Review: Natural Posterior Network: Deep Bayesian Uncertainty for Exponential Family Distribution

Shin Yun Seop

January 02, 2024

Seoul national university, statistics, IDEA LAB

Uncertainty

- Aleatoric uncertainty: data uncertainty, irreducible uncertainty.(cannot be reduced even if additional data is input, etc measurement error)
- Epistemic uncertainty: model uncertainty, reducible uncertainty. (if additional data is input then it can be reduced)

Uncertainty

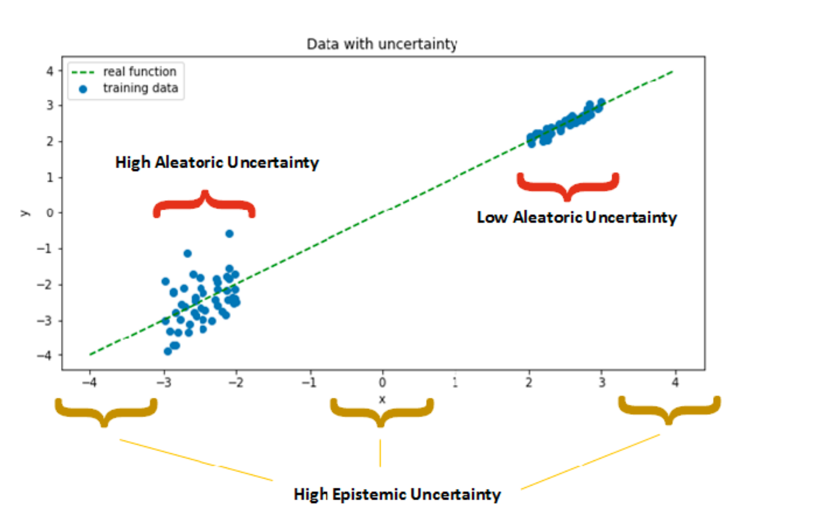

Figure 1: Type of uncertainty

Related Work

- Sampling-based methods: For example, ensemble, dropout based on bayesian neural network. $\Rightarrow$ Computation issue
- Sampling-free methods: They model uncertainty at the weight and/or activation levels $\Rightarrow$ Constrained to specific architecture

Natural Posterior Network

- It applies to many common supervised learning task type. (Classification, Regression, Count prediction)
- For every input, it predicts the parameters of the posterior over the target exponential family distribution.
- It requires only a single forward pass at testing time.

Flexible, Reliable, Fast & Practical

Bayes rule

Theorem

Bayes rule:

$$\mathbb{Q}(\theta|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\theta)\mathbb{Q}(\theta)$$

where, $\mathbb{P}(\mathcal{D}|\theta)$ *is the target distribution of the target data* $\mathcal{D}$ *given its parameter* θ*, and* $\mathbb{Q}(\theta)$ *and* $\mathbb{Q}(\theta|\mathcal{D})$ *are the prior and posterior distributions, respectively, over the target distribution parameters.*

Exponential family

- Exponential family cover a wide range of target variables like discrete, continuous, counts or spherical coordinates.
- Parameters, density functions and statistics of exponential family can often be evaluated in closed-form.

Exponential family

Definition

Formally, an exponential family distribution on a target variable $y \in \mathbb{R}$ with natural parameters $\theta \in \mathbb{R}^L$ can be denoted as

$$\mathbb{P}(y|\theta) = h(y)exp(\theta^T u(y) - A(\theta))$$

where $h : \mathbb{R} \to \mathbb{R}$ is base measure, $A : \mathbb{R}^L \to \mathbb{R}$ and $u : \mathbb{R} \to \mathbb{R}^L$ the sufficient stastistics.

Exponential family

Theorem

An exponential family distribution always admits a conjugate prior, which often also is a member of the exponential family

$$\mathbb{Q}(\theta \mid \boldsymbol{\chi}, n) = \eta(\boldsymbol{\chi}, n) \exp\left(n\theta^T \boldsymbol{\chi} - nA(\theta)\right)$$

where $\eta(\boldsymbol{\chi}, n)$ is a normalization coefficient, $\boldsymbol{\chi} \in \mathbb{R}^L$ are prior parameters and $n \in \mathbb{R}^+$ is the evidence.

Exponential family(Continue)

Theorem

Given a set of N target observations $\left\{y^{(i)}\right\}_i^N$, *it is easy to compute a closed-form Bayesian update,*

$$\mathbb{Q}\left(\theta \mid \chi^{\text{post}}, n^{post}\right) \propto \exp\left(n^{post}\,\theta^T\chi^{post} - n^{post}\,A(\theta)\right)$$

where $\chi^{post} = \frac{n^{prior}\,\chi^{prior} + \sum_j^N \boldsymbol{u}\left(y^{(j)}\right)}{n^{prior} + N}$ *and* $n^{post} = n^{prior} + N$.
Also we can show that $\boldsymbol{\chi} = \mathbb{E}_Y(u(Y))$.
(Brown, 1986; Diaconis & Ylvisaker, 1979)

Posterior parameter update

- NatPN extends the Bayesian treatment of a single exponential family distribution prediction by predicting an individual posterior update per input.
- Distinguish between the chosen prior parameters $\boldsymbol{\chi}^{prior}, n^{prior}$ shared among sample, and the additional predicted parameter $\boldsymbol{\chi}^{(i)}, n^{(i)}$ dependent on the input $x^{(i)}$ leading to the updated posterior parameters.
- The updated posterior parameters per one input are followed:

$$\chi^{\text{post },(i)} = \frac{n^{\text{prior }} \chi^{\text{prior }} + n^{(i)} \chi^{(i)}}{n^{\text{prior }} + n^{(i)}}, \quad n^{\text{post },(i)} = n^{\text{prior }} + n^{(i)}$$

Model setting

- An arbitrary encoder f_ϕ maps the input $x^{(i)}$ onto a low-dimensional latent vector $z^{(i)} = f_\phi(x^{(i)}) \in \mathbb{R}^H$.
- A linear decoder g_ψ is trained to output the parameter update $\chi^{(i)} = g_\psi(z^{(i)}) \in \mathbb{R}^L$.
- A single normalized density(typically, radial flow or masked auto regressive flow are used) is trained to output the evidence update $n^{(i)} = N_H \mathbb{P}(z^{(i)}|\omega)$.
- N_H is hyper parameter depending on H. On paper authors recommend $\left\{e^{\frac{1}{2}H}, e^H, e^{\log(\sqrt{4\pi})H}\right\}$.
- So, to train the model need to optimize ϕ, ψ, ω.

Optimization

- Minimizing the Bayesian loss function.

$$\mathcal{L}^{(i)} = -\underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(i)} \sim \mathbb{Q}^{\text{post },(i)}} \left[\log \mathbb{P}\left(y^{(i)} \mid \boldsymbol{\theta}^{(i)} \right) \right]}_{\text{(i)}} - \underbrace{\mathbb{H}\left[\mathbb{Q}^{\text{post },(i)} \right]}_{\text{(ii)}}$$

  where $\mathbb{H}\left[\mathbb{Q}^{\text{post },(i)}\right]$ denotes the entropy of the predicted posterior distribution $\mathbb{Q}^{\text{post },(i)}$.

- This loss is guaranteed to be optimal when the predicted posterior distribution is close to the true posterior distribution $\mathbb{Q}^*\left(\boldsymbol{\theta} \mid \boldsymbol{x}^{(i)}\right)$ i.e. $\mathbb{Q}^{\text{post, },(i)} \approx \mathbb{Q}^*\left(\boldsymbol{\theta} \mid \boldsymbol{x}^{(i)}\right)$.

Optimization(Continue)

- The term (i) is the expected likelihood under the predicted posterior distribution.
- The term (ii) is an entropy regularizer acting as a prior which favors uninformative distributions $\mathbb{H}\left[\mathbb{Q}^{\text{post },(i)}\right]$ with high entropy.
- In our case, we assume the likelihood $\mathbb{P}\left(y^{(i)} \mid \boldsymbol{\theta}^{(i)}\right)$ and the posterior $\mathbb{Q}^{\text{post },(i)}$ to be member of the exponential family so we can calculate it in closed form.

Optimization(Continue)

Likelihood $\mathbb{P}$	Conjugate Prior $\mathbb{Q}$	Parametrization Mapping m	Bayesian Loss (Eq. 5)
$y \sim \mathrm{Cat}(\boldsymbol{p})$	$\boldsymbol{p} \sim \mathrm{Dir}(\boldsymbol{\alpha})$	$\boldsymbol{\chi} = \boldsymbol{\alpha}/n$ $n = \sum_c \alpha_c$	**(i)** $= \psi(\alpha_{y*}^{(i)}) - \psi(\alpha_0^{(i)})$ **(ii)** $= \log B(\boldsymbol{\alpha}^{(i)}) + (\alpha_0^{(i)} - C)\psi(\alpha_0^{(i)}) - \sum_c (\alpha_c^{(i)} - 1)\psi(\alpha_c^{(i)})$
$y \sim \mathcal{N}(\mu, \sigma)$	$\mu, \sigma \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$	$\boldsymbol{\chi} = \begin{pmatrix} \mu_0 \\ \mu_0^2 + \frac{2\beta}{n} \end{pmatrix}$ $n = \lambda = 2\alpha$	**(i)** $= \frac{1}{2}\left(-\frac{\alpha}{\beta}(y-\mu_0)^2 - \frac{1}{\lambda} + \psi(\alpha) - \log\beta - \log 2\pi\right)$ **(ii)** $= \frac{1}{2} + \log\left((2\pi)^{\frac{1}{2}}\beta^{\frac{3}{2}}\Gamma(\alpha)\right) - \frac{1}{2}\log\lambda + \alpha - (\alpha + \frac{3}{2})\psi(\alpha)$
$y \sim \mathrm{Poi}(\lambda)$	$\lambda \sim \Gamma(\alpha, \beta)$	$\chi = \alpha/n$ $n = \beta$	**(i)** $= (\psi(\alpha) - \log\beta)y - \frac{\alpha}{\beta} - \sum_{k=1}^{y}\log k$ **(ii)** $= \alpha + \log\Gamma(\alpha) - \log\beta + (1-\alpha)\psi(\alpha)$

Figure 2: Examples of Exponential Family Distributions where $\psi(x)$ and $B(x)$ denote Digamma and Beta function, respectively.

Uncertainty estimation

- Aleatoric uncertainty: The entropy of the target distribution $\mathbb{P}(y|\theta)$ was used to estimate the aleatoric uncertainty. i.e. $\mathbb{H}\left[\mathbb{P}(y|\theta)\right]$
- Epistemic uncertainty: The entropy of the posterior distribution $\mathbb{Q}\left(\theta \mid \chi^{\mathrm{post}}, n^{\mathsf{post}}\right)$ was used to estimate the epistemic uncertainty.

Overview

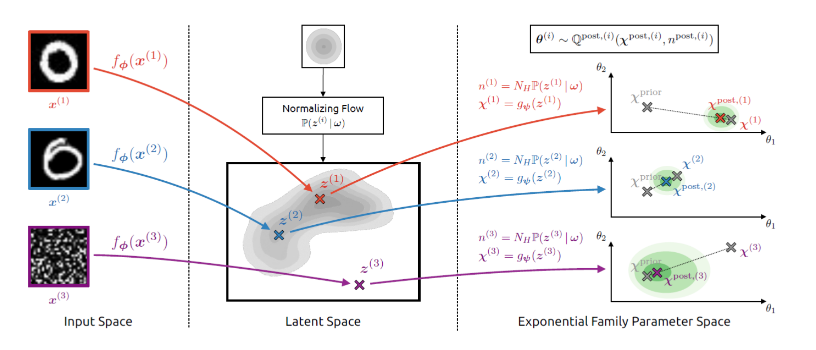

Figure 3: The right figure show epistemic uncertainty estimation. Third observation is highly uncertain.

Limitation

- NatPN is capable of detecting OOD samples only with respect to the considered task and requires labeled examples during training.
- This is because NatPN does not perform OOD detection directly on the input but rather fits a normalizing flow on a learned space.
- For example, NatPN likely fails to detect a change of image color if the task aims at classifying object shapes and the latent space has no notion of color.

Reference

1. Charpentier, Bertrand, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. "Natural Posterior Network: Deep Bayesian Uncertainty for Exponential Family Distributions." ArXiv.org (2021): ArXiv.org, 2021. Web.
2. Diaconis, Persi, and Donald Ylvisaker. "Conjugate Priors for Exponential Families." The Annals of Statistics 7.2 (1979): 269-81. Web.
3. Brown, Lawrence D. "Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory." Lecture Notes-monograph Series 9 (1986): I-279. Web.