Student ID : 20194293
Name : Go, Kyeong Ryeol

[AI 602] The Functional Neural Process

1. Paper Summary

Bayesian neural networks posit a prior distribution over the weights of the networks and through inference they can represent their uncertainty in the posterior distributions. However, the choice of the prior is quite difficult and the inference over the weights of a neural network can be a daunting task. As an alternative, a stochastic process can be adopted which posit distributions over functions. Here, gaussian process is well-used one which encode any inductive bias in the form of a covariance structure among the datapoints, but it is not flexible enough for high dimensional problems and costly for training and inference as it scales cubically with the size of the dataset.

The author presents Functional Neural Process (FNP) which operate by building a graph of dependencies among local latent variables rather than requiring the explicit global latent variables like conventional neural process families. This prevents the information loss when extracting a global information shared across the datapoints and the sufficient conditions for the stochastic process from the Kolmogorov Extension, exchangeability and consistency, are satisfied.

First, the reference set R is chosen from the data inputs and with the remaining set M, it is embedded to a latent space to compute the local latent variables u. Then the graphs of dependencies, A and G, are constructed where A is a bipartite graph from R to M and G is the directed acyclic graph among the points in R. Here, each element of the adjacency matrix is sampled from the Bernoulli variable whose parameter is computed from the kernel. This is where the term 'functional' comes from so that the concept of similarity among the latent embeddings automatically defines the unique functional space as RKHS. After that, the final latent variables z is computed solely from the reference set R and its corresponding output y_R according to the dependency information extracted in the graph A and G. As a result, the prediction on the data output y can be computed as follows.

$$\text{FNP}: \sum_{A,G} \int p_\theta(u|X)p(A,G|u)p_\theta(y,z|R,A,G)\,du\,dz\,dy_R$$

FNP^+ is a modified version of FNP where the prediction on y is also conditional on the latent embedding u, which are empirically shown to be useful for extrapolation. This is mainly due to the nature of the kernel where it ends up with an uninformative standard normal prior over z if the points locate too far from the reference set R.

As the derivation of the maximum likelihood estimator requires too many marginalization of the variables, the author propose the variational learning scheme with the variational distribution $\mathrm{q}_\phi(u,A,G,z|X) = \mathrm{p}_\theta(\mathrm{u}|\mathrm{X})\mathrm{p}(\mathrm{A},\mathrm{G}|\mathrm{u})\mathrm{q}_\phi(z|X)$. Sharing $\mathrm{p}_\theta(u|X)p(A,G|u)$ with the joint probability $\mathrm{p}_\theta(u,A,G,z,y|X)$ enables the ELBO to be decomposed into simple form as $\mathrm{L_R} + L_{M|R}$ where each term corresponds to the set of the data inputs R and M, respectively. One last note is that $\mathrm{L_{M|R}}$ can be decomposed to |M| independent sums by its i.i.d nature which enables the minibatch optimization.

$$\mathrm{L_R}: E_{p_\theta(u_R,G|R)}\left[E_{q_\phi(z_R|R)}\left[log\, p_\theta(z_R|par_G(R,y_R)p_\theta(y_R|z_R)) - log\, q_\phi(z_R|R)\right]\right]$$

$$L_{M|R}: E_{p_\theta(u,A|X)}\left[E_{q_\phi(z_M|M)}\left[log\, p_\theta(z_M|par_A(R,y_R)p_\theta(y_M|z_M) - log\, q_\phi(z_M|M)\right]\right]$$

2. In-depth Discussions

Here I would like to offer 2 discussion points. To begin with, is it truly scalable? Personally speaking, contrast to what author insists, I don't think so as the cardinality of the reference set must get larger as the dimension of the data gets higher. That may be the reason why the experiments were conducted only on the simple datasets such as toy 1d regression, MNIST, and CIFAR10. Next, what would be the critical cost for constructing the graph for modeling the dependency? Actually, as a prior work, there was a research called 'Attentive neural process' which also models the dependency among the data points using the attention mechanism. However, since it has a global latent variable, it can be easily transferrable to the other domain while FNP may not. This is critical when applied to the reinforcement learning settings which are well explored area by the Bayesian neural network as these uncertainty measure can motivate exploration of the algorithm and therefore boost the convergence.