of training examples. Assume all the examples are normalized, i.e., $\|x_i\|_2 \leq 1$. We assume that there exists an classifier $u \in \mathbb{R}^d$ with $\|u\|_2^2 = 1$ that is able to classified all the training examples in $\mathcal{D}$ with a margin $0 \leq \gamma \leq 1$, i.e., $y_i u^\top x_i \geq \gamma$ for any (x_i, y_i) in $\mathcal{D}$. The number of mistake M made by CELLIP when learning from $\mathcal{D}$ (Algorithm) is upper bounded by*

$$M \leq \frac{2\log(1-a) + 2\log\gamma - \log(1+(1-a)\gamma)}{\log\left(1 - a^2\gamma^2/(1+(1-a)\gamma)^2\right)}$$

Address Inseparable Case : an improved ellipsoid method

Can not cast online learning as a feasibility problem since no classifier can classify all the instances correctly

- Treat w_t and P_t as summarization of received training examples
- w_{t+1} is a linear combination of the training examples received in the first t trials

$$P_{t+1}^{-1} = \frac{1}{1-\alpha_t^2} P_t^{-1} + \frac{2\alpha_t}{(1-\alpha_t)^2(1-\alpha_t)} g_t g_t^\top$$

$$P_{t+1}^{-1} = \theta_0 P_1^{-1} + \sum_{i=1}^{t} \theta_i g_i g_i^\top \propto \theta_0 P_1 + \sum_{i=1}^{t} \xi_i x_i x_i^\top$$

where θ_i and ξ_i are functions of $\{\alpha_j\}_{j=i}^{t}$.

- P_t^{-1} is a weighted covariance matrix that stores the second order information of training examples

Address Inseparable Case : Improved Ellipsoid Method

- Modify the updating equation for P_t as

$$P_{t+1} = \frac{1}{1-c_t}(P_t - c_t P_t g_t g_t^\top P_t)$$

where $c_t \in [0, 1]$.

- Set $c_t = cb^{t-1}$ where $0 \leq c, b \leq 1$ are two constants.

$$P_{t+1}^{-1} = (1-c_t)P_t^{-1} + c_t g_t g_t^\top$$

- P_{t+1}^{-1} is a mixture of matrices P_t^{-1} and $g_t g_t^\top$.
- Given $c_t = cb^{t-1}$, P_{t+1} is a weighted sum of $x_i x_i$ where the weight for $x_i x_i$ decays exponentially in t
- Vary c and $b \rightarrow$ adjust "memory" of P_t. The smaller b is, the shorter the memory is

Improved Ellipsoid Method (IELLIP) for Online Learning

INPUT:

- $\gamma \geq 0$: the desired classification margin
- $0 \leq c, b \leq 1$: parameters controlling the memory of online learning

INITIALIZE: $w_1 = \mathbf{0}$ and $P_1 = I_d$

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

  receive an instance x_t

  predict its class label: $\hat{y}_t =$ sign$(w_t^\top x_t)$

  receive correct class label y_t

  if $y_t \neq \hat{y}_t$ **then**

    compute w_{t+1} and P_{t+1} using the modified updating rule

  else

    $w_{t+1} \leftarrow w_t$ and $P_{t+1} \leftarrow P_t$

  end if

end for

Mistake Bound for the Improved Ellipsoid Method

- Measure the progress of online learning by

$$q_t = (u - w_t)^\top P_t^{-1}(u - w_t)$$

where u is some optimal classifier; P_t^{-1} measures the distance between u and w_t

Theorem 3 *Let $\mathcal{D} = \{(x_i, y_i), i = 1, 2, \ldots, T\}$ be the set of training examples. Let u be the optimal classifier with norm $|u|_2^2 = 1$. Assume all the examples are normalized, i.e., $\|x_i\|_2 \leq 1$. If c and b satisfy $c + b < 1$, the number of mistakes made by IELLIP is upper bounded by*

$$M \leq \frac{1}{\gamma^2} + \frac{2}{\gamma}\frac{1-b}{1-b-c}\sum_{i=1}^{T} l_i(u)$$

where $l_i(u) = \max(0, \gamma - u^\top x_i)$.

Extend the Ellipsoid Method to Multi-label Learning

Follow the framework by (Crammer et al.)

- Given x assigned to a subset of classes Y
- Weight vectors for K classes $w_i \in \mathbb{R}^d, i = 1, \ldots, K$
- Classification margin $\eta(W; x, Y) = \min_{z \in Y} w_z^\top x - \max_{z \notin Y} w_z^\top x$
- Loss function $l(W; x, Y) = \max(0, \gamma - \eta(W; x, Y))$
- Construct vector $v = (w_1, \ldots, w_K)$
- Define class indices $a_i = \max_{y \notin Y_i} w_y^\top x_i$, and $b_i = \min_{y \in Y_i} w_y^\top x_i$

  for misclassified (x_i, Y_i), i.e., $\eta(W; x_i, Y_i) \leq 0$

Extend the Ellipsoid Method to Multi-label Learning

- Construct a big vector $z_i \in \mathbb{R}^{K\times d}$

$$z_i^j = \begin{cases} x_i^k & j = (b_i - 1)d + k \\ -x_i^k & j = (a_i - 1)d + k \\ 0 & \text{otherwise} \end{cases}$$

- Construct a half plane $\mathcal{P}_t$ for each misclassified example z_t

$$P_t = \{v \in \mathbb{R}^{K\times d} | \alpha_t - (v - v_t)^\top g_t \leq 0\}$$

  where α_t and g_t are identical the expressions in (1) except that $y_t x_t$ is replaced by z_t.
- Directly extend to multi-label learning, by definition of classifier v, misclassified example z_i, α_t and g_t

Outline

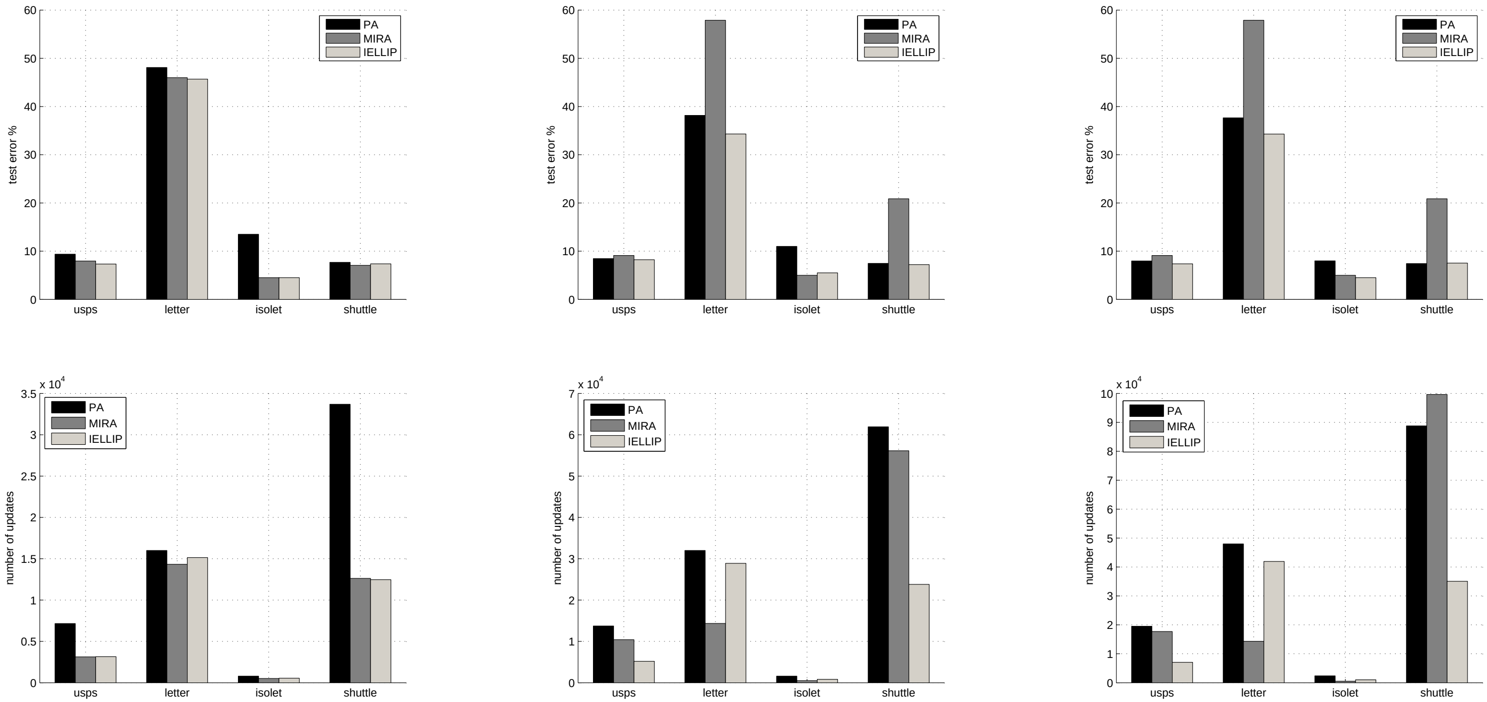
- Motivation
- Algorithm
- **Evaluation**
  - Datasets, Baseline Methods & Evaluation Metrics
  - Results of Multiclass Classification

Experiment Setup

- We evaluate IELLIP, since CELLIP cannot handle inseparable cases and outperformed by IELLIP
- Initialize P_1 as identity matrix at the scale of 0.1; randomly initialize $\mathbf{w}$ around the origin
- Datasets : USPS, UCI Letter, UCI Isolet, UCI Shuttle
- Baselines : Online Passive-Aggressive Algorithm(PA)(Crammer et al., 06) and Margin Infused Relaxed Algorithm (MIRA)(Crammer & Singer, 03)
- All use linear classifiers. Margin $= 0.1$
- Test error: # mistake made on a given sequence normalized by its length

Results of Multiclass Classification

- PA better than generalized Perceptron algorithms due to the aggressiveness (large margins)
- test error of IELLIP better than the best of PA and MIRA.
- a smaller # updates by IELLIP to achieve better test error than PA and MIRA.
- IELLIP is more efficient than baselines



Conclusion

- This work is the **first** attempt to **explicitly** represent **version space**
- Represent the version space by the ellipsoid method, capturing all classifiers consistent with training examples
- Same mistake bounds with perceptron up to a const. factor
- Shape matrix stores more information of training examples, and provides additional controls
- Geralized to multi-label learning
- Empirical effectiveness of IELLIP, compared with state-of-the-art online learners