Notes on the article : Hyperspherical Variational Auto-Encoders

November 14, 2023

1 Expected project outcomes

No report, only a 7-minute presentation : un **très court résumé de la contribution proposée dans l'article** et surtout vous concentrer sur une **critique en lien avec le cours**. Si c'est un article théorique, qu'apporte-t-il de plus que ce qu'on a vu en cours et est-ce un apport ou paie-t-on le prix ailleurs avec d'autres conditions ? **La preuve a-t-elle un schéma classique ?** etc.. Si l'article est méthodologique, **qu'en pensez-vous ? Pouvez-vous reproduire facilement les résultats de l'article avec votre implémentation ?**

2 Contribution

The Gaussian distribution is widely used in Variational Autoender models but it fails to model data with a latent hyperspherical structure. The authors replace it with the Von Mises-Fisher (vMF) distribution leading to a hyperspherical latent space. Their experiments show that :

- hyperspherical VAE, or $\mathcal{S}$-VAE, is more suitable to capture data with a hyperspherical latent structure,
- while outperforming a normal, $\mathcal{N} - VAE$, in low dimensions on other data types.

Some data types are better explained through spherical reprensentations like *directional data*?? (e. g. protein structure, observed wind directions).

Few machine learning techniques explicitly account for the intrinsically spherical structure of some data in the modeling process. Uniform distribution on the hypersphere is conveniently recovered as a special cas of the vMF.

3 Variational Auto encoders

In the "Generative" algorithm setup, we have access to examples X distributed according a unknown distribution $P_{gt}(X)$ and we aim at learning a model $P_\theta(X)$ which we can sample from, and which as similar as possible to $P_{gt}(X)$.
VAEs do make an approximation, but the error introduced by this approximation is arguably small given high-capacity models.

Notations :

$z \in \mathbb{R}^M$: latent variables
x : a vector of D observed variables
$p_\phi(x, z)$: a parameterized model of the joint distribution
$q(z)$: approximate posterior distribution
$q_\psi(z|x;\theta)$: parameterized by a NN that
outputs a probability distribution of z for each data points

We want to maximize the log-likelihook of the data by maximizing the Evidence Lower Bound (ELBO).

$$\mathcal{L}(\phi, \psi) = \mathbb{E}_{q_\psi(z|x;\theta)}[logp_\phi(x|z)] - KL(q_\psi(z|x;\theta)||p(z))$$

The first term is the reconstruction loss and the second term is the KL divergence between the approximate posterior and the prior $p(z)$ which is usually $\mathcal{N}(0, 1)$. Both the prior $p(z)$ and the posterior are defined as normal distributions.

Limitation of gaussian distribution :

- Low dimensions : origin gravity
- High dimensions : soap bubble effect **The standard Gaussian distribution in high dimensions tends to resemble a uniform distribution on the surface of a hypersphere, with the vast majority of its mass concentrated on the hyperspherical shell.**

4 Replacing Gaussian with Von-Mises Fisher

Analogous to the Gaussian distribution, the vMF is parameterized by $\mu \in \mathbb{R}^m$, the mean direction, and $\kappa \in \mathbb{R}^+$, the concentration parameter around mu. For $\kappa = 0$, the vMF distribution represents a Uniform distribution on the hypersphere. For a random unit vector $z \in \mathbb{R}^m$, $(z \in \mathcal{S}^{m-1})$, the vMF distribution is defined as :

$$q(z|\mu, \kappa) = C_m(\kappa)exp(\kappa\mu^T z)$$
$$C_m(\kappa) = \frac{\kappa^{m/2-1}}{(2\pi)^{m/2} I_{m/2-1}(\kappa)}$$

where $||\mu||^2 = 1$, and I_v denotes the modified Bessel function of the first kind of order v. Using the vMF distribution as an approximate posterior, we are able to place a uniform prior on the latent space. The KL divergence to

optimize writes :

$$KL(q_\psi(z|x;\theta)||p(z)) = KL(vMF(\mu,\kappa)||U(S^{m-1})) \\ = \kappa\frac{I_{m/2}(k)}{I_{m/2-1}(k)} + logC_m(\kappa) - log\left(\frac{2\pi^{m/2}}{\Gamma(m/2)}\right)^{-1}$$

Since the KL term does not depend on μ, this is only optimized in the reconstruction term. The above expression cannot be handled by automatic differentiation packages because of the modified Bessel function. Thus, we derive the gradient of the KL divergence with respect to κ.

- **To sample from vMF** using the procedure of Ulrich
  sampling $q(z|e_1,\kappa)$ with $e_1 = (1,0,\ldots,0)$ using acceptance-rejection sampling
  apply an orthogonal transformation U so that the transformed sampled is distributed according to $q(z|\mu,\kappa)$
  The Householder reflection simply finds an orthonormal transformation $U(\mu)$ that maps vector e_1 to μ.
  apply U to z' obtained from w and another sample v from a uniform distribution **The sampling technique does not suffer from the curse of dimensionality as we first sample from a univariate distribution**

- **Extend the reparametrization trick to vMF**
  Lemma 2 shows that the distribution can be reparameterized

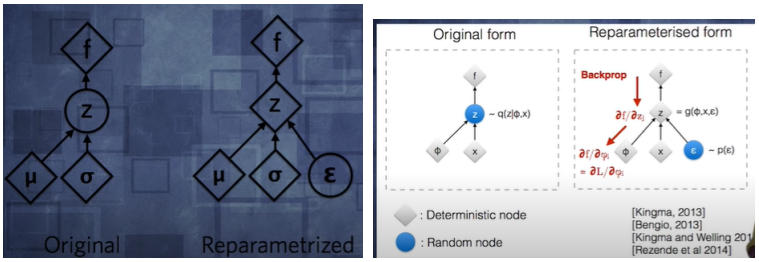
- In high dimension, the surface area collapse to zero (*vanishing surface problem*) $\rightarrow$ unstable behavior of this model in high dimensions

variational inference : optimization method that aims at estimating probability distributions with a simpler one

acceptance-rejection scheme : ???

Householder reflection : ???

Reparameterization trick : How to used backpropagation through the sampling node ?
Reparameterization trick enables backpropagation through the sampling node by introducing a deterministic variable (named accepted variable?) ϵ and a differentiable function g such that a sample $z \sim q_\psi(z|x;\theta)$ is expressed as $z = g(\epsilon, \theta, x)$. Indeed, after sampling *epsilon* during forward pass, it is fixed.



5 Experiments

- they show that $\mathcal{S}$-VAE outperform $\mathcal{N}$-VAE in recovering a hyperspherical latent structure
  to investigate theoretical properties of $\mathcal{S}$-VAE

  1. generate samples from three vMFs on the circles, mapped in higher dimension
  2. train an auto-encoder a $\mathcal{N}$-VAE, a $\mathcal{S}$-VAE and a disentangled $\mathcal{N}$-VAE

- thorough comparision with $\mathcal{N}$-VAEs on the MNIST dataset throught an unsupervised and semi-supervised learning task
  $\mathcal{S}$-VAE create a better separable latent representation to enhance classification (compare on a dataset that does not have a clear hyperspherical latent structure)

  1. reconstruction task using MNIST :
  2. compute ELBO, KL, negative reconstruction error, and marginal log-likelihood (LL) on test data

- $\mathcal{S}$-VAEs improve link prediction performance for three citation network datasets in combination with a *Variational Graph Auto-encoder* (VGAE)
  data with a non-Euclidean latent representation of low dimensionality : https://paperswithcode.com/paper/variational-graph-auto-encoders

  1. considere undirected and unweighted graphs
  2. popular citation network datasets : Cora, Citeseer and Pubmed. goal : recover link between node
  3. models trained on an incomplete version of the dataset where parts of the citation links (edges) have been removed while all nodes features are kept
  4. evaluation : area under the ROC curve (AUC) and average precision (AP)
  5. load dataset : av pyg

link prediction : ?
citation network dataset : ?

6 Conclusion

7 My feedbacks / comments / critics

Code with pytorch freely available on: https://github.com/nicola-decao/s-vae-pytorch

From first meeting held on 2023-11-08 : the project will consist in three important parts :

- introduction
- present variational auto-encoder (from the paper on *Tutorial on VAE*)
- present methods involved in the replacement of gaussians with vMF distributions (sampling and reparametrization trick)

- reproduce one experiment from the article, preferably the third one about link prediction on graph : *it is reproduceable with our limited means ?*
- conclusion

Possible problems : what criticism related to the course can we make ?