Predictive Uncertainty Estimation via Prior Networks

Ksenofontov Gregory

MIPT

December 18, 2022

Previous approaches

Bayesian class:

1. performance depends on the form of approximation and the nature of the prior distribution of parameters
2. implicitly model distributional uncertainty through model uncertainty

Non-Bayesian class:

1. minimizing its KL-divergence to both a in-domain posterior and a out-of-domain posterior
2. conflate distributional uncertainty with data uncertainty

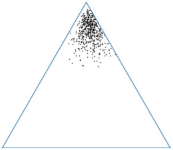
Bayesian class

Consider a distribution $p(\mathbf{x}, y)$ over input features x, labels y and classification model $P(y = \omega_c|\mathbf{x}^*, \theta)$, trained on $D = \{\mathbf{x}_j, y_j\}_{j=1}^N \sim p(\mathbf{x}, y)$. So, in Bayesian framework the uncertainty is:

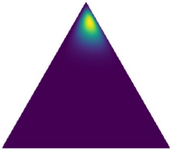
$$P(\omega_c|\mathbf{x}^*, D) = \int P(\omega_c|\mathbf{x}^*, \theta)p(\theta|D)d\theta$$

where $P(\omega_c|\mathbf{x}^*, D)$ - data uncertainty, $p(\theta|D)$ - model uncertainty

$$P(\omega_c|\mathbf{x}^*, D) \approx \frac{1}{M}\sum_{i=1}^{M} P(\omega_c|\mathbf{x}^*, \theta^{(i)}), \theta^{(i)} \sim q(\theta)$$



(a) Ensemble

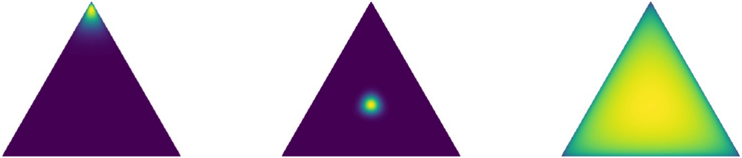


(b) Distribution

Prior Networks

$$P(\omega_c|\mathbf{x}^*, D) = \int\int p(\omega_c|\mu)p(\mu|\mathbf{x}^*, \theta)p(\theta|D)d\mu d\theta$$

where $p(\omega_c|\mu)$ - data uncertainty, $p(\mu|\mathbf{x}^*, \theta)$ - distributional uncertainty



(a) Confident Prediction (b) High data uncertainty (c) Out-of-distribution

Dirichlet Prior Networks

Considering marginalization of θ in last equation:

$$\int p(\omega_c|\mu) \int \left[p(\mu|\mathbf{x}^*, \theta) p(\theta|D) d\theta \right] d\mu = \int p(\omega_c|\mu) p(\mu|\mathbf{x}^*, D)$$

However, marginalization is generally intractable, so:

$$p(\theta|D) = \delta(\theta - \hat{\theta}) \to p(\mu|\mathbf{x}^*, D) \approx p(\mu|\mathbf{x}^*, \hat{\theta}))$$

Dirichlet Prior Networks

$$p(\mu|\mathbf{x}^*, \hat{\theta}) = Dir(\mu|\alpha), \alpha = \mathbf{f}(\mathbf{x}^*; \hat{\theta})$$

So, posterior over class labels will be given by the mean of the Dirichlet:

$$P(\omega_c|\mathbf{x}^*, \hat{\theta}) = \int p(\omega_c|\mu)p(\mu|\mathbf{x}^*, \hat{\theta}) = \frac{\alpha_c}{\alpha_0}$$

where $\alpha_c > 0, \alpha_0 = \sum_{c=1}^{K} \alpha_c$.
If an exponential output function is used ($\alpha_c = e^{z_c}$), the the probability of label is the output of softmax:

$$P(\omega_c|\mathbf{x}^*, \hat{\theta}) = \frac{e^{z_c(\mathbf{x}^*)}}{\sum_{k=1}^{K} e^{z_k(\mathbf{x}^*)}}$$

Now let's consider loss function:

$$L(\theta) = \mathbb{E}_{p_{in}(x)}[KL[Dir(\mu|\hat{\alpha})||p(\mu|\mathbf{x}, \theta)]] + \mathbb{E}_{p_{out}(x)}[KL[Dir(\mu|\tilde{\alpha})||p(\mu|\mathbf{x}, \theta)]]$$

where $\hat{\alpha}$ - in-distribution targets, $\tilde{\alpha}$ - out-of-distribution targets.

Uncertainty Measures

1. Max probability:

$$P = max_c P(\omega_c|\mathbf{x}^*, D)$$

2. Entropy:

$$H[P(y|\mathbf{x}^*, D)] = -\sum_{c=1}^{K} P(\omega_c|\mathbf{x}^*, D)\ln(P(\omega_c|\mathbf{x}^*, D))$$

3. Mutual Information between y and θ (MI):

$$I[P(y, \theta|\mathbf{x}^*, D)] = H[E_{p(\theta|D)}P(y|\mathbf{x}^*, \theta)] - E_{p(\theta|D)}H[P(y|\mathbf{x}^*, \theta)]$$

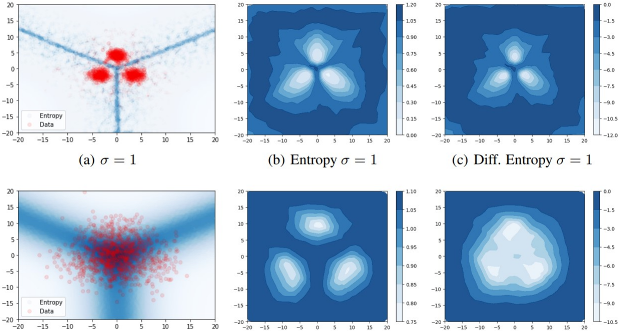
4. Mutual Information between y and μ (MI):

$$I[P(y, \mu|\mathbf{x}^*, D)] = H[E_{p(\mu|\mathbf{x}^*, D)}P(y|\mu)] - E_{p(\mu|\mathbf{x}^*, D)}H[P(y|\mu)]$$

5. Differential entropy:

$$H[p(\mu|\mathbf{x}^*, D)] = -\int_{S^{K-1}} p(\mu|\mathbf{x}^*, D)\ln p(\mu|\mathbf{x}^*, D)d\mu$$

Synthetic experiments



(a) $\sigma = 1$ (b) Entropy $\sigma = 1$ (c) Diff. Entropy $\sigma = 1$

(d) $\sigma = 4$ (e) Entropy $\sigma = 4$ (f) Diff. Entropy $\sigma = 4$

MNIST and CIFAR-10 experiments

Table 1: MNIST and CIFAR-10 misclassification detection

Data	Model	AUROC Max.P	AUROC Ent.	AUROC M.I.	AUROC D.Ent.	AUPR Max.P	AUPR Ent.	AUPR M.I.	AUPR D.Ent.	% Err.
MNIST	DNN	98.0	98.6	-	-	26.6	25.0	-	-	**0.4**
	MCDP	97.2	97.2	96.9	-	33.0	29.0	27.8	-	**0.4**
	DPN	**99.0**	98.9	98.6	92.9	**43.6**	39.7	30.7	25.5	0.6
CIFAR10	DNN	92.4	92.3	-	-	48.7	47.1	-	-	**8.0**
	MCDP	**92.5**	92.0	90.4	-	48.4	45.5	37.6	-	**8.0**
	DPN	92.2	92.1	92.1	90.9	**52.7**	**51.0**	**51.0**	45.5	8.5

Table 2: MNIST and CIFAR-10 out-of-domain detection

Data ID	Data OOD	Model	AUROC Max.P	AUROC Ent.	AUROC M.I.	AUROC D.Ent.	AUPR Max.P	AUPR Ent.	AUPR M.I.	AUPR D.Ent.
MNIST	OMNI	DNN	98.7	98.8	-	-	98.3	98.5	-	-
		MCDP	99.2	99.2	99.3	-	99.0	99.1	99.3	-
		DPN	**100.0**	**100.0**	99.5	**100.0**	**100.0**	**100.0**	97.5	**100.0**
CIFAR10	SVHN	DNN	90.1	90.8	-	-	84.6	85.1	-	-
		MCDP	89.6	90.6	83.7	-	84.1	84.8	73.1	-
		PN	98.1	98.2	98.2	**98.5**	97.7	97.8	97.8	**98.2**
CIFAR10	LSUN	DNN	89.8	91.4	-	-	87.0	90.0	-	-
		MCDP	89.1	90.9	89.3	-	86.5	89.6	86.4	-
		DPN	94.4	94.4	94.4	**94.6**	93.3	**93.4**	**93.4**	93.3
CIFAR10	TIM	DNN	87.5	88.7	-	-	84.7	87.2	-	-
		MCDP	87.6	89.2	86.9	-	85.1	87.9	83.2	-
		DPN	94.3	94.3	94.3	**94.6**	94.0	94.0	94.0	**94.2**