Notation

The training dataset $\mathcal{D} = \{x_j, y_j\}_{j=1}^N$

$p(\omega_c|x^*, \mathcal{D})$ denotes a shorthand for $p(y = \omega_c|x^*, \mathcal{D})$, the predictive distribution over label space given the input and training dataset.

Let $\mu = [p(y = \omega_0|x*), \ldots, p(y = \omega_K|x*)]$

Motivation

1. Estimating how uncertain an AI system is in its prediction is important to improve the safety of such systems.

2. Uncertain can come from 3 sources: model uncertainty, data uncertainty and distributional uncertainty. Identifying the sources of uncertainty is important as different actions can be taken depending on the sources of uncertainty.

Background Knowledge - Simplex

1. The $k+1$ points $\mu_0, \ldots, \mu_K \in R^{K+1}$ are affinely independent if $\mu_1 - \mu_0, \ldots, \mu_K - \mu_0$ are linearly independent.

2. Suppose the $k+1$ points $\mu_0, \ldots, \mu_K \in R^{K+1}$ are affinely independent, the simplex determined by the $k+1$ points $\mu_0, \ldots, \mu_K$ is the set of points:

$$C = \{\sum_i^K \theta_i k_i | \sum_i^K \theta_i = 1 \text{ and } \theta_i \geq 0, \forall i\}$$

3. The probability simplex is the simplex formed from the $k+1$ standard unit vector.

$$\{x \in R^{K+1} | \sum_i^K x_i = 1 \text{ and } x_i \geq 0, \forall i\}$$

4. The vector of probabilities assigned to each possible label by the predictive distribution μ can be seen as a point in the probability simplex.

Background Knowledge - Dirichlet Distribution

1. The form of the Dirichlet distribution as used in the paper is:

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{c=1}^K \Gamma(\alpha_c)} \prod_{c=1}^K \mu_c^{\alpha_c - 1}, \quad \alpha_c > 0, \quad \alpha_0 = \sum_{c=1}^K \alpha_c$$

2. The parameter of a Dirichlet is a vector $alpha = [alpha_0, \ldots, \alpha_K]$.

3. α_0 is referred to as the precision of the distribution. The higher the numerical value of α_0 is, the sharper the distribution is.

4. The Dirichlet distribution is the conjugate prior of the categorical distribution.

Type of Uncertainty

1. Model uncertainty (epistemic uncertainty) measures the uncertainty in estimating the model parameter from the data. Model uncertainty is reducible as the size of the training data increases.

2. Data uncertainty (aleatoric uncertainty) is irreducible uncertainty which arises from the natural complexity of the data, such as class overlap, homoscedastic and heteroscedastic noise.

3. Distributional uncertainty arises due to mismatch between the training and test distributions.

Prior Networks

1. This work explicitly parameterizes a distributions over distributions on a simplex, i.e. the distribution $p(\mu|x^*, \theta)$. The desirable behaviors of this distribution is as followed.

2. When the model is confident is its prediction for x^*, $p(\mu|x^*, \theta)$ should output a sharp distribution centered on one of the corners of the simplex. This indicates that there is one specific label that the model is certain is the correct label.

3. When there exists high level of data uncertainty for x^*, the model should output a sharp distribution focused on the center of the simplex. This indicates that the model has seen many other inputs from the same distribution as x^*. However, the correct label for those other inputs are inherently noisy. Thus, more than one labels have high probability of being the correct model under the ground truth predictive distribution.

4. When there exists a high level of distributional uncertainty for x^*, the model should output a flat distribution over the simplex. This indicates that the model knows it has not seen many inputs from the same distribution as x^*. Thus, absent any information to make a well-informed prediction, the safest prediction is to consider all members of the simplex occurring with equal probability. This is referred to as the principle of insufficient reason.

5. Prior Networks seeks to model the three different sources of uncertainty separately. $p(\omega_c|\mu)$ models data uncertainty, $p(\mu|x^*, \mathcal{D})$ models distributional uncertainty, $p(\theta|\mathcal{D})$ models model uncertainty.

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

6. The integral over $p(\theta|\mathcal{D})$ is estimated by a single sample, denoted by $\hat{\theta}$.

7. $p(\mu|x^*, \theta)$ takes the form of a Dirichlet distribution parameterized by a vector α where $\alpha = f(x^*; \theta)$. The Prior Network referred to in this work parameterizes this Dirichlet distribution.

8. The posterior distribution over the class labels is given by the mean of Dirichlet:

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

9. If an exponential is used to normalize the entries of α, i.e. $\alpha_c = e^{z_c}$, then the probability of a label as assigned by the posterior distribution is:

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

10. Thus, standard DNN for classification with softmax output function can be viewed as predicting the expected categorial distribution over a Dirichlet prior.

11. The mean of the Dirichlet is insensitive to arbitrary scaling of α. The precision α_0 is thus degenerate under cross entropy loss training.

12. However, we want to control for α_0, which controls how sharp the distribution over μ is. They thus change the training loss to explicitly train a DPN to yield a sharp or flat distribution around the expected categorical depending on the input data.

13. The loss function is:

$$\mathcal{L}(\boldsymbol{\theta}) = E_{\mathrm{p_{in}}(\boldsymbol{x})}[KL[\mathrm{Dir}(\boldsymbol{\mu}|\hat{\boldsymbol{\alpha}})\|\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x};\boldsymbol{\theta})]] + E_{\mathrm{p_{out}}(\boldsymbol{x})}[KL[\mathrm{Dir}(\boldsymbol{\mu}|\tilde{\boldsymbol{\alpha}})\|\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x};\boldsymbol{\theta})]]$$

14. $\tilde{\alpha} = 1, \forall c \neq 0$. For the out-distribution data, the DPN is trained to output the parameter of a flat Dirichlet distribution.

15. $\hat{\theta}_0$ is a hyper-parameter. For the remaining entries, $\hat{\mu}_c = \begin{cases} 1-(K-1)\epsilon & \text{if } \delta(y=\omega_c)=1 \\ \epsilon & \text{if } \delta(y=\omega_c)=0 \end{cases}$

16. For in-distribution data, the DPN is trained to output a sharp distribution centered at the corresponding class.

17. The loss function requires samples from the out-of-domain distribution. But, the true out-of-domain distribution is unknown.

Uncertainty measures

1. $\mathcal{P} = \max_c \mathrm{P}(\omega_c|\boldsymbol{x}^*;\mathcal{D})$

2. $\mathcal{H}[\mathrm{p}(y|\boldsymbol{x}^*;\mathcal{D})] = -\sum_{c=1}^{K} \mathrm{P}(\omega_c|\boldsymbol{x}^*;\mathcal{D}) \ln(\mathrm{P}(\omega_c|\boldsymbol{x}^*;\mathcal{D}))$ (The entropy of the predictive distribution)

3. Depending on what sources of uncertainties are marginalized out, different measures of uncertainties exist.

4. $\mathcal{H}[\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D})] = -\int_{\mathcal{S}^{K-1}} \mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D}) \ln(\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\mathcal{D}))d\boldsymbol{\mu}$ (differential entropy). The differential entropy is maximized when all categorial distribution is equiprobable. Differential entropy is well suited to measuring distributional uncertainty as it can be low even when the expected categorial under the Dirichlet prior has high entropy.

5. No marginalization. Use the full equation $p(\omega_c|x^*,\mathcal{D}) = \int\int p(\omega_c|\mu)p(\mu|x^*,\theta)p(\theta|\mathcal{D})d\mu d\theta$.

Experiment - synthetic data

1. A synthetic experiment was designed to illustrate the limitation of using uncertainty measures derived from $\mathrm{P}(\omega_c|\boldsymbol{x}^*;\mathcal{D})$ to detect out-of-distribution sample.

2. The main difference between the entropy and the differential entropy shows up when there is a large degree of class overlap.

3. The entropy is high in both region of class overlap and far from training data, making it difficult to distinguish between out-of-distribution samples and in-distribution samples at the decision boundary.

4. In contrast, the differential entropy is low over the whole region of training data and high outside, allowing the in-distribution region to be clearly distinguished from the out-of-domain region.

Experiments - misclassification detection

1. It is unclear how exactly the experiments were run here. Will need to read code to see what exactly was done.

Experiments - out-of-distribution detection

1. It is unclear how exactly the experiments were run here. Will need to read code to see what exactly was done.

2. On the dataset they tested on, there was little benefit in using the differential entropy. They hypothesize that this is because the inputs for the different classes are well-separated. When they added noise to the input to create artificial class overlap, differential entropy shows significant advantages compared to the entropy and mutual information measure.

Comments

1. What exactly is the input and output of the Prior Network? It seems that the Prior Network parameterizes f. Then, how does its output depend on $\hat{\theta}$?

2. If the precision α_0 controls how sharp the distribution $p(\mu|x^*,\theta)$ is and is set as a hyper-parameter, instead of being learnt, then they're fixing the degree of distributional uncertainty inherent in the model, aren't they?