Learning Drifting Data Using Selective Sampling

March 23, 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

Problem setup:

- $\boldsymbol{x}_t \in R^d$
- $y_t \in \{\pm 1\}$

Assumptions:

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

Problem Setting

- At each round t instance $\boldsymbol{x}_t$ is observed
- Prediction $\hat{y}_t$ is issued
- Regret R_t is suffered
- True label y_t can be queried

Linear classification is used to issue prediction:

$$\hat{y}_t = \mathrm{sign}\left\{\boldsymbol{w}_t^\top \boldsymbol{x}_t\right\} \quad (1)$$

RLS Estimator

$\boldsymbol{w}_t$ would be the RLS estimator (Chesa-Bianchi at all. 2004, 2006, 2009) solving the following problem:

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

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

RLS Estimator

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:

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

RLS Estimator

Equivalent formulation:

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

Where:

- $A_t = \left(I + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right) \in R^{d \times d}$
- $b_t = \sum_{i=1}^{n} y_i \boldsymbol{x}_i \in R^d$

A_t can be viewed as covariance or "confidence " matrix

BBQ Algorithm - Querying Labels

Cesa-Bianchi, Gentile, Orabona 2009:

Selective sampling algorithm -

- Set $\kappa \in (0, 1)$
- Calculate $r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t$
- If $r_t > t^{-\kappa}$ label y_t is queried
- If $r_t \leq t^{-\kappa}$ label y_t remains unknown

RLS Estimator and BBQ Algorithm Properties

Assuming standard, no switch setting, $(\boldsymbol{u} = \boldsymbol{v})$:

- Logarithmic cumulative regret R_T:

$$R_T \leq O\left(d \ln T\right) \tag{5}$$

- Reduced number of queried labels N_T:

$$N_T \sim O\left(dT^{\kappa} \ln T\right) \tag{6}$$

- Controlled estimator bias B_t:

$$B_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t - \mathrm{E}\left[\boldsymbol{w}_t^\top \boldsymbol{x}_t\right] \leq r_t + \sqrt{r_t} \tag{7}$$

Effect of Switch

- Switch from $\boldsymbol{u}$ to $\boldsymbol{v}$ at round τ increases bias bound:

$$B_t \leq r_t + \sqrt{r_t} + N_\tau \|\boldsymbol{v} - \boldsymbol{u}\| \sqrt{r_t} \tag{8}$$

- The bias B_t controls the regret R_T. So switch at round τ increases regret bound:

$$R_T \leq O\left(\left\{\|\boldsymbol{v} - \boldsymbol{u}\|^2 \tau^{2\kappa} \left(d \ln \tau\right)^2 + 1\right\} d \ln T\right) \tag{9}$$

Using Selective Sampling to Detect Switch

- When switch occurs low cumulative regret can no longer be expected
- Selective sampling approach measures estimator's confidence regarding prediction on give instance x_t
- When confidence on instance x_t is high prediction should be close to optimal
- Evaluating prediction on "high confidence" instances can be used to detect change

Using Selective Sampling to Detect Switch

Confidence factor $r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t$:

- Small r_t - high confidence regarding instance $\boldsymbol{x}_t$
- Large r_t - high uncertainty (low confidence) regarding instance $\boldsymbol{x}_t$

r_t controls both the bias B_t and the instantaneous regret R_t:

- If r_t is large, low regret R_t can not be assured, switch or no switch.
- If r_t is small, low regret R_T should be expected. Unless a switch had occurred...

Using Selective Sampling to Detect Switch

Main idea - evaluate performance on instances with small r_t to detect switch.

Bad performance will indicate that switch had occurred.

- Performance cannot be evaluated comparing prediction $\hat{y}_t$ to label y_t due to noise
- Even if optimal classifier $\boldsymbol{v}$ is known - error probability will be $\frac{1-\left|\boldsymbol{v}^\top \boldsymbol{x}_t\right|}{2}$
- Prediction will be evaluated comparing to optimal classifier $\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right|$

Windowed Demo Classifiers

- Problem - optimal classifier $\boldsymbol{v}$ is unknown.
- Solution - estimate optimal classifier $\boldsymbol{v}$ with demo classifier h_t.
- Demo classifier h_t constructed from a window of last L rounds and should estimate $\boldsymbol{v}$ well enough
- Performance of $\boldsymbol{w}_t$ comparing to h_t will evaluate $\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right|$ and indicate possible switch

Construction of Windowed Demo Classifier

- Parameter - initial window length $L_0 > 0$
- Calculate window length $L_t = L_0 + \sqrt{t}$
- At round t select a window of last L_t instances
- Calculate $A_{L_t} = \left(I + \sum_{l=t-L_t}^{t-1} \boldsymbol{x}_l \boldsymbol{x}_l^\top \right), b_{L_t} = \sum_{l=t-L_t}^{t-1} y_l \boldsymbol{x}_l$
- Construct demo classifier $h_t = \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right)^{-1} b_{L_t}$

Resolution of Windowed Demo Classifier

- Demo classifier h_t constructed at round t from a window of last L_t instances
- Demo classifier h_t will be used to evaluate next KL_t instances
- Next demo classifier h_{t_next} will be constructed at round $KL_t + 1$
- Only $\frac{T}{K}$ labels will be queried
- Switch detection resolution reduced from L_t to KL_t

Algorithm for Detecting Switch

- Calculate estimator $\boldsymbol{w}_t = A_t^{-1} b_t$
  - Where $A_t = \left(I + \sum_{i=1}^{N_t} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right), b_t = \sum_{i=1}^{N_t} y_i \boldsymbol{x}_i$
- Calculate demo classifier $h_t = \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right)^{-1} b_{L_t}$
  - $A_{L_t} = \left(I + \sum_{l=m_t-L_t}^{m_t} \boldsymbol{x}_l \boldsymbol{x}_l^\top \right), b_{L_t} = \sum_{l=m_t-L_t}^{m_t} y_l \boldsymbol{x}_l$
- Calculate $r_t = \boldsymbol{x}_t^\top A_t^{-1} \boldsymbol{x}_t$
- Calculate $r_{L_t} = \boldsymbol{x}_t^\top \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top \right)^{-1} \boldsymbol{x}_t$

Algorithm for Detecting Switch

- Parameter - $\delta \in (0,1)$
- Calculate $\delta_t = \frac{\delta}{t(t+1)}$
- Calculate $C_t = \left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$
- Calculate:

$$K_t = \left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)\sqrt{r_t} + r_t + \left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)\sqrt{r_{L_t}} + r_{L_t}$$

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

Algorithm's Main Result

- If $C_t > K_t$
  - If switch occurred it is detected
  - If no switch occurred $\Pr\left[C_t > K_t\right] \leq (1 - 2\delta)$ - to be proved
- If $C_t \leq K_t$
  - If no switch occurred, no change applied to standard setting
  - If a switch occurred and undetected as $C_t \leq K_t$, additional regret caused would be small - to be proved

Algorithm's Main Result

Main result:

1. If a switch occurs - algorithm detects it, or assures it causes small harm
2. No switch occurs - no false detection

Proving Main Result

Proof structure as follows:

1. Proving undetected switch will cause low regret:
   - Bounding instantaneous regret
   - Summing to cumulative regret
2. Proving probability for false positives is small

Proving Main Result - Instantaneous Regret Bound

Instantaneous regret R_t controlled by the term $|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t|$:

$$R_t = \Pr\left[y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t < 0\right] - \Pr\left[y_t \boldsymbol{v}^\top \boldsymbol{x}_t < 0\right] \leq$$
$$\varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t| < \varepsilon\}} + \Pr\left[\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \geq \varepsilon\right] \tag{10}$$

$|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t|$ can be bounded by triangle inequality:

$$\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \leq \left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right| + \left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$$
$$= C_t + \left|\boldsymbol{v}_t^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right| \tag{11}$$

Proving Main Result - Instantaneous Regret Bound

- C_t is bounded by K_t as a switch was not detected:

$$C_t \leq K_t = \left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)\sqrt{r_t} + r_t + \left(\sqrt{2\ln\frac{2}{\delta_t}} + 1\right)\sqrt{r_{L_t}} + r_{L_t} \tag{12}$$

- From properties of RLS estimator (Cesa-Bianchi at all) applied to demo classifier h_t:

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

With probability $1 - \delta_t$.

Proving Main Result - Instantaneous Regret Bound

Combining bounds on C_t and on $\left|\boldsymbol{v}^\top \boldsymbol{x}_t - h_t^\top \boldsymbol{x}_t\right|$:

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

Proving Main Result - Instantaneous Regret Bound

Using identities:

- $\Pr[A] = \mathrm{E}[I_A]$
- $I_{\{x<1\}} \le e^{1-x}$

final bound instantaneous regret R_t achieved:

$$
\begin{aligned}
R_t &\le \varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t|<\varepsilon\}} + \Pr\left[\left|\boldsymbol{w}_t^\top \boldsymbol{x}_t - \boldsymbol{v}^\top \boldsymbol{x}_t\right| \ge \varepsilon\right] \qquad (15) \\
&\le \varepsilon I_{\{|\boldsymbol{v}^\top \boldsymbol{x}_t|<\varepsilon\}} + 2\exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_{L_t}}\right\} + 2\exp\left\{1 - \frac{\beta_\varepsilon}{r_{L_t}}\right\} \\
&+ \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_t}\right\} + \exp\left\{1 - \frac{\beta_\varepsilon}{r_t}\right\} + \delta_t
\end{aligned}
$$

Proving Main Result - Cumulative Regret Bound

Cumulative regret R_T is given by:

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

Cumulative regret R_T will be bounded by summing over bound of instantaneous regret R_t.

Calculation outline:

1. Summation over the r_t terms - separate calculation for:
   - Rounds t for which with $r_t \le t^{-\kappa}$
   - Rounds t for which with $r_t > t^{-\kappa}$
2. Summation over the r_{L_t} terms.
3. Deriving final bound

Proving Main Result - Cumulative Regret Bound

Summation over r_t terms - for rounds with $r_t > t^{-\kappa}$ -

- Identity $\exp\{-x\} \leq \frac{1}{ex}$ gives:

$$\sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_t}\right\} \leq \frac{1}{\alpha_{\varepsilon,T}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} r_t$$

- The result $r_t \leq \left(1 - \frac{\det A_{t-1}}{\det A_t}\right)$ (Cesa-Bianchi at. all 2004) yields:

$$\frac{1}{\alpha_{\varepsilon,t}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} r_t \leq \frac{1}{\alpha_{\varepsilon,t}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \left(1 - \frac{\det A_{t-1}}{\det A_t}\right)$$

Proving Main Result - Cumulative Regret Bound

Summation over r_t terms - for rounds with $r_t > t^{-\kappa}$ -

- Identity $1 - x \leq -\ln x$ (for $x \leq 1$) gives:

$$\frac{1}{\alpha_{\varepsilon,t}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \left(1 - \frac{\det A_{t-1}}{\det A_t}\right) \leq -\frac{1}{\alpha_{\varepsilon,t}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \ln\left(\frac{\det A_{t-1}}{\det A_t}\right)$$

- Computing the sum will give final expression:

$$-\frac{1}{\alpha_{\varepsilon,t}} \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \ln\left(\frac{\det A_{t-1}}{\det A_t}\right) \leq \frac{1}{\alpha_{\varepsilon,t}} \left\{ d \ln T - \ln\left(\det A_{T_1}\right) \right\}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_t terms - for rounds with $r_t \leq t^{-\kappa}$ -

- Substituting $r_t \leq t^{-\kappa}$ and replacing sum with integral yields:

$$\sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_t}\right\} \leq e \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{-\frac{\alpha_{\varepsilon,t}}{t^{-\kappa}}\right\}$$

$$\leq e \int_{T_1}^{T} \exp\left\{-\alpha_{\varepsilon,T} t^{\kappa}\right\} dt =$$

$$= \frac{e}{\kappa \left(\alpha_{\varepsilon,T}\right)^{\frac{1}{\kappa}}} \left(\Gamma\left\{\frac{1}{\kappa}, \alpha_{\varepsilon,T} T_1^{\kappa}\right\} - \Gamma\left\{\frac{1}{\kappa}, \alpha_{\varepsilon,T} T^{\kappa}\right\}\right)$$

- Last equality follows from the identity:

$$\int \exp\{a z^s\}\, dz = -\frac{z(-a z^s)^{-\frac{1}{s}}}{s} \Gamma\left\{\frac{1}{s}, -a z^s\right\} \qquad (17)$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

Matrix Chernoff bound - for a series of random, i.i.d PSD matrices $Z_k \in R^{d \times d}$ holds:

$$\Pr\left[\lambda_{\min}\left\{\sum_k Z_k\right\} \leq (1-\gamma)\,\mu_{\min}\right] \leq d\left(\frac{e^{-\gamma}}{(1-\gamma)^{(1-\gamma)}}\right)^{\frac{\mu_{\min}}{\rho}} \tag{18}$$

where:

- $\gamma \in (0,1)$
- $\mu_{\min} = \lambda_{\min}\left\{\sum_k \mathrm{E}\left[Z_k\right]\right\}$
- $\lambda_{\max}\left\{\mathrm{E}\left[Z_k\right]\right\} \leq \rho$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

- Assumption: smallest eigenvalue of covariance matrix grows linearly - $\lambda_{\min}\left\{\sum_{k=1}^{L} \mathrm{E}\left[\boldsymbol{x}_k \boldsymbol{x}_k^\top\right]\right\} \sim O\left(\frac{L}{d}\right)$
- Using Chernoff matrix bound on $Z_k = \boldsymbol{x}_k \boldsymbol{x}_k^\top$, under the above assumption, yields:

$$\lambda_{\min}\left\{A_{L_t}\right\} = \lambda_{\min}\left\{I + \sum_{k=1}^{L_t} \boldsymbol{x}_k \boldsymbol{x}_k^\top\right\} > (1-\gamma)\frac{L_t}{d} + 1 \quad (19)$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

Using the bound and identities below:

- For unit normed $\boldsymbol{x}$: $\boldsymbol{x}^\top M \boldsymbol{x} \leq \lambda_{max}\{M\}$
- $\lambda_{max}\{M\} = \frac{1}{\lambda_{min}\{M^{-1}\}}$
- $\lambda_{\min}\{A_{L_t}\} > (1-\gamma)\frac{L_t}{d}$

we get:

$$r_{L_t} = \boldsymbol{x}_t^\top \left(A_{L_t} + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1} \boldsymbol{x}_t \leq \frac{d}{(L_t+2)(1-\gamma)} \tag{20}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

- Replacing $L_t = L_0 + \sqrt{t}$ into the bound would yield:

$$r_{L_t} \leq \frac{d}{\left(L_0 + \sqrt{t} + 2\right)(1-\gamma)} \tag{21}$$

- Substituting bound r_{L_t} into the sum over regret R_T bound:

$$\sum_{t=T_1}^{T} \exp\left\{1 - \frac{\alpha_{\varepsilon,t}}{r_{L_t}}\right\} \leq e \sum_{t=T_1, r_t > t^{-\kappa}}^{T} \exp\left\{-\hat{\alpha}_{\varepsilon,t}\left(L_0 + \sqrt{t}\right)\right\}$$

Proving Main Result - Cumulative Regret Bound

Summation over r_{L_t} terms -

Now replacing sum with integral and solving as before yields:

$$e\sum_{t=T_1}^{T}\exp\left\{-\hat{\alpha}_{\varepsilon,t}\left(L_0+\sqrt{t}\right)\right\}\leq$$
$$\frac{2e}{\left(\tilde{\alpha}_{\varepsilon,T}\right)^2}\left(\Gamma\left\{2,\tilde{\alpha}_{\varepsilon,T}\left(T_1-L_0\right)^{\frac{1}{2}}\right\}-\Gamma\left\{2,\tilde{\alpha}_{\varepsilon,T}\left(T-L_0\right)^{\frac{1}{2}}\right\}\right)$$

Proving Main Result - Cumulative Regret Bound

- Summing all developed bounds yields:

$$R_T \leq O\left(d\left\{\ln T\right\}^2\right) \tag{22}$$