Student ID : 20194293
Name : Go, Kyeong Ryeol

[AI 602] Variational Inference with Normalizing Flows

1. Paper Summary

Approximate posterior inference is the key interest for many Bayesian researchers. Variational inference is the well-known technique that transforms the inference problem into the optimization domain, which suffices to find the optimal parameter of some parametric family of distribution that can well describe the data distribution. However, there is a significant challenge on setting the right family of the distribution which can vary task-by-task. Moreover, in the sense of tractability, very restricted forms were only allowed such as the factorization gaussians when using the mean field variational inference, which ends up underestimating the variance as the dependency across the elements are ignored.

The author proposed to use the normalizing flow which transforms a simple probability distribution through a series of invertible mappings. The final output turns out to be highly flexible enough to contain the true posterior as one solution. Most importantly, this is based on the change of variable theorem as follow.

Change of variables theorem : $\mathrm{q}(\mathrm{z}') = q(z)\left|\det\frac{\partial f^{-1}}{\partial z'}\right| = q(z)\left|\det\frac{\partial f}{\partial z}\right|^{-1} \quad where\ z' = f(z)$

$$\mathrm{z_K} = f_K \circ \ldots \circ f_2 \circ f_1(z_0) \rightarrow \ln q_K(z_K) = \ln q_0(z_0) - \sum_{k=1}^{K} \ln\left|\det\frac{\partial f_k}{\partial z_{k-1}}\right|$$

Therefore, the normalizing flow is nothing but a sequence of expansions and contractions on the initial distribution $\mathrm{q}_0(\cdot)$. When using appropriate transformation, any complex, multi-modal distribution can be obtained even with very simple initial distribution. However, since computing both the Jacobian determinant and its gradients scales $\mathrm{O(LD^3)}$, it can not choose any transformation. The author suggests two transformations as follows which can be computed in linear time. Then, the overall computation scales at most quadratic.

$$\mathrm{f(z) = z + uh}\left(\mathrm{w^T}z + b\right) \text{ or } \mathrm{f(z) = z + \beta h(\alpha, r)(z - z_0)}$$

2. In-depth discussions

Here, I propose one discussion point.

1. Is there any way to explicitly measure the tightness of the ELBO? Approximate posterior inference problems basically want the model distribution to be close to the true posterior distribution. However, since the true posterior is unknown, the distance between two distributions cannot be analytically computed. Therefore, many papers in this domain justify their ideas in qualitative way through visualization. Maybe, Importance Weighted AutoEncoder (IWAE) may give an answer to this problem, which newly design a new bound which can almost surely converge to the true log likelihood if infinitely many samples are used.