Online Clustering

Marco Jiralerspong Andjela Mladenovic (Group 2)
University of Montreal

Introduction

We examine the problem of online clustering within a regret framework and expand on the work of [1].

- We attempt to **replicate** the results of [1] and find our implementations yields **relevant differences**.
- We expand the experiments to include a **Follow the Leader (FTL)** implementation of a **Gaussian Mixture Model (GMM)**.
- We examine the performance of FTL GMM on **natural datasets** and the **counter example**.

Follow The Leader

Want to minimize regret with respect to best fixed action:

$$R(\omega_{1:T}) = \sum_{t=1}^{T} f_t(\omega_t) - \min_{u \in S} \sum_{t=1}^{T} f_t(u) \tag{1}$$

Follow the Leader:

$$\omega_t = \arg\min_{\omega \in S} \sum_{i=1}^{t-1} f_i(\omega) \tag{2}$$

Online k-Means and Online GMM

Online k-means

- The online algorithm at time t maintains a set of k candidate cluster centres, $C_t = \{c_{t,1}, \ldots, c_{t,k}\}$ before observing the datum x_t that arrives at time t with the goal of minimizing regret:

$$\text{regret}_T = \sum_{t=1}^{T} l_t(C_t, x_t) - \min_{C:|C|=k} \sum_{t=1}^{T} \min_{c \in C} ||x_t - c||_2^2 \tag{3}$$

where the loss incurred by the algorithm at time step t is:

$$l_t(C_t, x_t) = \min_{c \in C_t} ||x_t - c||_2^2 \tag{4}$$

Online GMM

- The online algorithm at time t maintains a set of k Gaussians with parameters, $\theta_t = \{(\pi_{t,1}, \mu_{t,1}, \Sigma_{t,1}), \ldots, (\pi_{t,k}, \mu_{t,k}, \Sigma_{t,k})\}$ before observing the datum x_t that arrives at time t with the goal of minimizing regret:

$$\text{regret}_T = \sum_{t=1}^{T} -\log p(x_t|\theta_t) - \min_{\theta} \left(\sum_{t=1}^{T} -\log p(x_t|\theta) \right) \tag{5}$$

where the loss incurred by the algorithm at time step t is:

$$l_t(\theta_t, x_t) = -\log p(x_t|\theta_t) = -\log\{\sum_{k=1}^{K} \pi_{t,k} \mathcal{N}(x_t|\mu_{t,k}, \Sigma_{t,k})\} \tag{6}$$

Natural Dataset Experiments

Replicated experiments for FTL k-means

While implementation details are sparse in [1], **we cannot replicate logarithmic regret on MNIST with FTL k-means**. For other datasets, our results are similar, up to a constant factor.

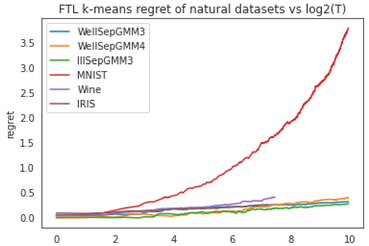

Figure 1: Our implementation

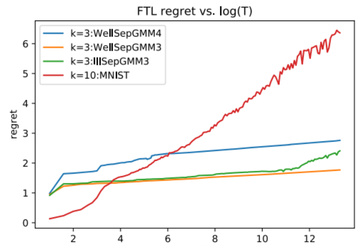

Figure 2: Implementation from [1]

Replicated experiments using FTL GMM

We extend the work in [1] to **a FTL implementation of a simplified GMM** (either spherical or diagonal covariances). We use (5) to compute the regret of our implementation instead of the L2 distance-based regret used in k-means.

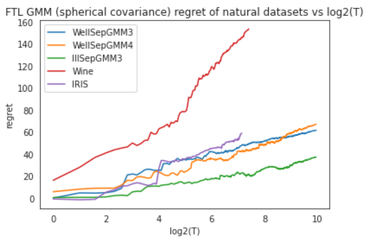

Figure 3: Spherical covariance

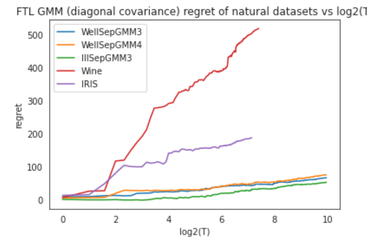

Figure 4: Diagonal covariance

FTL GMM on MNIST

The GMM with a diagonal covariance approaches **logarithmic regret on MNIST.**

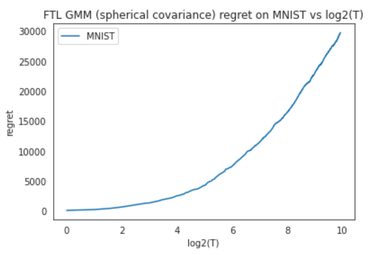

Figure 5: Spherical covariance

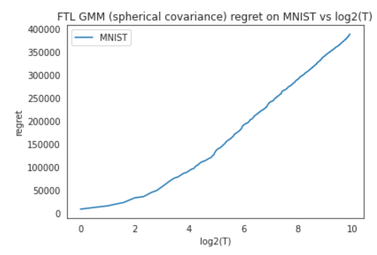

Figure 6: Diagonal covariance

Linear Regret - FTL GMM and K-means

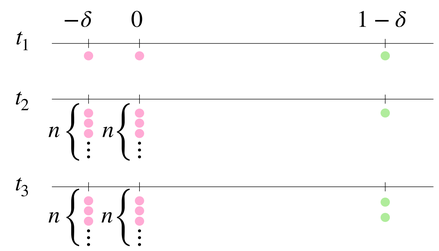

Figure 7: Constructed counter example

Experiments for FTL k-means and FTL GMM - linear regret

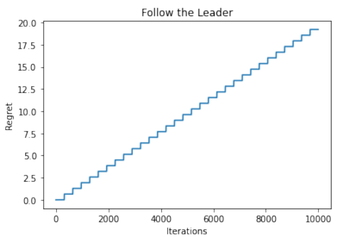

Figure 8: FTL k-means - Linear Regret

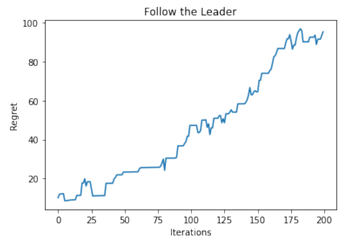

Figure 9: FTL GMM - Linear Regret

Future Work

- Construct and prove theoretically that the counter example yields linear regret for FTL GMM.
- Explore Follow the Regularized Leader and see if better regret is achievable.
- Propose a novel online algorithm that takes into account the nature of GMM and k-means without just relying on FTL.

Acknowledgements

This work was done under the supervision of Gauthier Gidel.
Scikit-learn was used for implementations on natural datasets.

References

[1] Vincent Cohen-Addad et al. “Online k-means Clustering”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1126–1134.