PAC-Bayesian Online Clustering

Le li, Benjamin Guedj, Sbastien Loustau

Mathis Linger, Selim Dekali

ENSAE

Table of contents

Notation

Notation

- $(x_t)_{1:T}$: online dataset, where $x_t \in \mathbb{R}^d$
- K_t: nb of clusters
- $\hat{c}_t = (\hat{c}_{t,1}, \hat{c}_{t,2}, ...\hat{c}_{t,K_t})$: clusters location, depending on past information $(x_s)_{1:(t-1)}$ and $(\hat{c}_s)_{1:(t-1)}$
- When x_t is newly revealed, the instantaneous loss:
  $$\ell(\hat{c}_t, x_t) = \underset{1 \leq k \leq K_t}{min} |\hat{c}_{t,k} - x_t|_2^2$$

Notation

- $\mathscr{C} = \bigcup_{k=1}^{p}$
- q: discrete probability distribution on the set [1,p]:=1,..p
  for any k $\in$ [1,p], let π_k, the probability distribution on $\mathbb{R}^{dk}$
- For any c $\in \mathscr{C}$, we define $\pi(c)$, as
  $$\pi(c) = \sum_{k\in[1,p]} q(k)\mathbb{1}_{\{c\in\mathbb{R}^{dk}\}}\pi_k(c)$$

Notation

- $c \in \mathscr{C}$ a partition of $\mathbb{R}^d, \pi \in \mathbb{P}(\mathscr{C})$ a quasi prior over this set
- $\lambda > 0$: inverse temperature parameter
- At each time t, we observe x_t and a random partition $\hat{c}_{t+1} \in \mathscr{C}$ is sampled from the quasi-posterior:
  $d\hat{\rho}_{t+1} \propto \exp\{-\lambda S_t(c)\} d\pi(c)$
- Cumulative loss:
  $S_t(c) = S_{t-1}(c) + \ell(c, x_t) + \frac{\lambda}{2}\{\ell(c, x_t) - \ell(\hat{c}_t, x_t)\}^2$

Sparcity Regret Bounds

Sparcity Regret Bounds

Algo 1: The PAC-Bayesian online Clustering algorithm

1: Input parameters: $p > 0, \pi \in \mathscr{P}(\mathscr{C}), \lambda > 0$ and $S_0 = 0$

2: Initialization: Draw $\hat{c}_1 \sim \pi$

3: For $t \in [1, T-1]$:

4: Get the data x_t

5: Draw $\hat{c}_{t+1} \sim \hat{\rho}_{t+1}(c)$ where $d\hat{\rho}_{t+1} \propto \exp\{-\lambda S_t(c)\} d\pi(c)$, and

$$S_t(c) = S_{t-1}(c) + \ell(c, x_t) + \frac{\lambda}{2}\{\ell(c, x_t) - \ell(\hat{c}_t, x_t)\}^2$$

6: End for

Sparcity Regret Bounds

Theorem 1:

For any $(x_t)_{1:T} \in \mathscr{R}^{dT}$, any quasi prior $\pi \in \mathscr{P}(\mathscr{C})$, any $\lambda > 0$,

the procedure described in Algo 1 satisfies:

$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots\hat{\rho}_t)} \ell(\hat{c}_t, x_t) \leq$
$\underset{\rho \in \mathscr{P}_\pi(\mathscr{C})}{inf} \Big\{ \mathbb{E}_{c\sim\rho}[\sum_{t=1}^{T} \ell(c, x_t)] + \frac{\mathcal{K}(\rho,\pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots\hat{\rho}_T)} \mathbb{E}_{c\sim\rho} \sum_{t=1}^{T} [\ell(c, x_t) - \ell(\hat{c}$

Sparcity Regret Bounds

The regret bound could be refine when :

- $q(k) = \frac{\exp -\eta k}{\sum_{i=1}^{p} \exp -\eta i}$, with $\eta > 0$.
- $d\pi_k(c, R) = \left(\frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}} \right) \frac{1}{(2R)^{dk}} \left\{ \prod_{j=1}^{k} \mathbb{1}_{\{\mathbb{B}_d(2R)\}}(c_j) \right\} dc$

Corollary 1:

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots \hat{\rho}_t)} \ell(\hat{c}_t, x_t) \leq \underset{k \in [1,p]}{inf} \left\{ \underset{c \in \mathscr{C}(k,R)}{inf} \sum_{t=1}^{T} \ell(c, x_t) + \frac{dk}{2\lambda} \log \frac{8R^2 \lambda T}{d=2} + \frac{\eta}{\lambda} k \right\} + \left(\frac{\log p}{\lambda} + \frac{d}{2\lambda} + \frac{\lambda T C_1^2}{2} \right)$$

where $C_1 = (2R + max_{t=1..T} |x_t|_2)^2$

Sparcity Regret Bounds

The below calibration yields a sublinear remainder term: :

- $\lambda = \frac{d+2}{2\sqrt{T}R^2}$

Corollary 2:

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots\hat{\rho}_t)} \ell(\hat{c}_t, x_t) \leq \inf_{k\in[1,p]} \left\{ \inf_{c\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(c, x_t) + k\frac{dR^2}{d+2}\sqrt{T}\log 4\sqrt{T} + k\frac{2R^2\eta}{d+2}\sqrt{T} \right\} + \left(\frac{2R^2 \log p}{d+2} + \frac{dR^2}{d+2} + \frac{(d+2)C_1^2}{4R^2} \right)\sqrt{T}$$

Hence, if there exist k^*, and $c^* \in \mathscr{C}(k^*, R)$, which achieve the infimum:

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots\hat{\rho}_t)} \ell(\hat{c}_t, x_t) - \sum_{t=1}^{T} \ell(c^*, x_t) \leq Jk^*\sqrt{T}\log T$$

J: constant depending on d,R, $\log p$ and C_1^2

Then the regret of the expected cumulative loss is sublinear in T.

Adaptative Sparcity Regret Bounds

Adaptative Sparcity Regret Bounds

T is usually unknown, prompting us to choose $\lambda = \lambda_t$

Algo 1: The adaptative PAC-Bayesian online Clustering algorithm

1: Input parameters: $p > 0, \pi \in \mathscr{P}(\mathscr{C}), (\lambda_t)_{0:T} > 0$ and $S_0 = 0$

2: Initialization: Draw $\hat{c}_1 \sim \pi$

3: For $t \in [1, T-1]$:

4: Get the data x_t

5: Draw $\hat{c}_{t+1} \sim \hat{\rho}_{t+1}(c)$ where $d\hat{\rho}_{t+1} \propto \exp\{-\lambda_t S_t(c)\} d\pi(c)$, and

$$S_t(c) = S_{t-1}(c) + \ell(c, x_t) + \frac{\lambda_{t-1}}{2}\{\ell(c, x_t) - \ell(\hat{c}_t, x_t)\}^2$$

6: End for

Adaptative Sparcity Regret Bounds

Theorem 2:

For any $(x_t)_{1:T} \in \mathscr{R}^{dT}$, any quasi prior $\pi \in \mathscr{P}(\mathscr{C})$,
if $(\lambda_t)_{0:T}$ a non-increasing sequence of positive numbers,

the procedure described in Algo 2 satisfies:

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots \hat{\rho}_t)} \ell(\hat{c}_t, x_t) \leq$$
$$\underset{\rho \in \mathscr{P}_\pi(\mathscr{C})}{inf} \Big\{ \mathbb{E}_{c \sim \rho} [\textstyle\sum_{t=1}^{T} \ell(c, x_t)] + \frac{\mathcal{K}(\rho, \pi)}{\lambda_T} + \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots \hat{\rho}_T)} \mathbb{E}_{c \sim \rho} \Big[\textstyle\sum_{t=1}^{T} \frac{\lambda_{t-1}}{2} [\ell(c, x_t) -$$

Adaptative Sparcity Regret Bounds

keeping previous setting for q and π_k, with $\eta \geq 0$ and $R \geq \underset{t=1..T}{max} |x_t|_2$
The below adaptative calibration for any $t \in [1, T] and \lambda_0 = 1$: :

- $\lambda_t = \frac{d+2}{2\sqrt{t}R^2}$

Corollary 2:

Then the algorithm 2 satisfies:
$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots \hat{\rho}_t)} \ell(\hat{c}_t, x_t) \leq$
$\underset{k \in [1,p]}{inf} \left\{ \underset{c \in \mathscr{C}(k,R)}{inf} \sum_{t=1}^{T} \ell(c, x_t) + \frac{dkR^2}{d+2} \sqrt{T} \log 4\sqrt{T} + k \frac{2R^2 \eta}{d+2} \sqrt{T} \right\} +$
$\left(\frac{2R^2 \log p}{d+2} + \frac{dR^2}{d+2} + \frac{(d+2)C_1^2}{2R^2} \right) \sqrt{T}$

The adaptative Algorithm 2 is supported by a sparcity regret bound with rate $\sqrt{T} \log T$.

The PACO algorithm

The PACO algorithm

- Since direct sampling from the quasi-posterior $\hat{\rho}_t$ is isually not possible, we will focus on a stochastic approximation, called PACO.
- Approximate $\hat{\rho}_t$ through MCMC, favoring local move.
- States of interest of the MC $(k^{(n)}, c^{(n)})_{0 \leq n \leq N}$,
  where $k^{(n)} \in [1, p]$ and $c^{(n)} \in \mathbb{R}^{dk^{(n)}}$
- At each iteration, from $(k^{(n)}, c^{(n)})$ to proposal state (k', c')
  Hence $c^{(n)} \in \mathbb{R}^{dk^{(n)}}$, and $c' \in \mathbb{R}^{dk'}$ may be of different dimensions $(k' \neq k^{(n)})$
  We create auxilary vectors ν_1, ν_2 to compensate for dimensional difference (d_1, d_2) s.t. $dk^{(n)} + d_1 = dk' + d_2$

The PACO algorithm

- Let $\rho_{k'}(., c_{k'}, \tau_{k'})$ denote the multivariate Student distribution on $\mathbb{R}^{dk'}$:
  $\rho_{k'}(c, c_{k'}, \tau_{k'}) = \prod_{j=1}^{k'} \left\{ C_{\tau_{k'}}^{-1} \left(1 + \frac{|c_j - c_{k',j}|_2^2}{6\tau_{k'}^2}\right)^{-\frac{3+d}{2}} \right\} dc$,
  where $C_{\tau_{k'}}^{-1}$ is the normalizing constraint
- First a local move from $k^{(n)}$ to k' is proposed by choosing $k' \in [k^{(n)} - 1, k^n + 1]$ with probability $q(k^{(n)}, .)$
- Next, choosing $d_1 = dk'$, $d_2 = dk^{(n)}$, we sample ν_1 from $\rho_{k'}$
- Finally, the pair (ν_2, c') is obtained by $(\nu_2, c') = g(\nu_1, c^{(n)})$, where $g : (x, y) \in \mathbb{R}^{dk'} x \mathbb{R}^{dk^{(n)}} \to (y, x) \in \mathbb{R}^{dk^{(n)}} x \mathbb{R}^{dk'}$

The PACO algorithm

Algo 3: PACO

1: Initialization: (λ_t)

2: For $t \in [1, T-1]$:

3: Initialization: $(k^{(0)}, c^{(0)}) \in [1, p] x \mathbb{R}^{dk^{(0)}}$

4: For $n \in [1, N-1]$:

5: Sample $k' \in [k^n - 1, kn + 1]$ from $q(k^{(n)}, .) = 1/3$

6: Let $c' \leftarrow$ standard k-means output.

7: Let $\tau' = 1/\sqrt{pt}$.

8: Sample $\nu_1 \sim \rho_{k'}(., c_{k'}, \tau_{k'})$.

9: Let $(\nu_2, c') = g((\nu_1, c^{(n)})$.

10: Accept the move $(k^{(n)}, c^{(n)}) = (k', c')$ with probability

$$\alpha\left[(k^{(n)}, c^{(n)}), (k', c')\right] = min\left\{1, \frac{\hat{\rho}_t(c')q(k', k^{(n)})\rho_{k^{(n)}}(c^{(n)}, c_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(c^{(n)})q(k^{(n)}, k')\rho'(c', c_{k'}, \tau_{k'})}\right\}$$

11: Else $(k^{n+1}, c^{n+1}) = (k^n, c^n)$

12: End for

13: Let $\hat{c}_t = c^{(N)}$.

14: End for

Numercial studies