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, Multiclass, unbalanced

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 $\boldsymbol{w}$) The hinge function is denoted by $[x]_+ = max\{0, x\}$. Sets of indices are denoted by capital curly letters (e.g. $\mathcal{J}$). We denote samples which arrive from the k class using superscript (e.g. $\boldsymbol{x}^k$). Subscript will denote the time point the samples is introduced (e.g. $\boldsymbol{x}_t^k$ is a vector from the k class at time t).

We are interested in the case where at each time point t we receive a batch of k_t sample and than choose how to update the vector weights $\boldsymbol{w}$. Specifically, we are interest in the case where each of the k_t samples arrives from a different class. 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. The tradeoff is controled by the hyperparameter C.

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w}_t||^2 + C\sum_{k=1}^{k_t}\xi^k \\ \text{subject to} \quad & 1 - \boldsymbol{w}^T\boldsymbol{x}^k y^k \leq \xi^k, \;\; k = 1, \ldots, k_t. \end{aligned}$$

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. In a way the other classes are controlling the step size that is made. Much like the case in multibatch stochastic gradient descent, one of the benefits is that the steps are more moderate and there is less variability in each small step. Only here we are using the variability between different classes and not the variability within the class.

3. Average Classification Loss

We first start by examining the case where we have only two classes. We denote our classes with X^+ and X^-.

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}||\boldsymbol{w}-\boldsymbol{w}_t||^2 + C(\xi^+ + \xi^-) \\ \text{subject to} \quad & 1 - \boldsymbol{w}^T\boldsymbol{x}^+ \leq \xi^+ \\ & 1 + \boldsymbol{w}^T\boldsymbol{x}^- \leq \xi^- \\ & 0 \leq \xi^+, \xi^- \end{aligned}$$

The dual problem is

$$\begin{aligned} \underset{\alpha^+,\alpha^-}{\text{maximize}} \quad & \frac{1}{2}||\alpha^+\boldsymbol{x}^+ - \alpha^-\boldsymbol{x}^-||^2 + \alpha^+(1-\boldsymbol{w}_t^T\boldsymbol{x}^+) + \alpha^-(1+\boldsymbol{w}_t^T\boldsymbol{x}^-) \\ \text{subject to} \quad & 0 \leq \alpha^+, \alpha^- \leq C, \end{aligned}$$

and the relation between the dual variables and the primial variables is: $\boldsymbol{w} = \boldsymbol{w}_t + \alpha^+\boldsymbol{x}^+ - \alpha^-\boldsymbol{x}^-$.

First notice that by adding the two constrains we get $1 - \boldsymbol{w}^T\frac{(\boldsymbol{x}^+ - \boldsymbol{x}^-)}{2} \leq \frac{\xi^+ + \xi^-}{2}$. By replacing the two constrain with this new constrain we get a type of problem that we refer to as PA_{AUC} Keshet et al. (2009). By solving this problem we aim at insuring that the $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ are ranked correctly ($\boldsymbol{w}^T\boldsymbol{x}^- \leq \boldsymbol{w}^T\boldsymbol{x}^+$). By solving the problem with the two constrains, we are not only guarantied that we $\boldsymbol{x}^+$ will be classified higher than $\boldsymbol{x}^-$ but also that we classify the two samples correctly.

AUC is often used in the case where we aim at classifying between two imbalanced classes. In such scenarios, minimizing the number of overall error is usually not helpful. For example, in the case where we have 1 positive for every 99 negative examples, a classifier which always predict negative will make 1% errors. Usually, this type of classifier in not what we are interested in. AUC can be interpreted as the probability that a positive samples will be ranked higher than a negative sample. It is easy to see how this definition of AUC as a probability ranker helps in the imbalanced case: By maximizing the AUC, we always treat a pair of samples, one positive and one negative in a balanced manner, regardless of their distribution in the overall population. This insures that the classes are treated as equals and that they are both equally represented to our classifier. However, maximizing the AUC

insures only the order of the two samples; we are not guarantied a correct classification. And, is it so often happen that algorithms which maximize the AUC fail to provide correct classification and a dynamic threshold term is introduced. Many times, when we are in the scenario where we want to classify a pair of imbalanced classes, what we actually want is to minimize the the mean number of classification errors in the first class and to minimize number of classification errors in the second class. First we will show that the AUC-error (1 - AUC) can be bounded by the mean number of mistakes made in the first class and the mean number of mistakes made in the second class. This will later help us develop an update that is both good at classification and can be used in the scenario of imbalanced classes.

Theorem 1 *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

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

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

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

■

We will next show that we can provide bounds for the sum of the mean number of errors using several update rules.

Theorem 3: Show that classification is correct and $\boldsymbol{x}^+ > 0$ while $\boldsymbol{x}^- < 0$ after the update - this is not correct since we are balancing with w_t from the former step. This would have been correct if we would have posed this is a feasibility problem (perceptron style).

4. Update rules

We are interested in solving a passive aggressive style problem only that we are shown k_t examples at time t. Specifically, 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.

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w}_t||^2 + C \sum_{k=1}^{k_t} \xi^k \\ \text{subject to} \quad & 1 - \boldsymbol{w}^T \boldsymbol{x}^k \boldsymbol{y}^k \leq \xi^k, \;\; k = 1, \ldots, k_t. \end{aligned}$$

The dual problem is

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

This problem does not have a closed analytic solution, So we aim to maximize the dual function iterativly. One way of solving this is using *dual coordinate ascent* (DCA) on all the k_t samples. DCA will make iteration until convergence. However as noticed by ..., we can also make a small advancement that is not optimal but that will advance us in the right direction. For example the passive aggressive algorithm can be thought of as single iteration of DCA where we update each dual variable only once. We suggest that this can be extended by advancing with a step that is made using the joined information of several samples. We choose a set of indices $\mathcal{J}$ to increase using the **same** step.

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

We derive the following τ

$$\tau = max(L_b, min(U_b, \frac{|\mathcal{J}| - \boldsymbol{w}_t^T \sum\limits_{j \in \mathcal{J}} \boldsymbol{x}^j y^j}{||\sum\limits_{j \in \mathcal{J}} \boldsymbol{x}^j y^j||^2}))$$

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 $0 \leq \alpha_j \leq C$. 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 $\boldsymbol{x}^{\mathcal{J}} = \sum\limits_{j \in \mathcal{J}} \boldsymbol{x}^j y^j$ 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 τ. 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. Multiclass and Multilabel

We are interested in the case where we have more than 2 classes and these classes are unbalanced. For this multliclass scenario we follow the *mutliclassAUC* suggested by Hand and Till (2001) We define $AUC_{all\ pairs}$ by:

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

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 positives.

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

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 extension to multilabel is stright forward: At time t select a sample from each of the K classes. Then, for each class we balance the positive and the negative samples. Let

P_t^k denote the set of samples at time t that are labeled positivly with class k and N_t^k the set of samples that are not label with class k.

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

The update rule is derived 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.

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

Calibrated separation ranking loss was proposed by Yuhong Guo
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

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

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

$$\frac{1}{|X^+|} \sum_{\boldsymbol{x}^+ \in X^+} \mathbb{1}_{\boldsymbol{w}^T \boldsymbol{x}^+ \leq 0} \; + \; \frac{1}{|X^-|} \sum_{\boldsymbol{x}^- \in X^-} \mathbb{1}_{0 \leq \boldsymbol{w}^T \boldsymbol{x}^-} \; = \; E(M^+) + E(M^-) \quad (2)$$

■

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 $1 - AUC_{all\ pairs}$ *can be bounded by:*

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

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 positives.*

Proof

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

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

$$\frac{1}{K} \sum_{k=1}^{K} \left(E_{X^k}(M_{\boldsymbol{w}^k}) + \frac{1}{K-1} \sum_{l \neq k} E_{X^l}(M_{\boldsymbol{w}^k}) \right) \quad (3)$$

■

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 inequality 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 :

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

$$max\{1/C, R^2\}(\; 2C \sum_{\substack{x^+ \in X^+ \\ x^- \in X^-}} l^* + ||\mathbf{u}||^2 \;) \quad (4)$$

Dividing by the number of samples we get:

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

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

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

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

■

The next result extend this inequality 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:

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

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

Rearranging we get:

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

$$max\{1/C, R^2\} \left(2C E_{X^k}[l^*] + \frac{2C}{K-1} \left(\sum_{l \neq k} E_{X^l}[l^*] \right) + ||\mathbf{u}||^2 \frac{1}{|X^k|} (\sum_{l \neq k} \frac{1}{|X^l|}) \right) \quad (7)$$

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

David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2): 171–186, November 2001. ISSN 1573-0565. doi: 10.1023/A:1010920819831. URL http://link.springer.com/article/10.1023/A:1010920819831.

Joseph Keshet, David Grangier, and Samy Bengio. Discriminative keyword spotting. *Speech Communication*, 51(4):317–329, April 2009. ISSN 01676393. doi: 10.1016/j.specom.2008.10.002. URL http://dl.acm.org/citation.cfm?id=1507768.1507929.

Dale Schuurmans Yuhong Guo. Adaptive Large Margin Training for Multilabel Classification. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.228.6549.