Clustering applied to Hidden Markov Models

Ibrahim KADDOURI

Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay

Abstract

We consider the clustering of observations of a **Hidden Markov Model (HMM)** comprised of **two hidden states** and **discrete emissions**. Clustering amounts to the reconstruction of the hidden states using the observations by minimizing (in average) the number of misclassified observations. Both **online** and **offline** clustering are studied.

- We show that the **empirical plug-in Bayes classifier** using consistent estimators of the model parameters is **efficient** in the sense that its risk is equivalent to the Bayes risk for large samples.
- We identify the **asymptotic Bayes risk** using some forgetting properties of Markov Chains.
- We exhibit **upper and lower bounds** on the asymptotic Bayes risk using the model parameters.

Motivation

- Clustering is usually applied to heterogeneous data coming from different populations. These are usually modeled by a **mixture model**. However, without any additional assumption, these models **are not identifiable**. Inference on parameters and clustering become impossible.
- Assuming in addition that the data is derived from a HMM the model becomes identifiable and inference and clustering algorithms can be used as in [2].

Mathematical setting

- The hidden states $(X_k)_{k\in\mathbb{N}}$ are assumed to form a **Markov chain**.
- The observations $(Y_k)_{k\in\mathbb{N}}$ are **independent conditionally on the hidden states** and $Y_k \mid X_k = j \sim f_j$.
- The Markov chain $(X_k)_{k\in\mathbb{N}}$ will be assumed to have two hidden states, initial distribution ν and transition matrix:

$$Q = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

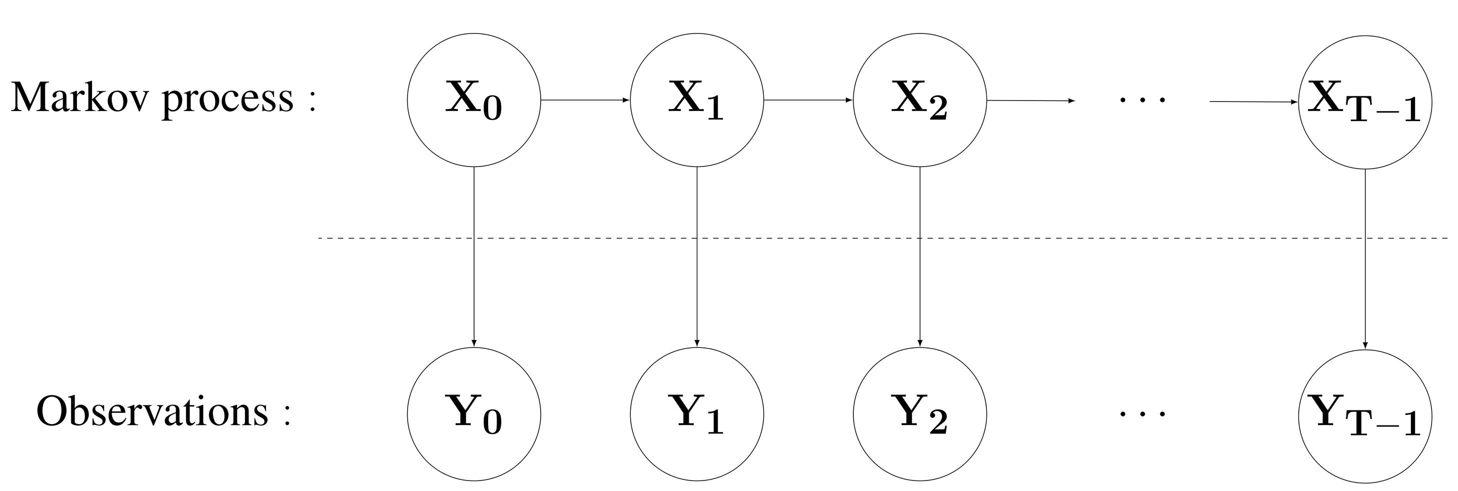

Figure 1: A hidden Markov model.

Offline vs online frameworks

Denote $\theta = (\nu, Q, f_0, f_1)$ and consider the loss function $L(x_{1:n}, x'_{1:n}) = \frac{1}{n}\sum_{k=1}^{n} 1_{x_k \neq x'_k}$. We study the risk of clustering observations in two frameworks:

- **Offline**: All observations are used in the classification procedures. Classifiers are of the form: $h(Y_{1:n}) = (h_k(Y_{1:n}))_{1\leq k\leq n}$.
  - **Offline risk**: $\mathcal{R}^{\text{Offline}}_{n,\text{HMM}}(h) = \mathbb{E}_\theta\left[\frac{1}{n}\sum_{k=1}^{n} 1_{X_k\neq h_k(Y_{1:n})}\right]$
  - **Offline Bayes risk**: $\mathcal{R}^{\star,\text{Offline}}_{n,\text{HMM}} = \mathbb{E}_\theta[\frac{1}{n}\sum_{k=1}^{n}\min_{x=0,1}\mathbb{P}_\theta(X_k\neq x \mid Y_{1:n})]$
- **Online**: Classification can use only past observations. Classifiers are of the from: $h(Y_{0:n-1}) = (h_k(Y_{0:k-1}))_{1\leq k\leq n}$
  - **Online risk**: $\mathcal{R}^{\text{Online}}_{n,\text{HMM}}(h) = \mathbb{E}_\theta\left[\frac{1}{n}\sum_{k=1}^{n} 1_{X_k\neq (h_k(Y_{0:k-1}))}\right]$
  - **Online Bayes risk**: $\mathcal{R}^{\star,\text{Online}}_{n,\text{HMM}} = \mathbb{E}_\theta[\frac{1}{n}\sum_{k=1}^{n}\min_{x=0,1}\mathbb{P}_\theta(X_k\neq x \mid Y_{0:k-1})]$

Efficiency of plug-in empirical Bayes classifier

Theorem 1 *Assume:*

- *The initial distribution ν is the **stationary distribution**.*
- $\delta = \min_{i,j} Q_{i,j} > 0$
- $(f_0(y), f_1(y)) = (p_0^y(1-p_0)^{1-y}, p_1^y(1-p_1)^{1-y})$
- $c^\star = \min(p_0, p_1, 1-p_0, 1-p_1) > 0$.

Then:

$$\mathcal{R}^{\text{Offline}}_{n,\text{HMM}}(\hat{h}) - \mathcal{R}^{\star,\text{Offline}}_{n,\text{HMM}} \leq \frac{4(1-\delta)}{\delta^2}\mathbb{E}_\theta\left[\frac{1}{n(1-\rho)}\|\nu-\hat{\nu}\|_2 + \Big(1/(1-\rho) + 1/(1-\hat{\rho})\Big)\left(\|Q-\hat{Q}\|_F + \frac{2}{c^\star}\max_{x=0,1}|p_x-\hat{p}_x|\right)\right]$$

$$\mathcal{R}^{\text{Online}}_{n,\text{HMM}}(\hat{h}) - \mathcal{R}^{\star,\text{Online}}_{n,\text{HMM}} \leq \frac{8(1-\delta)}{\delta^2}\mathbb{E}_\theta\left[\frac{1}{n(1-\rho)}\|\nu-\hat{\nu}\|_2 + \Big(1/(1-\rho) + 1/(1-\hat{\rho})\Big)\left(\|Q-\hat{Q}\|_F + \frac{1}{c^\star}\max_{x=0,1}|p_x-\hat{p}_x|\right)\right] + 2\mathbb{E}_\theta\left[\|Q-\hat{Q}\|_F\right]$$

where $\rho = \frac{1-2\delta}{1-\delta}$, $\hat{\rho} = \frac{1-2\hat{\delta}}{1-\hat{\delta}}$, $\hat{\delta} = \min_{i,j}\hat{Q}_{i,j} > 0$ and $\hat{h}$ is the empirical Bayes classifier using estimates of the model parameters $\hat{\theta} = (\hat{\nu}, \hat{Q}, \hat{p}_0, \hat{p}_1)$.

- Note that there is no need for consistent estimators of the initial distribution.
- Existence of consistent estimators of (Q, f_0, f_1) is ensured (cf. [1]).

$\rightarrow$ Plugging-in consistent estimators of the model parameters is thus an **efficient procedure**.

Define the following two quantities:

$$\mathcal{R}^{\star,\text{Offline}}_{\infty,\text{HMM}} = \mathbb{E}_\theta[\min(\mathbb{P}_\theta(X_0 = 1 \mid Y_{-\infty:+\infty}), \mathbb{P}_\theta(X_0 = 0 \mid Y_{-\infty:+\infty}))]$$
$$\mathcal{R}^{\star,\text{Online}}_{\infty,\text{HMM}} = \mathbb{E}_\theta[\min(\mathbb{P}_\theta(X_1 = 1 \mid Y_{-\infty:0}), \mathbb{P}_\theta(X_1 = 0 \mid Y_{-\infty:0}))]$$

Theorem 2 *Under the same assumptions:*

$$\left|\mathcal{R}^{\star,\text{Offline}}_{n,\text{HMM}} - \mathcal{R}^{\star,\text{Offline}}_{\infty,\text{HMM}}\right| \leq \frac{2}{n(1-\rho_0)}$$
$$\left|\mathcal{R}^{\star,\text{Online}}_{n,\text{HMM}} - \mathcal{R}^{\star,\text{Online}}_{\infty,\text{HMM}}\right| \leq \frac{\rho_1}{2n}\frac{1}{1-\rho_0}$$

where $\rho_0 = \frac{1-2\delta}{1-\delta}$, $\rho_1 = 1-2\delta$ and $\delta = \min_{i,j} Q_{i,j} > 0$.

Bounds on the asymptotic Bayes risk

We introduce the following parametrization :

$$\phi(\theta) = \left(\frac{q-p}{q+p},\quad 1-p-q,\quad \|f_0-f_1\|_\infty\right)$$

Theorem 3 *- The asymptotic Bayes risk for online clustering verifies:*

$$\mathcal{R}^{\star,\text{Online}}_{\infty,\text{HMM}} \leq \frac{1}{2} - \frac{|\phi_1|}{2}$$

- Assuming in addition that $\min(f_0, f_1) \geq c$ and that $|\phi_2| \leq \frac{c}{12}\wedge\frac{c^2}{4}$, then one has:

$$\mathcal{R}^{\star,\text{Online}}_{\infty,\text{HMM}} \geq \frac{1}{2} - \frac{|\phi_1|}{2} - \frac{(1-\phi_1^2)\phi_2^2\phi_3}{2c}$$

Note that the upper bound $\frac{1}{2} - \frac{|\phi_1|}{2}$ corresponds to the Bayes risk reached by the **majority class classifier**.

Conclusion and future directions

This work clarifies some features of the asymptotic behavior of the Bayes risk and empirical estimation procedures. However, some aspects still need to be studied:

- Extension of the results to the **nonparametric setting**.
- Bounds on the asymptotic Bayes risk in the **offline framework**.
- **Matching the upper and lower bounds** in order to understand how much better is the HMM Bayes classifier compared to majority class classifier.

References

[1] Kweku Abraham, Ismaël Castillo, and Elisabeth Gassiat. Multiple testing in nonparametric hidden markov models: An empirical bayes approach. *Journal of Machine Learning Research*, 2021.

[2] Olivier Cappé, Eric Moulines, and Rydén Tobias. *Inference in Hidden Markov Models*. Springer, 2005.

Scan this for more details $\rightarrow$

Acknowledgments: This work was done under the supervision of professors Elisabeth Gassiat and Zacharie Naulet. I would like to express my warm thanks and gratitude for proposing this subject to me, for all the fruitful discussions we had and for their availability all the time.