Decision-Making with Auto-Encoding Variational Bayes

Cyriaque Rousselot
Master Data Science
Université de Lille
cyriaque.rousselot@etu.univ-lille.fr

Gabriel Loiseau
Master Data Science
Université de Lille
gabriel.loiseau@etu.univ-lille.fr

Abstract

Variational Inference can provide approximate inference with good results. Variational AutoEncoders provide state-of-the art results in [3] many applications. Previous works used the variational distribution as a proposal distribution for sampling and introduced estimators. The question this article is trying to answer is : can the probability distribution found with variational Bayes be a good estimation of a posterior for Bayesian Decision Making ?

1 Variational Auto-Encoder

An Auto-Encoder with Variational Bayes [2] is a hierarchical Bayesian model. Variational Auto-Encoder models suppose the existence of a hidden latent variable $\mathbf{z}$ that could explain some data $\mathbf{x}$.

In the Variational inference framework, an intractable estimation of a posterior is reduced to a simpler problem. A family of parametrized tractable distributions $q_\phi(z|x)$ is considered, and an optimisation problem arise.

In the special case of Variational auto-encoder, The VAE usually tries to solve the ELBO problem : finding in a family of $q_\theta(x)$ one that minimize the KL divergence and thus maximize the ELBO :

$$\log p_\theta(\mathbf{x}) = \underbrace{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z})} \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{q_\phi(\mathbf{z})}}_{\mathrm{ELBO}[q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z},\mathbf{x})]} + \underbrace{\mathrm{KL}[q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z} \mid \mathbf{x})]}_{\text{divergence}}. \tag{1}$$

2 Decision Making formulation

Let us consider a state space $\mathcal{S}$, an action space $\mathcal{A}$, a loss $\mathcal{L}(a, z)$ with $a \in \mathcal{A}$ and $s \in \mathcal{S}$. There exist a model $p_\theta(x, z)$ defined over the observations of x and latent variable z. By Subjective expected utility, the Bayes optimal action is chosen by minimizing :

$$a^*(x) \in \underset{a \in \mathcal{A}}{\arg\min} \mathbb{E}_{p_\theta(z|x)} \mathcal{L}(z, a) \tag{2}$$

The article generalize over all functions $f(x)$ depending on the loss

$$a^*(x) = \underset{a \in \mathcal{A}}{\arg\min} \mathbb{E}_{p_\theta(z|x)} f(x) \tag{3}$$

This action choice require the estimation of the posterior $p_\theta(z|x)$. The main focus of this article is how to use the variational distribution $q_\phi(z|x)$ as a surrogate for $p_\theta(z|x)$ and how to do it efficiently.

All our experiments are available in this repository : https://github.com/GabrielLoiseau/Decision-Making-VAE

Existing Estimators Several estimator of $\mathcal{Q}(f,x) := \mathbb{E}_{p_\theta(z|x)} f(x)$ are proposed in the literature. The plugin estimator $\hat{\mathcal{Q}}^n_{\mathrm{P}}$ is obtained by sampling z_i from $q_\theta(z|x)$

$$\hat{\mathcal{Q}}^n_{\mathrm{P}}(f,x) = \frac{1}{n}\sum_{i=1}^{n} f\left(z_i\right) \tag{4}$$

The self-normalized importance sampling estimator [1] is obtained similarly but with weights $w(x,z) := p_\theta(x,z)/q_\phi(z \mid x)$.

The weight are computed through the use of another loss than the ELBO loss for the variational Auto Encoder, the IWELBO. They are used quantify the relevance of the sample with respect to the true distribution.

$$\hat{\mathcal{Q}}^n_{\mathrm{IS}}(f,x) = \frac{\sum_{i=1}^{n} w\left(x, z_i\right) f\left(z_i\right)}{\sum_{j=1}^{n} w\left(x, z_j\right)} \tag{5}$$

It is known that those estimators have bad asymptotic behaviors. Several others estimators such as those coming from χ^2 loss are also studied in the paper.

To clarify our report, we will focus on the effect of choosing the ELBO loss or the IWELBO loss. We will also choose two specific experiments where the results appears more clearly.

3 Working with pPCA

Since it can be complicated to do theory on untractable problems, the authors focus on a problem with closed form solution, the probabilistic PCA. With those assumptions on the distribution of the model, they are able to show a concentration bound between the estimator with Importance sampling and the true posterior. They are able to quantify the closeness of the estimator to the real posterior.

Setup of the problem In this context, the likelihood has a closed form, giving nice properties for estimating boundaries.

$$\begin{aligned} &p(z) \sim \mathcal{N}(0, I) \\ &p(x \mid z) \sim \mathcal{N}(W(z) + \mu, \sigma^2 I) \quad f \in F \quad c > 0 \\ &q_\phi(z|x) \sim \mathcal{N}(\mu_q, D_q(x)) \end{aligned} \tag{6}$$

$\mu_q, D_q(x)$ are estimated by the neural network.

Sufficient Sample Size Theorem If the number of importance sampling particles n satisfies $n = \beta \exp\{\Delta KL(p_\theta || q_\phi)\}$ for some $\beta > \log t^*(x)$,

$$\mathbb{P}\left(\left|\hat{\mathcal{Q}}^n_{IS}(f,x) - \mathcal{Q}(f,x)\right| \geq \frac{2\sqrt{3\kappa}}{\beta^{1/8\gamma} - \sqrt{3}}\right) \leq \frac{\sqrt{3}}{\beta^{1/8\gamma}} \tag{7}$$

Experiments results By taking a toy example for $f_\nu(z) = \mathbb{1}\left\{e_1^\top z \geq \nu\right\}$, we get the cumulative density function of a Gaussian distribution. our experiments shows that the estimator is quite close.

4 Working on MNIST

On the MNIST Dataset, we come up with another toy decision problem. The action $a(x)$ is either 0 if the corresponding number is a 9 or 1 if the corresponding number is not a 9. The loss $L(a,c)$ is

$$L(a,c) := a(x) * \mathbb{1}_{c=9} + (1 - a(x)) * \mathbb{1}_{c!=9}$$

The optimal action a* is

$$a^*(x) := argmin \int L(a(x), c) p_\theta(c|x)$$

In the VAE the loss is computed using

$$\mathbb{KL}\left(\mathcal{N}(\mu,\sigma) \parallel \mathcal{N}(0,1)\right) = \sum_{x \in X}\left(\sigma^2 + \mu^2 - \log\sigma - \frac{1}{2}\right) \tag{8}$$

5 Case study of the paper

The authors focus on the case of Variational Autoencoder applied to pPCA. Their main claim is that they justify the need for employing a different variational bound on the evidence for model fitting and for decision making.

Their theoretical procedure for decision making with VAE can be resumed by the following points :

- First, we fit multiple VAEs with different variational distribution (for example from the paper : IWAE, RWS, CHIVAE, WW)
- We keep the best model based on a surrogate of the likelihood
- Then we learn variational approximation to the model posterior
- Finally we can estimate the optimal decision from multiple importance sampling

They also tested their approach with MNIST using the rejection option in classification. The paper also presents a case study on controlling the posterior expected FDR for hypothesis testing on single cell RNA sequencing data.

6 Conclusion

We spent a lot of times trying to build the IWELBO from this paper and we didn't felt there was a lot of information in this paper to implement it. The VAE we built is perhaps different from the one used in the paper but we didn't found sufficient details on how exactly to replicate their experiments. It was also our first real experience with Pytorch.

7 Strong and weak points of the paper

About the strong points of the paper :

- First, the paper presents some good theoretical results of this problem of decision making with VAEs,
- Their experiments are well detailed within the sections, and they also provide the code to rerun those experiments.

Therefore, some weak aspects need to be considered :

- There is a loot of notions that are not directly presented in the paper, and without any expertise on the subject it is quite intense to get through without tackling many other papers about the subject.
- Also, the code present in their repository is also quite hard to get through, is has been made to be so optimized and packed that it is very complex to disassemble.

References

[1] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. *Importance Weighted Autoencoders*. 2016. arXiv: 1509.00519 [cs.LG].

[2] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].

[3] Arash Vahdat and Jan Kautz. *NVAE: A Deep Hierarchical Variational Autoencoder*. 2021. arXiv: 2007.03898 [stat.ML].