Repulsive Mixture Variational Embedding

Omid S. Solari[*] SOLARI@BERKELEY.EDU
Department of Statistics
University of California, Berkeley

Yulun Wu[*] YULUN_WU@BERKELEY.EDU
Department of Statistics
University of California, Berkeley

James B. Brown JBBROWN@LBL.GOV
Lawrence Berkeley National Laboratory and Department of Statistics
University of California, Berkeley

Editor:

Abstract

tentative achievements:

1. separable and stable latent mixtures
2. better performance on entangled clustering
3. cluster single cell data and map them to body

Keywords: Gaussian Mixtures, Clustering, Variational Auto-Encoders

1. Introduction

- review use cases of vae/gmvae
- review uses in genomics specifically

2. Background

- review k-means and gaussian mixture models
- review variational auto-encoder
- introduce repulsive priors
- set up the use for single cell data

mention the shortcomings of GMVAE.

∗. equally contributing authors, alphabetical order

3. Problem Statment

general definition of Gaussian mixture models

4. Model

4.1 Gaussian Mixture Variational Auto-Encoder

Dependency structure: (draw a graph)

4.1.1 Variational Lower Bound

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{v}_{latent}|\boldsymbol{v}_{observed})}\left[\log \frac{p_\theta(\boldsymbol{v}_{latent}, \boldsymbol{v}_{observed})}{q_\phi(\boldsymbol{v}_{latent}|\boldsymbol{v}_{observed})}\right] \tag{1}$$

4.1.2 Supervised

TODO: define notations

$$\begin{aligned}\mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})\right] &= -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}\right] + \log p_\theta(\boldsymbol{x},\boldsymbol{y}) \\ &= -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) + \log p_\theta(\boldsymbol{z}|\boldsymbol{y}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})\right] \\ &\quad - \log \frac{p(\boldsymbol{y})}{q_\phi(\boldsymbol{y}|\boldsymbol{x})} - \log q_\phi(\boldsymbol{y}|\boldsymbol{x}) + \log p_\theta(\boldsymbol{x},\boldsymbol{y})\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] + \log q_\phi(\boldsymbol{y}|\boldsymbol{x}) \\ - \mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{y})\right] - \left[q_\phi(\boldsymbol{y}|\boldsymbol{x}) \parallel p(\boldsymbol{y})\right]\end{aligned} \tag{3}$$

(Sometimes people add term $\alpha \log q_\phi(\boldsymbol{y}|\boldsymbol{x})$ to amplify classification loss.)

4.1.3 Unsupervised

$$\mathcal{D}\left[q_\phi(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})\right] = -\mathbb{E}_{q_\phi(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{q_\phi(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})}\right] + \log p_\theta(\boldsymbol{x}) \tag{4}$$

$$\begin{aligned}\mathcal{L}_{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] \\ - \mathbb{E}_{q_\phi(\boldsymbol{y}|\boldsymbol{x})}\mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{y})\right] - \mathcal{D}\left[q_\phi(\boldsymbol{y}|\boldsymbol{x}) \parallel p(\boldsymbol{y})\right]\end{aligned} \tag{5}$$

4.1.4 Estimation and Propagation

Sampling estimation for q_ϕ: TODO

Understanding the back-propagation for unsupervised case: $\mathbb{E}_{q_\phi(\boldsymbol{y}|\boldsymbol{x})}\mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{y})\right]$

Lemma 1 *In sampling estimation, The KL divergence term that regularizes latent variable z: $-\mathbb{E}_{q_\phi(\boldsymbol{y}|\boldsymbol{x})}\mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{y})\right]$ in our model is a prior-weighted K-Means Model on latent prior $p_\theta(z)$.*

Proof

$$
\begin{aligned}
-\mathbb{E}_{q_\phi(\boldsymbol{y}|\boldsymbol{x})}\,\mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{y})\right] &= -\sum_{\boldsymbol{y}} q_\phi(\boldsymbol{y}|\boldsymbol{x}) \cdot \frac{1}{s} \sum_{\boldsymbol{z_1} z_s \in q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})} \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}{p_\theta(\boldsymbol{z}|\boldsymbol{y})} \\
&= \frac{1}{s} \sum_{\boldsymbol{y}} q_\phi(\boldsymbol{y}|\boldsymbol{x}) \sum_{\boldsymbol{z}} \log p(\boldsymbol{z}|\mu_\theta(\boldsymbol{y}), \boldsymbol{I}) - const(p_\theta) \qquad (6) \\
&= -\frac{1}{2s} \sum_{\boldsymbol{y}} q_\phi(\boldsymbol{y}|\boldsymbol{x}) \sum_{\boldsymbol{z}} ||\boldsymbol{z} - \mu_\theta(\boldsymbol{y})||^2 - const(p_\theta)
\end{aligned}
$$

This is a K-Means model with the learnt likelihood of labels $q_\phi(\boldsymbol{y}|\boldsymbol{x})$ replacing the standard weight $1/|S_{\boldsymbol{y}}|$. ■

In SGD update, we won't be able to track the total number of data in each mixture. Alternatively, weighting the squared deviations in this fashion makes the learning of label distribution more responsible to the latent prior.

Remark 2 *Unsupervised RMVE reduces to VaDE in a simple case.*

When we assume conditional independence $q_\phi(\boldsymbol{y}, \boldsymbol{z}|(x)) = q_\phi(\boldsymbol{y}|\boldsymbol{x})q_\phi(\boldsymbol{z}|\boldsymbol{x})$, and only consider regularizing $\boldsymbol{z}$: $\mathcal{D}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]$, it reduces to VaDE.

note that the update of prior parameter is only in response to $p_\theta(\boldsymbol{z}|\boldsymbol{y})$

4.2 Bayesian Repulsive Prior

find a nice drawing software for neural nets that allows depicting repulsion!

Lemma 3 *some lemma or theorem on covariance repulsion! Use HSIC*

4.3 Regularization

5. Experiments

5.1 Simulation

Simple MNIST is sufficient

5.2 Single Cell Clustering

Use The Human Cell Atlas data to show that 1. we can cluster them 2. with a simple linear transformation, those cells are mappable to their physical place! A cool picture ensues! Weinstein et al. (2013)

6. Conclusion

References

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.