Evidental Focal Loss Derivation

Ruxiao Duan

1 Preliminaries

For a random variable X following a beta distribution:

$$X \sim \text{Beta}(\alpha, \beta) \tag{1}$$

in which the support of X is $(0, 1)$ and the distribution parameters $\alpha, \beta > 0$, the probability density function of X is given by

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \quad \forall x \in (0, 1) \tag{2}$$

in which

$$\mathrm{B}(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3}$$

where $\mathrm{B}(\cdot, \cdot)$ and $\Gamma(\cdot)$ denote beta function and gamma function, respectively.

As shown in [3], the expectation of logarithm can be expressed as

$$\mathbb{E}[\log X] = \int_0^1 (\log x) f(x; \alpha, \beta) dx = \int_0^1 (\log x) \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} dx = \psi(\alpha) - \psi(\alpha+\beta) \tag{4}$$

in which $\psi(\cdot)$ represents digamma function. Therefore, it can be derived that

$$\int_0^1 (\log x) \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} dx = \psi(\alpha) - \psi(\alpha+\beta) \tag{5}$$

$$\int_0^1 (\log x) x^{\alpha-1}(1-x)^{\beta-1} dx = \mathrm{B}(\alpha, \beta)(\psi(\alpha) - \psi(\alpha+\beta)) \tag{6}$$

Replacing β in Eq. (6) with $\beta + \gamma$ where $\gamma \geq 0$ is another constant, we have

$$\int_0^1 (\log x) x^{\alpha-1}(1-x)^{\beta+\gamma-1} dx = \mathrm{B}(\alpha, \beta+\gamma)(\psi(\alpha) - \psi(\alpha+\beta+\gamma)) \tag{7}$$

2 Problem Formulation

In the classification setting, each target label $\boldsymbol{y} = [y_1, y_2, \ldots, y_K]^\top$ is a one-hot vector in which K is the number of classes. (The sample index i is omitted for simplicity.) The label $\boldsymbol{y}$ follows a categorical distribution with parameters $\boldsymbol{\mu}$

$$\boldsymbol{y} \sim \text{Cat}(\boldsymbol{\mu}) \tag{8}$$

in which $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]^\top$ is the class probability vector with $\mu_j \in [0, 1]\ \forall j \in \{1, 2, \ldots, K\}$, $\sum_{j=1}^K \mu_j = 1$, and $\text{Cat}(\cdot)$ denotes categorical distribution. The probability that the sample belongs to class j is μ_j.

Evidential deep learning assumes that $\boldsymbol{\mu}$ is a random vector following a Dirichlet distribution:

$$\boldsymbol{\mu} \sim \text{Dir}(\boldsymbol{\alpha}) \tag{9}$$

in which the Dirichlet parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]^\top$ and $\alpha_j > 0\ \forall j \in \{1, 2, \ldots, K\}$. The Dirichlet strength $\alpha_0 = \sum_{j=1}^K \alpha_j$.

3 Cross-Entropy Loss

According to [2], in principle, we can define any loss function ℓ and compute its Bayes risk with respect to the class predictor, i.e.,

$$\mathcal{L} = \int \ell(\boldsymbol{y}, \boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\ell(\boldsymbol{y}, \boldsymbol{\mu})] \tag{10}$$

In their paper, they demonstrated two options of ℓ: cross-entropy loss and sum-of-squares loss. For the cross-entropy loss

$$\ell^{\mathrm{CE}}(\boldsymbol{y}, \boldsymbol{\mu}) = -\sum_{j=1}^{K} y_j \log \mu_j \tag{11}$$

thus the Bayes risk can be derived as

$$\begin{aligned}
\mathcal{L}^{\mathrm{CE}} &= \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\ell^{\mathrm{CE}}(\boldsymbol{y}, \boldsymbol{\mu})] &(12)\\
&= \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[-\sum_{j=1}^{K} y_j \log \mu_j] &(13)\\
&= -\sum_{j=1}^{K} y_j \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\log \mu_j] &(14)\\
&= -\sum_{j=1}^{K} y_j (\psi(\alpha_j) - \psi(\alpha_0)) &(15)\\
&= \sum_{j=1}^{K} y_j (\psi(\alpha_0) - \psi(\alpha_j)) &(16)
\end{aligned}$$

The derivation of expectation of logarithm from Eq. (14) to Eq. (15) can be found in [4].

4 Focal Loss

As introduced in [1], focal loss is defined as

$$\ell^{\mathrm{Focal}}(\boldsymbol{y}, \boldsymbol{\mu}) = -\sum_{j=1}^{K} y_j (1-\mu_j)^{\gamma} \log \mu_j \tag{17}$$

in which $\gamma \geq 0$ is a hyperparameter that can be adjusted. Note that when $\gamma = 0$, focal loss reduces to cross-entropy loss. In their paper, $\gamma = 2$ gave the best performance.

The focal version of evidential loss can be derived as follows.

$$\begin{aligned}
\mathcal{L}^{\mathrm{Focal}} &= \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\ell^{\mathrm{Focal}}(\boldsymbol{y}, \boldsymbol{\mu})] &(18)\\
&= \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[-\sum_{j=1}^{K} y_j (1-\mu_j)^{\gamma} \log \mu_j] &(19)\\
&= -\sum_{j=1}^{K} y_j \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[(1-\mu_j)^{\gamma} \log \mu_j] &(20)
\end{aligned}$$

Now we only need to separately calculate $\mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[(1-\mu_j)^{\gamma} \log \mu_j]$. We can make use of the fact that the marginal distribution of Dirichlet distribution is beta distribution:

$$\mu_j \sim \mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j) \tag{21}$$

Therefore,

$$\begin{aligned}
\mathcal{L}^{\text{Focal}} &= -\sum_{j=1}^{K} y_j \mathbb{E}_{\boldsymbol{\mu}\sim\text{Dir}(\boldsymbol{\alpha})}[(1-\mu_j)^\gamma \log \mu_j] && (22)\\
&= -\sum_{j=1}^{K} y_j \mathbb{E}_{\mu_j\sim\text{Beta}(\alpha_j,\alpha_0-\alpha_j)}[(1-\mu_j)^\gamma \log \mu_j] && (23)\\
&= -\sum_{j=1}^{K} y_j \int_0^1 (1-\mu_j)^\gamma (\log \mu_j) f(\mu_j;\alpha_j,\alpha_0-\alpha_j) d\mu_j && (24)\\
&= -\sum_{j=1}^{K} y_j \int_0^1 (1-\mu_j)^\gamma (\log \mu_j) \frac{\mu_j^{\alpha_j-1}(1-\mu_j)^{\alpha_0-\alpha_j-1}}{\mathrm{B}(\alpha_j,\alpha_0-\alpha_j)} d\mu_j && (25)\\
&= -\sum_{j=1}^{K} \frac{y_j}{\mathrm{B}(\alpha_j,\alpha_0-\alpha_j)} \int_0^1 (\log \mu_j)\mu_j^{\alpha_j-1}(1-\mu_j)^{\alpha_0-\alpha_j+\gamma-1} d\mu_j && (26)
\end{aligned}$$

This expression can be simplified using Eq. (7). Replacing x, α, and β in Eq. (7) by μ_j, α_j, and $\alpha_0-\alpha_j$ respectively, we have

$$\int_0^1 (\log \mu_j)\mu_j^{\alpha_j-1}(1-\mu_j)^{\alpha_0-\alpha_j+\gamma-1} d\mu_j = \mathrm{B}(\alpha_j,\alpha_0-\alpha_j+\gamma)(\psi(\alpha_j)-\psi(\alpha_0+\gamma)) \tag{27}$$

By simple substitution,

$$\begin{aligned}
\mathcal{L}^{\text{Focal}} &= -\sum_{j=1}^{K} \frac{y_j}{\mathrm{B}(\alpha_j,\alpha_0-\alpha_j)}\mathrm{B}(\alpha_j,\alpha_0-\alpha_j+\gamma)(\psi(\alpha_j)-\psi(\alpha_0+\gamma)) && (28)\\
&= -\sum_{j=1}^{K} \frac{y_j}{\frac{\Gamma(\alpha_j)\Gamma(\alpha_0-\alpha_j)}{\Gamma(\alpha_0)}} \frac{\Gamma(\alpha_j)\Gamma(\alpha_0-\alpha_j+\gamma)}{\Gamma(\alpha_0+\gamma)}(\psi(\alpha_j)-\psi(\alpha_0+\gamma)) && (29)\\
&= -\sum_{j=1}^{K} y_j \frac{\Gamma(\alpha_0)\Gamma(\alpha_0-\alpha_j+\gamma)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_0+\gamma)}(\psi(\alpha_j)-\psi(\alpha_0+\gamma)) && (30)\\
&= \sum_{j=1}^{K} y_j \frac{\Gamma(\alpha_0)\Gamma(\alpha_0-\alpha_j+\gamma)}{\Gamma(\alpha_0-\alpha_j)\Gamma(\alpha_0+\gamma)}(\psi(\alpha_0+\gamma)-\psi(\alpha_j)) && (31)
\end{aligned}$$

It can be observed that $\mathcal{L}^{\text{Focal}} = \mathcal{L}^{\text{CE}}$ when $\gamma = 0$.

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[2] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 2

[3] Wikipedia contributors. Beta distribution — Wikipedia, the free encyclopedia, 2023. [Online; accessed 27-July-2023]. 1

[4] Wikipedia contributors. Dirichlet distribution — Wikipedia, the free encyclopedia, 2023. [Online; accessed 27-July-2023]. 2