Online Learning and Active Learning

A comparative study with Support Vector Machine (SVM)

Ezukwoke K.I[1], Zareian S.J[2]

[1, 2] Department of Computer Science
Machine Learning and Data Mining
{ifeanyi.ezukwoke, samaneh.zareian.jahromi}@etu.univ-st-etienne.fr
University Jean Monnet, Saint-Etienne, France

Abstract

Passive aggressive online learning is an extension of Support Vector Machine (SVM) to the context of online learning for binary classification. In this paper we consider the application of the algorithm on LibSVM dataset available from UCI. We also work on an improved version of the online learning algorithm called **Active learning** and we compare both algorithm to that of LibSVM. We formalize and model the kernel versions of online and active learning algorithm wih experimental comparisons.

Keywords
Passive aggressive online learning, Active learning, Support Vector Machine (SVM).

1 INTRODUCTION

Online learning is a binary classification algorithm which is an extension of Support Vector Machine (SVM). Unlike SVM where we update the weights using batches of data samples or all datasamples; here we only update the weights usng one example at a time. At every iteration, we only update the weight associated with the datasample until we reach the last index.

2 BINARY CLASSICATION

Binary classification is a classification problem involving only two defined categories. Given set of datasamples $\{x_i, y_i\}^N$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ where $\mathcal{Y} \in \{-1, +1\}$ for a binary classification problem.

2.1 Passive Agrressive algorithm

We present an optimization problem of the passive aggressive algorithm

$$\begin{cases} \underset{\mathbf{w} \in \mathbb{R}^N}{\arg\min} \frac{1}{2} \|\mathbf{w} - \mathbf{w}_t\|^2 \\ s.t \quad y_t(\mathbf{w} \cdot \mathbf{x}_t) \geq 1 \end{cases} \tag{1}$$

where $\mathbf{w}$ is the weight update. By introducing the Lagrangian multipliers we have

$$\mathcal{L} = \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + \tau(1 - y(\mathbf{w} \cdot \mathbf{x}_t)) \tag{2}$$

Observing the KKT optimality conditions we have that

$$\nabla \mathcal{L} = 0 \tag{3}$$

So that

$$\nabla_{\mathbf{w}} \mathcal{L} = \mathbf{w} - \mathbf{w}_t - \tau y x \tag{4}$$

$$\mathbf{w} = \mathbf{w}_t + \tau y x \tag{5}$$

Substituting equation (5) into (2) we have that

$$\mathcal{L}(\tau) = \frac{1}{2}\tau^2 y^2 |\mathbf{x}\|^2 + \tau(1 - y(\mathbf{w} \cdot \mathbf{x}_t + yx^2)) \tag{6}$$

given that $y \in \{-1, 1\}$ we can rewrite the above as

$$\mathcal{L}(\tau) = \frac{1}{2}\tau^2 \|\mathbf{x}\|^2 + \tau(1 - y(\mathbf{w} \cdot \mathbf{x}_t + yx^2)) \tag{7}$$

$$\mathcal{L}(\tau) = -\frac{1}{2}\tau^2 \|\mathbf{x}\|^2 + \tau(1 - y(\mathbf{w} \cdot \mathbf{x}_t)) \tag{8}$$

By differentiating with respect to τ we have that

$$\nabla_\tau \mathcal{L} = -\tau\|\mathbf{x}\|^2 + (1 - y(\mathbf{w} \cdot \mathbf{x}_t)) = 0 \tag{9}$$

where

$$\tau = \frac{1 - y(\mathbf{w} \cdot \mathbf{x}_t)}{\|\mathbf{x}\|^2} \tag{10}$$

The above equation is the hard margin formulation for passive aggressive algorithm. By introducing a error term or slack variable, we can transform the hard margin formulation to a soft margin. We write the soft margin formulation of passive aggressive algorithm as

$$\begin{cases} \underset{\mathbf{w}\in\mathbb{R}^N, \xi\geq 0}{\arg\min} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + C\sum_{i=1}^{N} \xi \\ s.t \quad y_t(\mathbf{w} \cdot \mathbf{x}_t) \geq 1 + \sum_{i=1}^{N} \xi \\ \qquad \xi \geq 0 \end{cases} \tag{11}$$

Following the same procedure for deriving τ as in the case of the hard margin; we also have that,

$$\tau = min\left\{C, \frac{1 - y(\mathbf{w} \cdot \mathbf{x}_t)}{\|\mathbf{x}\|^2}\right\} \tag{12}$$

Algorithm 1: Online Passive-Aggressive Algorithm

```
Input  : $\mathbf{X}, y$
Output : $\mathbf{w}$
begin
    $\mathbf{w} \leftarrow \mathbf{w}^0$;
    for $t = 1, 2, \ldots N$ do
        receive instance: $\mathbf{x} \in \mathbb{R}^n$;
        predict $\hat{y} = sign(\mathbf{w}_t \cdot \mathbf{x})$;
        correct label: $y_t \in \{-1, 1\}$;
        loss
          $l_t = max(0, 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}))$;
        compute $\tau_t$;
        update: $\mathbf{w}_t \leftarrow \mathbf{w} + \tau y_t \mathbf{x}_t$
    end
end
```

where τ takes the different form

1. $\tau = \frac{l_t}{\|x_t\|^2}$

2. $\tau = min\left\{C, \frac{l_t}{\|x_t\|^2}\right\}$

3. $\tau = \frac{l_t}{\|x_t\|^2 + \frac{1}{2C}}$

2.2 Active Learning

The active learning version of algorithm 1's objective is to minimize the number of labels to query using a probabilistic criterion (**Bernoulli random distribution**). This way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data.

Algorithm 2: Active Learning version of Algorithm 1

```
Input  : $\mathbf{X}, y$
Output : $\mathbf{w}$
begin
    $\mathbf{w} \leftarrow \mathbf{w}^0$;
    for $t = 1, 2, \ldots, N$ do
        receive instance: $\mathbf{x}_t \in \mathbb{R}^n$;
        Let $\hat{p} = \mathbf{w}_t \cdot \mathbf{x}$;
        predict $\hat{y} = sign(\hat{p})$;
        Draw a Bernoulli random
          variable $Z_t \in \{0, 1\}$ of
          parameter $\delta/(\delta + |\hat{p}|)$;
        if $Z_t = 1$ then
            Get label $y_t \in \{-1, 1\}$;
            Loss
              $l_t = max(0, 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}))$
              compute $\tau_t$;
            update: $\mathbf{w}_t \leftarrow \mathbf{w} + \tau y_t \mathbf{x}_t$
        else
            $\mathbf{w}_t \leftarrow \mathbf{w}$
        end
    end
end
```

3 KERNEL METHODS

3.0.1 Kernel Passive-Aggressive Online Learning Algorithm

Kernel methods introduces non-linearity into our algorithm by projecting our data into a Hilbert space ($\mathcal{H}$). We implement the kernel passive-aggressive algorithm by considering that any classifier can be defined as a weighted sum of seen examples.

4 EXPERIMENT

4.1 Dataset

4.2 Performance Aanlysis