Finite Mixture Models: Beyond the EM Algorithm

Carlos Gomez-Uribe **(cgu)**, Viktor Krapivin **(krapivin)**, Grace Woods **(grawoods)**
CS 229: Machine Learning, Autumn 2019

Summary

Finite Mixture Models (FMMs), such as Gaussian mixtures, have been used and studied for several decades. The log-likelihood, $\ell(\theta)$, of FMMs is thought to be intractable, resulting in the Expectation-Maximization (EM) algorithm that maximizes a lower bound for $\ell(\theta)$. We propose online alternatives to EM that may result in better performance or accuracy. Some rely on the same lower bound as used by EM, and others on a different delta-method approximation of $\ell(\theta)$. We also find $\ell(\theta)$ is actually tractable, and leads to identical gradients as EM when the E-step is tight.

Model

$$z^{(i)} \sim \text{Categorical}(\phi) \qquad x^{(i)}|z^{(i)} = j \sim \text{Exp}(\eta_j) \longrightarrow \phi_j = \frac{e^{\theta_j}}{1 + \sum_{i=1}^{k-1} e^{\theta_j}}, \text{ for } j = 1, \ldots, k-1$$

Model parameters are Θ and the natural parameter, η

Approach

- Derive new approximate objective
- Derive gradients and Fisher information matrices for $\ell(\theta)$ for new objective
- Implement resulting SGD algorithms
- Compare accuracy and computation of different approaches

Model Fitting Approaches

- **EM**

$$P(x|z=\phi) = \prod_{j=1}^{k} p_j(x)^{\phi_j} \longrightarrow \text{ELBO}(x) = \sum_{j=1}^{k} q_j(x) \log \phi_j p_j(x)$$

- **Delta-Approximation Method**

$$\tilde{\ell}(x) \approx \underbrace{\log P(x|z=\phi)}_{\ell_0(x)} + \underbrace{\log\left(1 + \frac{\sum_{i,j=1}^{k} \sigma_{ij}^2 \partial_{z_i z_j}^2 P(x|z=\phi)}{2P(x|z=\phi)}\right)}_{\ell_2(x)}.$$

where $q_j(x) = \phi_j(x)\frac{P_j(x)}{P(x)}$, $\ell_0(x) = \sum_{j=1}^{k} \phi_j l_j(x), = \bar{l}(x)$, $\ell_2(x) = \log\left(1 + \frac{1}{2}\sigma_l^2(x)\right)$.

Algorithms

True objective or ELBO

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \lambda\phi_j^{(t)}(\alpha_j(x) - 1),$$
$$\eta_j^{(t+1)} = \eta_j^{(t)} + \lambda\phi_j^{(t)}\alpha_j(x)\Phi_j^{-1}e_j(x).$$

Using $\tilde{\ell}(x)$:

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \lambda\phi_j^{(t)}\left(\Delta_j(x) + \frac{\Delta_j^2(x) - \sigma_l^2(x)}{2 + \sigma_l^2(x)}\right),$$
$$\eta_j^{(t+1)} = \eta_j^{(t)} + \lambda\phi_j^{(t)}\left(1 + \frac{2\Delta_j(x)}{2 + \sigma_l^2(x)}\right)\Phi_j^{-1}e_j(x).$$

$$\bar{l}(x) = \sum_{i=1}^{k} \phi_i l_i(x) \longrightarrow \Delta_j(x) = l_j(x) - \bar{l}(x) \longrightarrow \sigma_l^2(x) = \sum_{i=1}^{k} \phi_i \Delta_i(x)^2$$

$$e(x) = T(x) - E[T(x)|\eta] = \begin{bmatrix} x - \mu \\ \text{vec}\left(xx' - \Sigma - \mu\mu'\right) \end{bmatrix}.$$

Second-Order Methods

- **Newton's Method**

$\ell(x)$: $F_\gamma = E\left[\frac{\partial\ell(x)}{\partial\gamma}\frac{\partial\ell(x)}{\partial\gamma}\right] \longrightarrow$

$$F_\theta^{(1:t+1)} = F_\theta^{(1:t)} + F_\theta,$$
$$F_{\eta_j}^{(1:t+1)} = F_{\eta_j}^{(1:t)} + F_{\eta_j},$$
$$\theta^{(t+1)} = \theta^{(t)} + \left(F_\theta^{(1:t+1)}\right)^{-1}\frac{\partial\ell(x)}{\partial\theta},$$
$$\eta_j^{(t+1)} = \eta_j^{(t)} + \left(F_{\eta_j}^{(1:t+1)}\right)^{-1}\frac{\partial\ell(x)}{\partial\eta_j}.$$

where $F_\theta = E\left[\frac{\partial\ell(x)}{\partial\theta}\frac{\partial\ell(x)}{\partial\theta}\right]$,
$F_{\eta_j} = E\left[\frac{\partial\ell(x)}{\partial\eta_j}\frac{\partial\ell(x)}{\partial\eta_j}\right]$

$\tilde{\ell}(x)$:

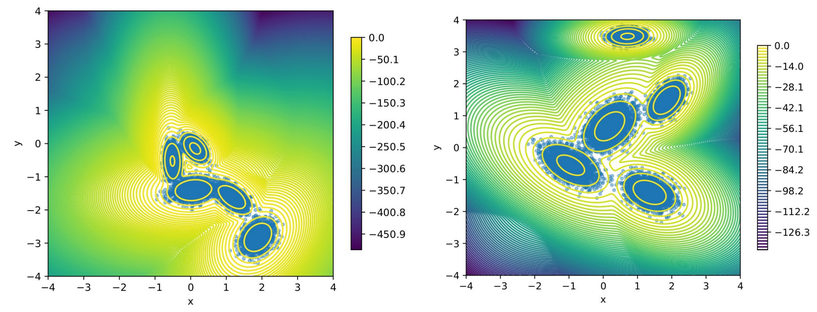
$$F_\theta = \Lambda E\left[\left(\Delta(x) + \frac{\Delta^2(x) - \sigma_l^2(x)}{2 + \sigma_l^2(x)}\right)\left(\Delta(x) + \frac{\Delta^2(x) - \sigma_l^2(x)}{2 + \sigma_l^2(x)}\right)'\right]\Lambda,$$
$$F_{\eta_j} = \phi_j^2\Phi_j^{-1}E\left[\left(1 + \frac{\Delta^2(x) - \sigma_l^2(x)}{2 + \sigma_l^2(x)}\right)^2 e_j(x)e_j'(x)\right]\Phi_j^{-1}.$$

Future Steps

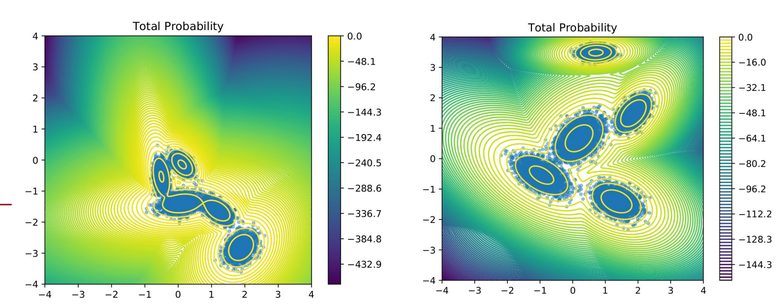
- Implement 2nd order learning methods (i.e. Newton's Method) via Fisher information and Monte Carlo Simulations
- Implement delta-approximation method for Variational Autoencoders (VAEs)

Results for SGD

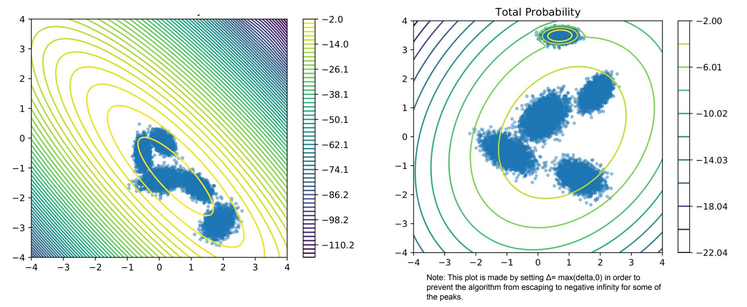
Generated Model Parameters:



True Objective or ELBO(x):



$$\ell(x) \approx \ell_0(x) + \ell_2(x)$$



1. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
2. Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin "Maximum likelihood from incomplete data via the EM Algorithm. "Journal of the Royal Statistical Society: Series B (Methodological) 39.1 (1977): 1-22.
3. Amari, Shun-ichi. "Neural learning in structured parameter spaces-natural Riemannian gradient." Advances in neural information processing systems. 1997.
4. Gómez-Uribe, Carlos Alberto, and Brian Karrer. "The decoupled extended Kalman filter for dynamic exponential-family factorization models." arXiv preprint arXiv:1806.09976 (2018).
5. McLachlan, Geoffrey, and Thriyambakam Krishnan. The EM algorithm and extensions. Vol. 382. John Wiley & Sons, 2007.
6. Puskorius, Gintaras V., and Lee A. Feldkamp. "Decoupled extended Kalman filter training of feedforward layered networks." Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on. Vol. 1. IEEE, 1991.
7. Y. Ollivier. Online natural gradient as a kalman filter. arXiv preprint arXiv:1703.00209, 2017.