Journal of Machine Learning Research 1 (2014) 1-48 Submitted 4/00; Published 10/00

Paired passive aggressive for ranking and classification

Author I AUTHOR1@SOMEWHERE
Department of
University of
City, WA 98195-4322, USA

Author II AUTHOR2@SOMEWHERE
Department of
University of
City, WA 98195-4322, USA

Editor: some editor

Abstract

This paper describes the ....
Keywords: Passive aggressive, AUC, MAP

1. Introduction

The importance of the problem

Related work

2. Problem Setting

In this section we introduce the notation used throughout the paper and describe our problem setting. Vectors are denoted by lower case bold face letters (e.g. $\boldsymbol{x}$ and $\boldsymbol{w}$) where the i^{th} element of the vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_i$. The hinge function is denoted by $[x]_+ = max\{0, x\}$. Sets are denoted by capital curly letters (e.g $\mathcal{J}$).

We are interested in the case where at each time point t we receive a batch of n_t sample and than choose how to update the vector weights $\boldsymbol{w}$. At each time point t we solve an optimization problem which performs a trade off between two things. First, it aims that the new solution $\boldsymbol{w}$ will be close to the former weight vector w_t. Second we prefer to classify all the samples provided at time t correctly with a margin of 1.

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w}_t||^2 + C\sum_{i=1}^{n_t} \xi_i \\ \text{subject to} \quad & 1 - \boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{y}_i \le \xi_i, \;\; i = 1, \ldots, n_t. \end{aligned}$$

There are some benefits of an update which uses several samples for the update. First, in cases where the data is unbalanced, using a balanced updating scheme that introduce an equal number of samples at time point t we can come up with guaranties both for the classification mistakes and for the AUC. Second, an update rule that uses several samples at a single time point t is internally tuned since the we need to advanced $\boldsymbol{w}$ in a way that is agreeable with the samples at time t.

3. Average Classification Loss

Define the loss and the problem

Derive the update rule

Theorem 1: the expected loss is less than the average loss

Theorem 2: 1-AUC is bounded

Theorem 3: Show that classification is correct and $wx+ > 0$ while $wx- < 0$ after the update

4. Double-slack

We are interested in solving a passive aggressive style problem only that we are shown k_t examples at time t. Specificly we are interested in the case that the samples arrive from different classes. We show that by using a balanced regiem we can provide bounds for the AUC and for a multiclass AUC. The specific method we choose to optimize the problem have a great deal of implication on the solution where different steps will yield different results.

$$\begin{aligned}
&\underset{\boldsymbol{w}}{\text{minimize}} && \frac{1}{2}||\boldsymbol{w}-\boldsymbol{w}_t||^2 + C\sum_{i=1}^{k_t}\xi_i \\
&\text{subject to} && 1-\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{y}_i \le \xi_i, \;\; i=1,\dots,k_t.
\end{aligned}$$

The dual problem is

$$\begin{aligned}
&\underset{\boldsymbol{\alpha}}{\text{maximize}} && \frac{1}{2}||\sum_{i=1}^{k_t}\alpha_i\boldsymbol{x}_i\boldsymbol{y}_i||^2 + \sum_{i=1}^{k_t}\alpha_i(1-\boldsymbol{w}_t^T\boldsymbol{x}_i\boldsymbol{y}_i) \\
&\text{subject to} && 0 \le \alpha_i \le C, \;\; i=1,\dots,k_t.
\end{aligned}$$

We aim to maximize the dual function so at each step we choose a set of indices $\mathcal{J}$ to increase using the same step.

$$\alpha_j = \alpha_j + \delta, \;\; j \in \mathcal{J}$$

We derive the following δ

$$\delta = max(L_b, min(U_b, \frac{|\mathcal{J}| - \boldsymbol{w}_t^T\sum_{j\in\mathcal{J}}\boldsymbol{x}_j y_j}{||\sum_{j\in\mathcal{J}}\boldsymbol{x}_j y_j||^2}))$$

Where $L_b = max_{j\in\mathcal{J}}(-\alpha_j)$ and $U_b = min_{j\in\mathcal{J}}(C-\alpha_j)$ which appear since each of the updated α_j needs to keep its constrain $0 \le \alpha_j \le C$. It is possible that a step that advances all $j \in \mathcal{J}$ does not exists because U_b could be smaller than L_b. We will always be able to perform at least one such step since we initialize α_i as zero. If we partition the set of samples and at each iteration we use a different partition, we are guaranteed that $L_b \leq U_b$ since we advance all the dual variables in each partition with the same steps.

We can think of this as updating a new vector $\boldsymbol{x}^{\mathcal{J}} = \sum_{j \in \mathcal{J}} \boldsymbol{x}_j y_j$ where $y^{\mathcal{J}} = 1$. Only here $\boldsymbol{x}^{\mathcal{J}}$ should be correct with a margin of $|\mathcal{J}|$. We than update $\boldsymbol{w}$ using $\boldsymbol{x}^{\mathcal{J}}$ with the step size δ. This analogy is not entirely correct because of the constrains we have on the δ.

Notice that by using multiple items in $\mathcal{J}$ we make a statement about their linear combination and not any of them individually. For example, when we update two items a positive sample $\boldsymbol{x}^+$ and a negative sample $\boldsymbol{x}^-$ forcing that their sum should be classified positive $0 \leq \boldsymbol{w}^T(\boldsymbol{x}^+ - \boldsymbol{x}^-)$ we actually argue about their order we say that their difference should be kept positive or that $\boldsymbol{w}^T\boldsymbol{x}^- \leq \boldsymbol{w}^T\boldsymbol{x}^+$.

Using various sets of $\mathcal{J}$ and various number of iteration at time t we propose several update rules:

In the case where at time t we are provided with two samples ($k_t = 2$), $\boldsymbol{x}^+, \boldsymbol{x}^-$.
I. PA-DCA - Iterate until convergence at each iteration choose a single sample ($|\mathcal{J}| = 1$).
II. PA-sequential - Iterate only once for $\boldsymbol{x}^+$ and than once for $\boldsymbol{x}^-$.
III. PA-AUC - Iterate only once using both samples $\mathcal{J} = \{\boldsymbol{x}^+, \boldsymbol{x}^-\}$.
IV. PA-correctMistakes - Iterate only once use only the samples that failed to achieve correct classification with the margin. $\mathcal{J} = \{\boldsymbol{x}^+, \boldsymbol{x}^-\}$ *or* $\{\boldsymbol{x}^+\}$ *or* $\{\boldsymbol{x}^-\}$.

In the case where we are presented k samples.
I. PA-DCA - Iterate until convergence at each iteration choose a single sample ($|\mathcal{J}| = 1$).
II. PA-sequential - Iterate only once for each sample.
III. PA-maxViolators - Iterate only once. Here $\mathcal{J}$ contains the positive sample that caused the highest loss and the negative sample that caused the highest loss.
IV. PA-correctMistakes - Iterate only once use only the samples that failed to achieve correct classification with the margin. $\mathcal{J} = \{i | 0 \leq l_{w_t}(\boldsymbol{x}_i, y_i)\}$.

Theorem 4: convergence of DCA

Theorem 5: classification errors –¿ number of mistakes

Theorem 6: As in the case of the classical passive aggressive if our step is not caped by C after the update we will correctly classify the samples. In case where we are caped by C the loss of the samples we choose to update will decrease. But we are not guarantied a correct classification.
Show that classification is correct and $wx+ > 0$ while $wx- < 0$ after the update when it is not caped.

Theorem 7: from Theorem 5 it follows that 1-AUC is bounded

5. Calibrated Multilabel Classification and Ranking

Calibrated separation ranking loss was proposed by **?**

We are interested in the case where we have more than 2 classses and these classes are unbalanced. For this multliclass scenario we follow the $mutliclassAUC$ suggested by ... We define $AUC_{allpairs}$ by:

$$AUC_{allpairs} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{K-1}\sum_{l\neq k}(AUC_{w^k}(C^k, C^l))$$

Where C_i denote all the samples from class i, and $AUC_{w^k}(C^k, C^l)$ is the AUC performed on the samples from the k class and samples from the l class using the classifier $\boldsymbol{w}^k$ train using the samples from the k class as positivies.

$$1 - AUC_{allpairs} \leq \frac{1}{K}(\sum_{k=1}^{K} E(M_{w^k}(C^k)) + \frac{1}{k-1}(\sum_{l\neq k} E(M_{w^k}(C^l)))$$

This suggest that at time t to train our classifier $\boldsymbol{w}^k$ we need to present to it a positive sample from the k class and average negative step from the other K-1 classes.

$$\boldsymbol{w}_t^k = \boldsymbol{w}_{t-1}^k + \alpha_t^k \boldsymbol{x}_t^k - \frac{1}{k-1}\sum_{l\neq k}\alpha_t^l \boldsymbol{x}_t^l$$

The update rule is derivied using $\mathcal{J} = \{i|0 \leq l_{w_t}(\boldsymbol{x}_i, y_i)\}$, we choose to update only the dual variables from the samples that we failed to be classified using the required margin.

$$\delta = max(L_b, min(U_b, \frac{|\mathcal{J}| - \boldsymbol{w}_t^T \sum_{j\in\mathcal{J}} \boldsymbol{x}_j y_j}{||\sum_{j\in\mathcal{J}} \boldsymbol{x}_j y_j||^2}))$$

Theorem 8: correct multilabels are above the incorrect set of labels

6. Experiments

Synthetic data

Discriminative keyword spotting with algo 1 and algo 2

Mutliclass classification evaluated using $AUC_{all\ pairs}$ and $AUC_{one\ vs\ all}$

LETOR3 data for ranking with algo 1 and algo 2

multi label - Reuters

Acknowledgments

We would like to acknowledge support for this project ....

Appendix A.

In this appendix we prove the following theorem from Section X.X:

Theorem *First we will show that the 1 - AUC is bounded by the mean of the errors in the first class and the mean number of errors in the second class.* $1-AUC \leq E(M^{+})+E(M^{-})$

Proof

$$1 - AUC = \frac{1}{|X^{-}||X^{+}|} \sum_{\substack{x^{+} \in X^{+} \\ x^{-} \in X^{-}}} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^{+} \leq \boldsymbol{w}^T \boldsymbol{x}^{-}} \quad \leq$$

$$\frac{1}{|X^{-}||X^{+}|} \sum_{\substack{x^{+} \in X^{+} \\ x^{-} \in X^{-}}} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^{+} \leq 0} \ + \ \mathbb{1}_{0 \leq \boldsymbol{w}^T \boldsymbol{x}^{-}} \quad =$$

$$\frac{1}{|X^{+}|} \sum_{\boldsymbol{x}^{+} \in X^{+}} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^{+} \leq 0} \ + \ \frac{1}{|X^{-}|} \sum_{\boldsymbol{x}^{-} \in X^{-}} \mathbb{1}_{0 \leq \boldsymbol{w}^T \boldsymbol{x}^{-}} \ = \ E(M^{+}) + E(M^{-}) \quad (1)$$

■

Next we will show that we mutlicalss AUC which uses the mean AUC of all pairs can also be bounded by mean classification mistakes.

Theorem $1 - AUC_{\textit{all pairs}}$ *can be bounded by:*
$1 - AUC_{\textit{all pairs}} \leq \frac{1}{K} \sum\limits_{k=1}^{K} \left(E_{X^k}(M_{\boldsymbol{w}^k}) + \frac{1}{k-1} (\sum\limits_{l \neq k} E_{X^l}(M_{\boldsymbol{w}^k}) \right)$
Where $E_{X^l}(M_{\boldsymbol{w}^k})$ *are the expected number of mistakes from class* l *that are made by the classifier that was trained to classify class* k *as postivies.*

Proof

$$1 - AUC_{\textit{all pairs}} = 1 - \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{l \neq k} (AUC_{\boldsymbol{w}^k}(X^k, X^l)) \quad \leq$$

$$\frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{l \neq k} E_{X^k}(M_{\boldsymbol{w}^k}) + E_{X^l}(M_{\boldsymbol{w}^k}) \quad =$$

$$\frac{1}{K} \sum_{k=1}^{K} \left(E_{X^k}(M_{\boldsymbol{w}^k}) + \frac{1}{K-1} \sum_{l \neq k} E_{X^l}(M_{\boldsymbol{w}^k}) \right) \quad (2)$$

■

Theorem *The sum of the average mistake in the two classes can be bounded*
$E_{X^{+}}[M] + E_{X^{-}}[M] \leq \ max\{1/C, R^2\}(\ 2C(E_{X^{+}}[l^*] + E_{X^{-}}[l^*]) + \frac{1}{|X^{-}||X^{+}|} ||\mathbf{u}||^2 \ )$

Proof. We use the inequalty from the PA paper which states that M - the number of mistakes made by introducing samples from X^{+} and samples from X^{-} can be bounded :

$$M = \sum_{\substack{x^+ \in X^+ \\ x^- \in X^-}} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^+ \leq 0} \ + \ \mathbb{1}_{0 \leq \boldsymbol{w}^T \boldsymbol{x}^-} \leq$$

$$max\{1/C, R^2\}(\ 2C \sum_{\substack{x^+ \in X^+ \\ x^- \in X^-}} l^* + ||\mathbf{u}||^2 \ ) \quad (3)$$

Dividing by the number of samples we get:

$$E_{X^+}[M] + E_{X^-}[M] = \frac{1}{|X^+|} \sum_{x^+ \in X^+} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^+ = \leq 0} \ + \ \frac{1}{|X^-|} \sum_{x^- \in X^-} \mathbb{1}_{0 \leq \boldsymbol{w}^T \boldsymbol{x}^-} =$$

$$\frac{1}{|X^+||X^-|} \sum_{\substack{x^+ \in X^+ \\ x^- \in X^-}} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^+ \leq 0} \ + \ \mathbb{1}_{0 \leq \boldsymbol{w}^T \boldsymbol{x}^-} \leq$$

$$\frac{1}{|X^+||X^-|} max\{1/C, R^2\}(\ 2C \sum_{\substack{x^+ \in X^+ \\ x^- \in X^-}} l^* + ||\mathbf{u}||^2 \ ) =$$

$$max\{1/C, R^2\}(\ 2C(E_{X^+}[l^*] + E_{X^-}[l^*]) + \frac{1}{|X^-||X^+|} ||\mathbf{u}||^2 \ ) \quad (4)$$

■

The next result extend this inequalty to the multiclass case. First we denote the set of samples from the k class using X^k. We are now interested in matching each samples from the k class with each sample from the other $K-1$ classes. By iterating the paired classes we get:

$$\frac{1}{K-1} \sum_{l \neq k} E_{X^k}[M] + E_{X^l}[M] \leq$$

$$\frac{1}{K-1} \sum_{l \neq k} max\{1/C, R^2\}(\ 2C(E_{X^k}[l^*] + E_{X^l}[l^*]) + \frac{1}{|X^k||X^l|} ||\mathbf{u}||^2 \ ) \quad (5)$$

Rearranging we get:

$$E_{X^k}[M] + \frac{1}{K-1} \sum_{l \neq k} E_{X^l}[M] \leq$$

$$max\{1/C, R^2\} \left(2C E_{X^k}[l^*] + \frac{2C}{K-1} \left(\sum_{l \neq k} E_{X^l}[l^*]) \right) + ||\mathbf{u}||^2 \frac{1}{|X^k|} (\sum_{l \neq k} \frac{1}{|X^l|}) \right) \quad (6)$$

By averaging the k different classifiers, each with it own vector $\boldsymbol{u}$ and l^* we can bound the size of $1 - AUC_{all\ pairs}$. ■

References