Overview

Goal: perform approximate posterior inference for finite and infinite Gaussian mixture models under time and memory constraints.

Method: variational inference with coresets.

Contributions: a novel coreset construction algorithm for posterior inference for BGMM and DPGMM.

Why Coresets?

Coresets are weighted subsets of the data, with strong performance guarantees for a specific problem. For a dataset $X$, a set of queries $Q = \{q_1, q_2, \ldots, q_n\}$, and a function $f(x, q)$, the coreset construction problem is to find a subset $C \subseteq X$ of size $|C| \leq \varepsilon |X|$ and weights $\gamma_x \geq 1$ for each $x \in X$ such that:

$$\sum_{x \in C} \gamma_x f(x, q) \geq (1 - \varepsilon) \sum_{x \in X} f(x, q)$$

Coresets are useful for large-scale machine learning problems, as they allow for efficient computation on a smaller subset of the data.

Coresets for VI in BGMM and DPGMM

Algorithm 1: Coreset for GMM

Input: $X$ data set, $M$ summary size, $\mu, \Sigma$.

1. for $x \in X$ do
   1.1. $\gamma_x = \frac{1}{\sum_{i=1}^{M} \gamma_i}$
   1.2. $q(x) = \frac{1}{\sum_{i=1}^{M} \exp\left(-\frac{1}{2} \left(\frac{||x - \mu_i||^2}{\Sigma_i}\right)\right)}$
   1.3. end for

2. $C \leftarrow \{x \in X : q(x) > \varepsilon \}$

3. for $x \in C$ do
   3.1. $\gamma_x = \gamma_x / \sum_{c \in C} \gamma_c$
   3.2. end for

4. return Coreset $C$

Coresets constructed for the log-likelihood can be used for posterior inference.

The size of previous coresets for GMM log-likelihood depends on $\kappa(\Sigma)$.

Integration over GMM parameter space $\theta = \{\nu_1, \mu_1, \Sigma_1, \ldots, \nu_T, \mu_T, \Sigma_T\}$ is problematic.

Theorem. For $M \in \Omega\left(\frac{D^3p^2}{\varepsilon^2}\right)$ Alg 1 returns $C$ s.t. w.p. $1 - \delta$:

$$|\phi(X) - \phi(C)| \leq \epsilon |\phi(X)| + \epsilon \sum_{i=1}^{M} \text{Tr}(\Sigma_i) \sum_{i=1}^{M} ||x_i - \mu_i||^2$$

where $\phi(X) = -\mathcal{L}(X) + n \log \sum_{i=1}^{T} \frac{1}{\sqrt{\text{det} \Sigma_i}} |x_i - \mu_i|^2$.

Corollary. Similar approximation guarantee holds for the ELBO.

Alg 1 can be used with weighted: CAVI, SVI, ADVI, BBVI.

Experiments and Discussion

- same coreset construction as for K-Means [1] works for BGMM and DPGMM
- coresets help if sampling distribution $q$ has low entropy -- the data is not evenly spread out
- coresets offer a N/M reduction in runtime and memory
- but VI converges with fewer iterations on the coreset

References