Learning Drifting Data Using Selective Sampling

April 1, 2014

Objectives

Approaching the problem of shifting concept in an on-line learning classification setting we set the following objectives:

1. Detect the switch
2. If switch is undetected - assure that the additional regret it causes is small
3. No false detections

Problem Setting

We work under the following assumptions:

- $y_t \in \{\pm 1\}$
- $\boldsymbol{x}_t \in R^d$
- for $t \leq \tau$ holds $\mathrm{E}[y_t] = \boldsymbol{u}^\top \boldsymbol{x}_t$
- for $t > \tau$ holds $\mathrm{E}[y_t] = \boldsymbol{v}^\top \boldsymbol{x}_t$
- $\|\boldsymbol{x}_t\| = \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$

BBQ Algorithm

We submit prediction:

$$\hat{y}_t = \operatorname{sign}\left\{\boldsymbol{w}_t^\top \boldsymbol{x}_t\right\} \tag{1}$$

$\boldsymbol{w}_t$ is our estimation to the optimal linear classifier obtained by solving the following problem:

$$\boldsymbol{w}_t = \min_{\boldsymbol{w}\in R^d}\left\{\sum_{i=1}^{n}\left(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i\right)^2 + \|\boldsymbol{w}\|^2\right\} \tag{2}$$

with $n = N_t$ being the number of queries issued until round $t-1$

BBQ Algorithm

The solution to equation 2 is:

$$\boldsymbol{w}_t = \left(I + S_{t-1}S_{t-1}^T + \boldsymbol{x}_t\boldsymbol{x}_t^\top\right)^{-1} S_{t-1}Y_{t-1} \tag{3}$$

where $S_{t-1} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in R^{d\times n}$ and $Y_{t-1} = (y_1, ..., y_n) \in R^n$.
Another formulation:

$$\boldsymbol{w}_t = A_t^{-1} b_t \tag{4}$$

where $A_t = I + \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{x}_t\boldsymbol{x}_t^\top$ and $b_t = \sum_{i=1}^{n} y_i\boldsymbol{x}_i$

BBQ Algorithm - Querying Labels

We define:

$$r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t \tag{5}$$

A query will be issued at round t if $r_t > t^{-\kappa}$.

If $r_t \leq t^{-\kappa}$ the value of the label y_t will remain unknown.

Effect of Switch on BBQ Algorithm

In the normal setting the BBQ algorithm works well - with logarithmic regret:

$$R_T \leq O\left(d \ln T\right) \tag{6}$$

while maintaining significantly reduced amount of quired labels:

$$N_T \sim dT^{\kappa} \ln T \tag{7}$$

However as switch of the optimal classifier from $\boldsymbol{u}$ to $\boldsymbol{v}$ at round τ increases regret bound:

$$R_T \leq O\left(\|\boldsymbol{v} - \boldsymbol{u}\|^2 \tau^{2\kappa} \left(d \ln \tau\right)^2 d \ln T\right) \tag{8}$$

Effect of Switch on BBQ Algorithm

The increase in the regret bound is due to increase in the bound of the classifier's bias, after the switch:

$$B_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t - \mathrm{E}\left[\boldsymbol{w}_t^\top \boldsymbol{x}_t\right] \leq r_t + \sqrt{r_t} + N_\tau \|\boldsymbol{v} - \boldsymbol{u}\| \sqrt{r_t} \qquad (9)$$

Instead of

$$B_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t - \mathrm{E}\left[\boldsymbol{w}_t^\top \boldsymbol{x}_t\right] \leq r_t + \sqrt{r_t} \qquad (10)$$

prior to the switch

Using Selective Sampling to Overcome Switch

Selective sampling concept gives us confidence on our prediction.

The term r_t controls and both the bias from the optimal classifier and the instantaneous regret:

- If r_t is large, then in any case, switch or none, we can not assure low regret.
- If r_t is small, we should suffer low regret - meaning our prediction should be close enough to the optimal prediction. Unless a switch had occurred...

Using Selective Sampling to Overcome Switch

Main idea - use instances with small r_t to detect switch. An "error" on such instances will be improbable and if it does occur- it must be due to a switch.

But what is an "error"? - even if we know the optimal classifier $\boldsymbol{u}$ the probability for a classification error is $\frac{1-|\boldsymbol{u}^\top \boldsymbol{x}_t|}{2}$. So error can only be considered in terms of distance from the optimal classifier.

Problem - the optimal classifier is unknown. So how can we check if our prediction is close enough to it?

Using Selective Sampling to Overcome Switch

Solution - estimate optimal classifier $\boldsymbol{v}$ with a demo classifier h_t constructed from a window of recent instances.

- If no switch occurred - $\boldsymbol{x}_t$ and h_t should give close predictions, as both are close in prediction to $\boldsymbol{v}$.
- If a switch occurred:
  - If $\boldsymbol{x}_t$ and h_t do not yield close predictions - we detect the switch
  - If $\boldsymbol{x}_t$ and h_t t yield close predictions - switch is insignificant and not much additional regret will be suffered

Construction of Demo Classifier

- Set $L_t = L_0 + \sqrt{t}$
- At round t select a window of last L_t instances
- Calculate $A_{L_t} = I + \sum_{l=t-L}^{t-1} \boldsymbol{x}_l \boldsymbol{x}_l^\top, b_{L_t} = \sum_{l=t-L}^{t} y_l \boldsymbol{x}_l$
- Construct $h_t = \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} b_{L_t}$

To save querying labels we set resolution classifier h_t for a window of KL_t next instances. At round $KL_t + 1$ we construct a new demo classifier, and so forth.

Algorithm for Detecting Switch

- Set $\delta_t = \frac{\delta}{t(t+1)}$
- Calculate $C_t = \left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$
- Calculate:

$$K_t = \sqrt{2r_t \ln \frac{2}{\delta_t}} + \sqrt{2r_{L_t} \ln \frac{2}{\delta_t}} + r_t + \sqrt{r_t} + r_{L_t} + \sqrt{r_{L_t}}$$

- If $C_t > K_t$ declare switch and restart classifier w_t from zero
- Else continue to next round

Algorithm for Detecting Switch

- If $C_t > K_t$ switch is detected and we overcome its effect
- If no switch occurred we can assure that $C_t \leq K_t$ and no false detections will be made
- If $C_t \leq K_t$ but a switch did occur - can we assure that it will cause no significant additional regret?

First we will show that indeed if $C_t \leq K_t$ we can assure low regret.
Later we will prove that the probability for a false positive is small.

Regret Calculation

The instantaneous regret is controlled by the term $|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t|$:

$$R_t = \Pr\left[y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t < 0\right] - \Pr\left[y_t \boldsymbol{v}^\top \boldsymbol{x}_t < 0\right] \leq$$
$$\varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t| < \varepsilon\}} + \Pr\left[\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \geq \varepsilon\right] \quad (11)$$

We can bound it by triangle inequality:

$$\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \leq \left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right| + \left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$$
$$= C_t + \left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right| \quad (12)$$

Regret Calculation

We already have a bound for C_t, as a switch was not detected. What about $\left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$?

From the bias bound on the BBQ classifier and by Hoefding bound we shall have:

$$\left|\boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right| \leq \sqrt{2 r_{L_t} \ln \frac{2}{\delta_t}} + r_{L_t} + \sqrt{r_{L_t}} \qquad (13)$$

With probability $1 - \delta_t$.

Regret Calculation

Combining given bound on C_t and equation 13 we have:

$$\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \leq \sqrt{r_t}\left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right) + r_t$$
$$+2\sqrt{r_{L_t}}\left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right) + 2r_{L_t} \tag{14}$$

Equation 14 together with the identity $I_{\{x<1\}} \leq e^{1-x}$ will allow us to bound the regret.

Regret Calculation

$$\Pr\left[\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \geq \varepsilon\right] \leq 2I_{\left\{r_{L_t}\left(\sqrt{2\ln\frac{2}{\delta_t}}+1\right)^2 \geq \frac{\varepsilon^2}{36}\right\}} \tag{15}$$

$$+2I_{\left\{r_{L_t}\geq\frac{\varepsilon}{6}\right\}} + I_{\left\{r_t\left(\sqrt{2\ln\frac{2}{\delta_t}}+1\right)^2 \geq \frac{\varepsilon^2}{36}\right\}} + I_{\left\{r_t\geq\frac{\varepsilon}{6}\right\}}$$

$$\leq 2\exp\left\{1 - \frac{\varepsilon^2}{36 r_{L_t}\left(\sqrt{2\ln\frac{2}{\delta_t}}+1\right)^2}\right\} + 2\exp\left\{1 - \frac{\varepsilon}{6 r_{L_t}}\right\}$$

$$+\exp\left\{1 - \frac{\varepsilon^2}{36 r_t\left(\sqrt{2\ln\frac{2}{\delta_t}}+1\right)^2}\right\} + \exp\left\{1 - \frac{\varepsilon}{6 r_t}\right\}$$

Regret Calculation

The cumulative regret is given by:

$$R_T = \sum_{t=1}^{T} R_t \tag{16}$$

We will sum over the terms of equation 16 to bound the regret.

We divide the summation to rounds where $r_t \leq t^{-\kappa}$ and rounds where $r_t \leq t^{-\kappa}$.

Regret Calculation

We use the identities: $1 - x \leq -\ln x$ (for $x \leq 1$) and $\exp\{-x\} \leq \frac{1}{ex}$, and the fact that:

$$r_t \leq 1 - \frac{\det A_{t-1}}{\det A_t} \tag{17}$$

to calculate the following sum:

$$\sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{1 - \frac{\varepsilon^2}{36 r_t \left(\sqrt{2 \ln \frac{2}{\delta_t}} + 1\right)^2}\right\} \leq$$

Regret Calculation

$$
\begin{aligned}
&\leq \frac{36\left(\sqrt{2\ln\frac{2}{\delta_T}}+1\right)^2}{\varepsilon^2}\sum_{t=T_1,r_t>t^{-\kappa}}^{T} r_t \\
&\leq \frac{36\left(\sqrt{2\ln\frac{2}{\delta_T}}+1\right)^2}{\varepsilon^2}\sum_{t=T_1,r_t>t^{-\kappa}}^{T} \left(1-\frac{\det A_{t-1}}{\det A_t}\right) \\
&\leq -\frac{36\left(\sqrt{2\ln\frac{2}{\delta_T}}+1\right)^2}{\varepsilon^2}\sum_{t=T_1,r_t>t^{-\kappa}}^{T} \ln\left(\frac{\det A_{t-1}}{\det A_t}\right) \\
&\leq \frac{16}{\varepsilon^2}\left\{d\ln T-\ln\left(\det A_{T_1}\right)\right\} \qquad (18)
\end{aligned}
$$

Regret Calculation

To sum over the $r_t \leq t^{-\kappa}$ bounds we use the following result:

$$\int \exp\{az^r\}\,dz = -\frac{z(-az^r)^{-\frac{1}{r}}}{r}\Gamma\{\frac{1}{r}, -az^r\} \qquad (19)$$

This yields:

$$\sum_{t=T_1, r_t \leq t^{-\kappa}}^{T} \exp\left\{1 - \frac{\varepsilon^2}{36 r_t \left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)^2}\right\} =$$

Regret Calculation

$$= \sum_{t=T_1, r_t \le t^{-\kappa}}^{T} \exp\left\{1 - \frac{\varepsilon^2 t^\kappa}{36\left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)^2}\right\} \le$$

$$\le e \int_{T_1}^{T} \exp\left\{-\frac{\varepsilon^2 t^\kappa}{36\left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)^2}\right\} dt =$$

Regret Calculation

$$= \frac{e36^{\frac{1}{\kappa}} \left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)^{\frac{2}{\kappa}}}{\kappa\varepsilon^{\frac{2}{\kappa}}}$$

$$\left[\Gamma\left\{\frac{1}{\kappa}, \frac{\varepsilon^2 T_1^\kappa}{36\left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)^2}\right\} - \Gamma\left\{\frac{1}{\kappa}, \frac{\varepsilon^2 T^\kappa}{36\left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)^2}\right\}\right]$$

The development of the sum $\sum_{t=T_1}^{T} \exp\left\{1 - \frac{\varepsilon}{6r_t}\right\}$ is identical up to constants.

Regret Calculation

We are left with summing over the r_{L_t} terms.