Learning Drifting Data Using Selective Sampling

March 19, 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 $\mathrm{E}\left[y_t\right] = \boldsymbol{u}^\top \boldsymbol{x}_t$
- for $t > \tau$ holds $\mathrm{E}\left[y_t\right] = \boldsymbol{v}^\top \boldsymbol{x}_t$
- $\|\boldsymbol{x}_t\| = \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$

BBQ Algorithm

We submit prediction:

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

$\boldsymbol{w}_t$ 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 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:

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

Instead of

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

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 the bias 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 instance 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 $\frac{1-\left|\boldsymbol{u}^\top \boldsymbol{x}_t\right|}{2}$. 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 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
- Set $A_{L_t} = I + \sum_{l=t-L}^{t-1} \boldsymbol{x}_l \boldsymbol{x}_l^\top, b_{L_t} = \sum_{l=t-L}^{t} y_l \boldsymbol{x}_l$
- 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 $C_t = |\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t|$
- 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 regret is controlled by the term $|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t|$:

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

We can bound it by triangle inequality:

$$|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t| \leq |\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t| + |\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t|$$
$$= C_t + |\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t| \tag{12}$$

Regret Calculation

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

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

$$|\boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t| \leq \sqrt{2 r_{L_t} \ln \frac{2}{\delta_t}} + r_{L_t} + \sqrt{r_{L_t}} \tag{13}$$

With probability $1 - \delta_t$.

Regret Calculation

Combining given bound on C_t and equation 13 we have:

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

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

Regret Calculation

$$
\begin{aligned}
&\Pr\left[|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t| \geq \varepsilon\right] \leq 2I_{\left\{\left(2r_{L_t}\ln\frac{2}{\delta_t}\right)\geq\frac{\varepsilon^2}{81}\right\}} \\
&+2I_{\left\{r_{L_t}\geq\frac{\varepsilon}{9}\right\}} + 2I_{\left\{r_{L_t}\geq\frac{\varepsilon^2}{81}\right\}} \\
&+I_{\left\{\left(2r_t\ln\frac{2}{\delta_t}\right)\geq\frac{\varepsilon^2}{81}\right\}} + I_{\left\{r_t\geq\frac{\varepsilon}{9}\right\}} + I_{\left\{r_t\geq\frac{\varepsilon^2}{81}\right\}} \\
&\leq \exp\{1-\frac{\varepsilon^2}{81\left(2r_t\ln\frac{2}{\delta_t}\right)}\} + \exp\{1-\frac{\varepsilon}{9r_t}\} + \exp\{1-\frac{\varepsilon^2}{81r_t}\} \\
&+2\exp\{1-\frac{\varepsilon^2}{81\left(2r_{L_t}\ln\frac{2}{\delta_t}\right)}\} + 2\exp\{1-\frac{\varepsilon}{9r_{L_t}}\} + 2\exp\{1-\frac{\varepsilon^2}{81r_{L_t}}\}
\end{aligned}
\tag{15}
$$

Regret Calculation

$$
\begin{aligned}
&\Pr\left[|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t| \geq \varepsilon\right] \leq 2I_{\left\{\left(r_{L_t}\left(\sqrt{2\ln\frac{2}{\delta_t}}+1\right)\ln\frac{2}{\delta_t}\right)\geq\frac{\varepsilon^2}{36}\right\}} \\
&+2I_{\left\{r_{L_t}\geq\frac{\varepsilon}{9}\right\}} + 2I_{\left\{r_{L_t}\geq\frac{\varepsilon^2}{81}\right\}} \\
&+I_{\left\{\left(2r_t\ln\frac{2}{\delta_t}\right)\geq\frac{\varepsilon^2}{81}\right\}} + I_{\left\{r_t\geq\frac{\varepsilon}{9}\right\}} + I_{\left\{r_t\geq\frac{\varepsilon^2}{81}\right\}} \\
&\leq \exp\{1-\frac{\varepsilon^2}{81\left(2r_t\ln\frac{2}{\delta_t}\right)}\} + \exp\{1-\frac{\varepsilon}{9r_t}\} + \exp\{1-\frac{\varepsilon^2}{81r_t}\} \\
&+2\exp\{1-\frac{\varepsilon^2}{81\left(2r_{L_t}\ln\frac{2}{\delta_t}\right)}\} + 2\exp\{1-\frac{\varepsilon}{9r_{L_t}}\} + 2\exp\{1-\frac{\varepsilon^2}{81r_{L_t}}\}
\end{aligned}
\tag{16}
$$