Posterior Regularization

1 Introduction

One of the key challenges in probabilistic structured learning, is the intractability of the posterior distribution, for fast inference. There are numerous methods proposed to approximate the posterior, so as to make it easy to work with. Posterior Regularization is one of the proposed methods to approximate joint distribution between a set of structured variables, and taking the constrains into account. In [1] there is a comprehensive review of the method and previous ideas is mentioned.

2 Modelling the problem

In Posterior Regularization, we are dealing with a doing inference over posterior probability, with considering constraints, as indirect supervision. In this context we define $\mathbf{X}$ as input observation variables, and $\mathbf{Y}$ as latent variables, which we want to make predictions about. One example problem could be POS tagging as an example of structured prediction, in which we see input sentences $\mathbf{x}$, and output tags $\mathbf{y}$. We assume having generative model defined using $p(\mathbf{Y})$ as prior knowledge on latent variables, likelihood to be $p(\mathbf{X}|\mathbf{Y})$, and marginal-likelihood or evidence to be $\mathcal{L}(\theta) = \ln p(\mathbf{X};\theta) = \ln \int_{\mathbf{Y}\in\mathcal{Y}} p(\mathbf{Y})p(\mathbf{X}|\mathbf{Y})d\mathbf{Y}$. However one can use this probabilistic modelling to learn in discriminative way.

3 Approximating the posterior

Now we want to use $q(\mathbf{Y})$ to approximate the real intractable posterior, $p(\mathbf{Y}|\mathbf{X};\theta)$. In fact, $q(\mathbf{Y})$ is a simpler parametric distribution with parameter γ which is easy to work with [1]. To approximate the true posterior $p(.)$ using the decomposed distribution $q(.)$ we should find a measure of distance

[1] Thus the right way is to denote it by $q(\mathbf{Y};\gamma)$, but it is common to drop the parameters.

between two functions, and also is practical in computational sense. One of the famous distance measure between functions is called Kullback-Leibler divergence or in short, KL-*divergence*.

Lemma 1. *We have*

$$\mathcal{L}(\theta) = \log p(\mathbf{X};\theta) = \mathcal{J}(q,\theta) + KL\left(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X};\theta)\right). \tag{1}$$

in which

$$\mathcal{J}(q,\theta) \stackrel{\triangle}{=} \int q(\mathbf{Y}) \log \frac{p(\mathbf{X},\mathbf{Y};\theta)}{q(\mathbf{Y})} d\mathbf{Y} \tag{2}$$

$$KL\left(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X};\theta)\right) \stackrel{\triangle}{=} \mathbb{E}_q\left[\log \frac{q(\mathbf{Y})}{p(\mathbf{Y}|\mathbf{X};\theta)}\right]$$

Proof. Now the Bayes rule we have,

$$\log p(\mathbf{X};\theta) = \log p(\mathbf{X},\mathbf{Y};\theta) - \log p(\mathbf{Y}|\mathbf{X};\theta)$$
$$\log p(\mathbf{X};\theta) = \log \frac{p(\mathbf{X},\mathbf{Y};\theta)}{q(\mathbf{Y})} - \log \frac{p(\mathbf{Y}|\mathbf{X};\theta)}{q(\mathbf{Y})}.$$

Multiplying two sides in $q(\mathbf{Y})$ we have,

$$q(\mathbf{Y}) \log p(\mathbf{X};\theta) = q(\mathbf{Y}) \log \frac{p(\mathbf{X},\mathbf{Y};\theta)}{q(\mathbf{Y})} - q(\mathbf{Y}) \log \frac{p(\mathbf{Y}|\mathbf{X};\theta)}{q(\mathbf{Y})}$$
$$\int q(\mathbf{Y}) \log p(\mathbf{X};\theta) d\mathbf{Y} = \int q(\mathbf{Y}) \log \frac{p(\mathbf{X},\mathbf{Y};\theta)}{q(\mathbf{Y})} d\mathbf{Y} - \int q(\mathbf{Y}) \log \frac{p(\mathbf{Y}|\mathbf{X};\theta)}{q(\mathbf{Y})} d\mathbf{Y}.$$

Note that in the left part of the above equation $p_{\mathbf{X}}(\mathbf{x})$ is not a function of $\mathbf{Y}$ and thus, $\int q(\mathbf{Y}) \log p_{\mathbf{X}}(\mathbf{X}|\theta) d\mathbf{Y} = \log p(\mathbf{X}|\theta)$. Also note that,

$$\begin{aligned}
\mathrm{KL}\left(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X};\theta)\right) &\stackrel{\triangle}{=} -\int q(\mathbf{Y}) \log \frac{p(\mathbf{Y}|\mathbf{X};\theta)}{q(\mathbf{Y})} d\mathbf{Y} \\
&= \int q(\mathbf{Y}) \log \frac{q(\mathbf{Y})}{p(\mathbf{Y}|\mathbf{X};\theta)} d\mathbf{Y} \\
&= \mathbb{E}_q\left[\log \frac{q(\mathbf{Y})}{p(\mathbf{Y}|\mathbf{X};\theta)}\right] \\
\Rightarrow \mathcal{L}(\theta) &= \mathcal{J}(q,\theta) + \mathrm{KL}\left(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X};\theta)\right).
\end{aligned}$$

□

Lemma 2. *For any given observation and latent variables we have the following inequality:*

$$\mathcal{L}(\theta) \geq \mathcal{J}(q, \theta)$$

Proof. I use Jensen's inequality here to show that $\mathcal{J}(q, \theta)$ a lower bound for the original likelihood:

$$\begin{aligned}
\mathcal{L}(\theta) = \ln p(\mathbf{X}; \theta) &= \ln \int p(\mathbf{X}, \mathbf{Y}; \theta) d\mathbf{Y} \\
&= \ln \int q(\mathbf{Y}) \frac{p(\mathbf{X}, \mathbf{Y}; \theta)}{q(\mathbf{Y})} d\mathbf{Y} \\
&\geq \int q(\mathbf{Y}) \ln \frac{p(\mathbf{X}, \mathbf{Y}; \theta)}{q(\mathbf{Y})} d\mathbf{Y} = \mathcal{J}(q, \theta)
\end{aligned}$$

This shows that $\exp[\mathcal{J}(q, \theta)]$ a lower bound for likelihood $p(\mathbf{X}; \theta)$:

$$\Rightarrow p(\mathbf{X}|\theta) \geq \exp[\mathcal{J}(q, \theta)] \text{ or } \mathcal{L}(\theta) \geq \mathcal{J}(q, \theta).$$

□

Corollary 1. *Maximizing $\mathcal{J}(q, \theta)$ (lower bound) will result in maximizing the likelihood $p(\mathbf{X}; \theta)$. For this reason, $\mathcal{J}(q, \theta)$ is sometimes called **ELBO** (Evidence Lower Bound).*

4 Posterior Regularization

We define the following to be the objective function to be maximized:

$$\begin{cases} \mathcal{J}(\theta, q) = \mathcal{L}(\theta) - \mathrm{KL}(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X}; \theta)) \\ \mathbb{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \leq \mathbf{b} \end{cases} \tag{3}$$

Remark 1. *Having the above inequality lets us to define the the equation (1), since KL-divergence is always a positive value.*

We define a set of constraints in the output space as following:

$$\mathcal{Q} = \{q(\mathbf{Y}) | \mathbb{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \leq \mathbf{b}\}$$

Note that in the definition of the constraints $\mathbb{E}_q[\phi(\mathbf{X}, \mathbf{Y})]$ is a function of $\mathbf{X}$ which means that we define our constraints over input observations, to follow the structure defined by inequality and in the feature function $\phi(\mathbf{X}, \mathbf{Y})$. This definition can encode hard and soft constraints in itself.

5 Relaxed form

To cut some slack on objective function and the constraint function, one could define extra slack variables:

$$\begin{cases} \mathcal{J}(\theta, q) = \mathcal{L}(\theta) - \mathrm{KL}\left(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X};\theta)\right) + \sigma\|\xi\|_\beta \\ \mathbb{E}_q\left[\phi(\mathbf{X},\mathbf{Y})\right] - \mathbf{b} \leq \xi \end{cases}$$

6 Learning and inference over the regularized posterior

Without considering the constraints, we can show that training consists of a two step procedure similar to EM algorithm:

1. Initialize θ, and variational parameters of $q(.)$.
2. Repeat until convergence
   (a) E-step: $q^{(t+1)} = \arg\max_q \mathcal{J}(q^{(t)}, \theta^{(t)})$
   (b) M-step: $\theta^{(t+1)} = \arg\max_\theta \mathcal{J}(q^{(t+1)}, \theta^{(t)})$

It is worthy to mention that minimizing the objective function is equivalent to maximizing the KL-*divergence*.

By considering the constraint, the above procedure is limited to those $q(.)$s that satisfy the constraint set $\mathcal{Q}$. One way to do so, it to constrain the "E-step" to $\mathcal{Q}$:

$$\text{E-step(modifiede): } q^{(t+1)} = \arg\max_{q\in\mathcal{Q}} \mathcal{J}(q^{(t)}, \theta^{(t)})$$

Instead of applying such a constraint which is hard to calculate, we can use the dual form of the "E-step" and the constraint inequality.

Theorem 1. *Consider the following problem:*

$$\begin{cases} \max_{q,\xi} KL\left(q(\mathbf{Y})||p(\mathbf{Y}|\mathbf{X};\theta)\right) \\ \mathbb{E}_q\left[\phi(\mathbf{X},\mathbf{Y})\right] - \mathbf{b} \leq \xi \end{cases} \tag{4}$$

With θ as constant parameters. The primal solution $q^*(\mathbf{Y})$ is unique since KL-divergence is strictly convex:

$$q^*(\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X};\theta)\exp\left\{-\lambda^*.\phi(\mathbf{X},\mathbf{Y})\right\}}{Z(\lambda^*)}$$

in which $Z(\lambda^*) = \int p(\mathbf{Y}|\mathbf{X};\theta)\exp\{-\lambda^*.\phi(\mathbf{X},\mathbf{Y})\}\,d\mathbf{Y}$ *is the normalizing function. The dual for the above program is as following:*

$$\max_{\lambda\geq 0}\left[-\mathbf{b}.\lambda - \ln Z(\lambda) - \epsilon\|\lambda\|_{\beta'}\right]$$

in which $\|.\|_\beta$ *is dual norm for* $\|.\|_{\beta'}$.

The proof can be found at [1].

7 Bibliographical notes

I used Jiayu Zhou's notes (`http://www.public.asu.edu/~jzhou29/slides/PR_Notes.pdf`) in writing this document.

References

[1] Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 99:2001–2049, 2010.