Papers Reading Notes

Yuanxun Zhang

Contents

1 Probabilistic Latent Semantic Analysis (Hofmann, 1999)

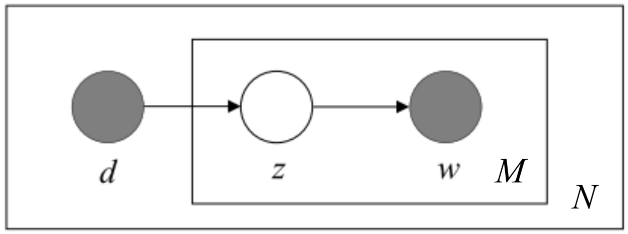

Figure 1: Graphical model representation of PLSA

As shown in Figure 1, the generative process in PLSA is as follows:

a) select a document d_i with probability $P(d_i)$

b) pick a latent class z_k with probability $P(z_k|d_i))$

c) generate a word w_j with probability $P(w_j|z_k))$

Then, the joint probability of PLSA model results in the expression,

$$P(d_i, w_j) = P(d_i)P(w_j|d_i), \qquad P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{1}$$

The modeling goal is to identify conditional probability mass functions $P(w_j|z_k)$. Formally, we can use a maximum likelihood formulation of the learning problem,

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_j) \tag{2}$$

Then, pluging Eq. 9 into Eq. 3, we got

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log\Big[P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)\Big] \\ &= \sum_{i=1}^{N} n(d_i)\log P(d_i) + \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\log \Big[\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)\Big] \end{aligned} \tag{3}$$

Where $n(d_i)$ denotes length of doc d_i, and $n(d_i, w_j)$ denotes the number of times the term w_j occurred in document d_i.

1.1 Inference with the EM Algorithm

Basically, to derive EM algorithm, we need to: a) define $Q(\theta, \theta^{(i)})$ function; b) In **E-step**, compute $Q(\theta, \theta^{(i)})$ function based on current parameter $\theta^{(i)}$; c) In **M-step**, re-estimate parameters $\theta^{(i+1)}$ which maximizes $Q(\theta, \theta^{(i)})$,

$$\theta^{(i+1)} = \arg\max_{\theta}(\theta, \theta^{(i)}) \tag{4}$$

1.1.1 Define Q function

The Q function is defined as the expectation of the complete-data log likelihood function $\log P(Y, Z|\theta)$ with respect of the posterior distribution of unobserved latent variables $P(Z|Y, \theta^{(i)})$, which is,

$$Q(\theta, \theta^{(i)}) = E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \tag{5}$$

Hence, in PLSA model, the complete-data log likelihood function will be Eq. 3. Because, in Eq. 3 the first term $\sum_{i=1}^{N} n(d_i)\log P(d_i)$ does not depends on latent variables z, we can ignore it. Then, the Q function can be defined as,

$$Q(\theta, \theta^{(i)}) = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} P(z_k|d_i, w_j)\log \left[P(w_j|z_k)P(z_k|d_i)\right] \tag{6}$$

1.1.2 E-Step

To compute $Q(\theta, \theta^{(i)})$, we just need to compute $P(z_k|d_i, w_j)$, which can be computed using Bayes rule,

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{P(d_i, w_j)} \tag{7}$$

$$= \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_k P(w_j|z_k)P(z_k|d_i)} \tag{8}$$

1.1.3 M-Step

In M-Step, we're going to find parameter $\theta^{(i+1)}$ that can maximize function Q. Because,

$$\sum_{j=1}^{M} P(w_j|z_k) = 1, \qquad \sum_{k=1}^{K} P(z_k|d_i) = 1 \tag{9}$$

So, the function $\mathcal{H}$ with Lagrange multipliers τ_k and ρ_i is,

$$\mathcal{H} = Q(\theta, \theta^{(i)}) + \sum_{k=1}^{K} \tau_k(1 - \sum_{j=1}^{M} P(w_j|z_k)) + \sum_{i=1}^{N} \rho_i(1 - \sum_{k=1}^{K} P(z_k|d_i)) \tag{10}$$

Then, first compute partial derivative of the function $\mathcal{H}$ with respect to the $P(w_j|z_k)$ and solve it when derivative is equal to zero.

$$\sum_{i=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)\frac{1}{P(w_j|z_k)} - \tau_k = 0 \tag{11}$$

or,

$$\sum_{i=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j) - \tau_k P(w_j|z_k) = 0 \tag{12}$$

the τ_k can be solved when combining $1 \leqslant j \leqslant M$,

$$\tau_k = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) P(z_k | d_i, w_j) \tag{13}$$

So, the $P(w_j | z_k)$ is

$$P(w_j | z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) P(z_k | d_i, w_j)} \tag{14}$$

Second, compute partial derivative of the function $\mathcal{H}$ with respect to the $P(z_k | d_i)$ and solve it when derivative is equal to zero.

$$\sum_{j=1}^{M} n(d_i, w_j) P(z_k | d_i, w_j) - \rho_i P(z_k | d_i) = 0 \tag{15}$$

And the ρ_i can be solved when combining $1 \leqslant k \leqslant K$, and $\sum_{k=1}^{K} P(z_k | d_i, w_j) = 1$,

$$\rho_i = \sum_{j=1}^{M} n(d_i, w_j) \tag{16}$$

So, the $P(z_k | d_i)$ is

$$P(z_k | d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{j=1}^{M} n(d_i, w_j)} \tag{17}$$

2 Finding scientific topics (Griffiths and Steyvers, 2004)

2.1 Derive the Eq.1 $P(\mathbf{w} | \mathbf{z})$

In paper, author use vector representation for $\mathbf{z}$ in Eq.1. For simplicity, I just use single topic assignment z_i instead of vector, and ϕ is multinomial distributions over the W

words for topic assignment z_i,

$$
\begin{aligned}
P(\mathbf{w}|z_i) &= \int P(\mathbf{w}, \phi|z_i)\ d\phi \\
&= \int P(\mathbf{w}|\mathbf{z}, \phi)P(\phi)\ d\phi \\
&= \int \frac{\Gamma(\sum_{i=1}^{W}\beta)}{\prod_{i=1}^{W}\Gamma(\beta)} \prod_{i=1}^{W} \phi_i^{\beta-1} \times \prod_{i=1}^{W} \phi_i^{n_{z_i}^{(w)}}\ d\phi \\
&= \int \frac{\Gamma(\sum_{i=1}^{W}\beta)}{\prod_{i=1}^{W}\Gamma(\beta)} \prod_{i=1}^{W} \phi_i^{n_{z_i}^{(w)}+\beta-1}\ d\phi \\
&= \frac{\Gamma(\sum_{i=1}^{W}\beta)}{\prod_{i=1}^{W}\Gamma(\beta)} \int \prod_{i=1}^{W} \phi_i^{n_{z_i}^{(w)}+\beta-1}\ d\phi \\
&= \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \int \prod_{i=1}^{W} \phi_i^{n_{z_i}^{(w)}+\beta-1}\ d\phi
\end{aligned}
\tag{18}
$$

in which $n_{z_i}^{(w)}$ is the number of times word w has been assigned to topic z_i. Because,

$$
\begin{aligned}
\int \prod_{i=1}^{W} \phi_i^{n_{z_i}^{(w)}+\beta-1}\ d\phi &= \frac{\prod_{i=1}^{W}\Gamma(n_{z_i}^{(w)}+\beta)}{\Gamma(\sum_{i=1}^{W} n_{z_i}^{(w)} + W\beta)} \\
&= \frac{\prod_{i=1}^{W}\Gamma(n_{z_i}^{(w)}+\beta)}{\Gamma(n_{z_i}^{(\cdot)} + W\beta)}
\end{aligned}
\tag{19}
$$

Then, the Equation 18 can be written as,

$$
P(\mathbf{w}|z_i) = \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \frac{\prod_{i=1}^{W}\Gamma(n_{z_i}^{(w)}+\beta)}{\Gamma(n_{z_i}^{(\cdot)} + W\beta)} \tag{20}
$$

When considering the whole T topic assignment $\mathbf{z}$, we get the same equation as shown in paper Eq.1.

$$
\begin{aligned}
P(\mathbf{w}|\mathbf{z}) &= \prod_{j=1}^{T} p(\mathbf{w}|z_j) \\
&= \Big(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\Big)^T \prod_{j=1}^{T} \frac{\prod_{i=1}^{W}\Gamma(n_j^{(w)}+\beta)}{\Gamma(n_j^{(\cdot)} + W\beta)}
\end{aligned}
\tag{21}
$$

In order to avoid numerical overflow,

2.2 Derive the Eq.5 $P(z_i = j|\mathbf{z}_{-i}, \mathbf{w})$

Because,

$$
P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})} \tag{22}
$$

Then,

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{P(\mathbf{w}, \mathbf{z}_{-i})} \tag{23}$$
$$= \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z})}{P(\mathbf{w}|\mathbf{z}_{-i})P(\mathbf{z}_{-i})}$$

So, we can put Eq.2 and Eq.3 of the original paper into Equation 23, and use Gamma function property $\Gamma(x+1) = x\Gamma(x)$ by cancellation of terms then,

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) = \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z})}{P(\mathbf{w}|\mathbf{z}_{-i})P(\mathbf{z}_{-i})} \tag{24}$$
$$= \frac{\left[\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^T \frac{\prod_{i=1}^W \Gamma(n_j^{(w)}+\beta)}{\Gamma(n_j^{(\cdot)}+W\beta)}\right] \times \left[\left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^D \frac{\prod_{j=1}^T \Gamma(n_j^{(d)}+\alpha)}{\Gamma(n_{\cdot}^{(d)}+T\alpha)}\right]}{\left[\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^T \frac{\prod_{i=1}^W \Gamma(n_{-i,j}^{(w)}+\beta)}{\Gamma(n_{-i,j}^{(\cdot)}+W\beta)}\right] \times \left[\left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^D \frac{\prod_{j=1}^T \Gamma(n_{-i,j}^{(d)}+\alpha)}{\Gamma(n_{-i,\cdot}^{(d)}+T\alpha)}\right]}$$
$$= \frac{n_{-i,j}^{(wi)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(di)} + \alpha}{n_{-i,\cdot}^{(di)} + T\alpha}$$

2.3 Model selection for computing $P(\mathbf{w}|T)$

In paper, author approximate $P(\mathbf{w}|T)$ by taking the harmonic mean of a set of values of $P(\mathbf{w}|\mathbf{z}^{(i)}, T)$ when $\mathbf{z}^{(i)}$ is sampled from the posterior $P(\mathbf{z}|\mathbf{w}, T)$, which means,

$$P(\mathbf{w}|T) \approx \left\{\frac{1}{m}\sum_{i=1}^{m} P(\mathbf{w}|\mathbf{z}^{(i)}, T)^{-1}\right\}^{-1} \tag{25}$$

Raftery *et al.* in papers (Newton and Raftery, 1994; Kass and Raftery, 1995) explain this idea by using the concept of importance sampling for model section.

In this example, we have several models $\{T_i : i = 10, 20, ..., 1000\}$, then Bayesian inference needs to compute the posterior probabilities given data $\mathbf{w}$,

$$P(T_i|\mathbf{w}) = \frac{P(\mathbf{w}|T_i)P(T_i)}{\sum_i^T P(\mathbf{w}|T_i)p(T_i)} \tag{26}$$

And the likelihood function $P(\mathbf{w}|T_i)$ is crucial component that needs to integrate out all topic assignment $\mathbf{z}$ then,

$$P(\mathbf{w}|T_i) = \int P(\mathbf{w}|\mathbf{z}, T_i)P(\mathbf{z}|T_i)\, d\mathbf{z} \tag{27}$$

So, the problem becomes how to approximate $P(\mathbf{w}|T_i)$.

Recall the basic Monte Carlo integration is to approximate $p(x) = \int p(x|\theta)p(\theta)d\theta$, when $p(\theta)$ is hard to integrate and the simple Monte Carlo approximation method is

$$\hat{I} = \frac{1}{m}\sum_{i=1}^{m} p(x|\theta^{(i)})) \tag{28}$$

However, the weakness of this simple method is that the estimation is dominated by a few large values of the small likelihood.

Another method (called importance sampling) is to generate samples $\{\theta^{(i)} : i = 1, ..., m\}$ from a proposal density function $q(\theta)$, and compute importance weight $w_i = \frac{p(\theta)}{q(\theta)}$. Then, the approximation is written as,

$$\hat{I} = \frac{\sum_{i=1}^{m} w_i p(x|\theta^{(i)})}{\sum_{i=1}^{m} w_i} \tag{29}$$

which is also known as importance sampling without normalization constants. Raftery *et al.* mentioned in papers (Newton and Raftery, 1994) that $q(\theta)$ can be approximately drawn from the their posterior density,

$$q(\theta) \approx p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{30}$$

Subsistution into Equation 29 yields, an an estimate for $p(x)$,

$$p(x) \approx \hat{p}(x) = \left\{ \frac{1}{m} \sum_{i=1}^{m} p(x|\theta^{(i)})^{-1} \right\}^{-1} \tag{31}$$

In this example, we need to approximate $P(\mathbf{w}|T_i)$, and we sample $\mathbf{z}^{(i)}$ from posterior distribution $P(\mathbf{w}|\mathbf{z}, T_i)$, then we got

$$P(\mathbf{w}|T_i) \approx \left\{ \frac{1}{m} \sum_{i=1}^{m} P(\mathbf{w}|\mathbf{z}^{(i)}, T_i)^{-1} \right\}^{-1} \tag{32}$$

3 On the importance of initialization and momentum in deep learning (Sutskever et al., 2013)

In this paper, authors mentioned two momentum-based optimization methods for deep learning: a). classical momentum (CM) and b). Nesterov's accelerated gradient (NAG).

3.1 Gradient Descent

The basic gradient descent is defined as,

$$\theta_{t+1} = \theta_t - \epsilon \nabla f(\theta_t) \tag{33}$$

where the θ is learning parameters or weights. and the ϵ is learning rate.

3.2 Classical Momentum (CM)

The CM method is defined as,

$$v_{t+1} = \mu v_t + (1-\mu)\nabla f(\theta_t) \tag{34}$$
$$\theta_{t+1} = \theta_t - \epsilon v_{t+1} \tag{35}$$

where ϵ is the learning rate, $\mu \in [0,1]$ is momentum coefficient. The notations are same with paper, but the equation is slightly different from the original paper.

3.3 Nesterov's Accelerated Gradient (NAG)

The NAG method is defined as,

$$v_{t+1} = \mu v_t + (1-\mu)\nabla f(\theta_t + \mu v_t) \tag{36}$$
$$\theta_{t+1} = \theta_t - \epsilon v_{t+1} \tag{37}$$

4 Auto-Encoding Variational Bayes (Kingma and Welling, 2013)

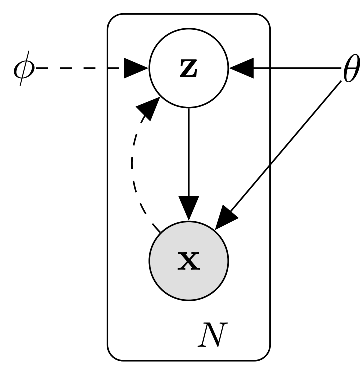

Figure 2: Variational inference of graphical model

4.1 The Variational Bound

Considering some dataset $X = \{x^{(i)}\}_{i=1}^N$ consisting of N i.i.d. samples of some continuous or discrete variable x. We assume that the data are generated by some random process, involving an unobserved continuous random variable z. Because, to compute $p(x)$ is intractable, that involves the integral of the marginal distribution $p(x) = \int p(z)p(x|z)dz$. Hence, to infer posterior density $p(z|x) = p(x|z)p(z)/p(x)$ is also intractable.

To solve this problem, authors introduce a recognition model $q(z|x)$ to approximate true posterior $p(z|x)$ and method to learn the recognition model parameters ϕ jointly with the generative model parameters θ. And the important definition of this paper is that they refer the recognition model $q_\phi(z|x)$ as a probabilistic ***encoder***, and generative model, and refer $p_\theta(x|z)$ as as a probabilistic ***decoder***. The Figure 3 illustrates the encoder-decoder framework in VAE.

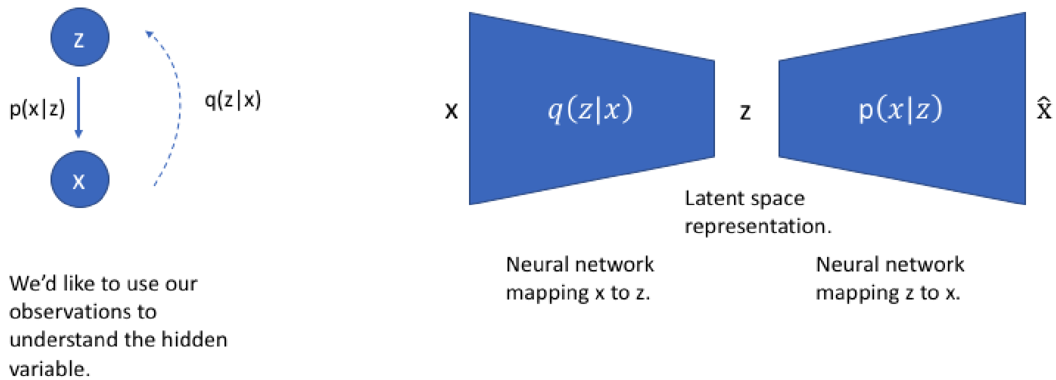

Figure 3: Encoder-Decoder framework in VAE

Then, this problem can be treated as optimization problem that is to minimize the divergence between $q_\phi(z|x)$ and $p_\theta(z|x)$, which is $D_{KL}(q_\phi(z|x)||p_\theta(z|x))$. Then,

$$\begin{aligned}
&D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \qquad (38)\\
&\quad= -\sum q_\phi(z|x)\log\frac{p_\theta(z|x)}{q_\phi(z|x)}\\
&\quad= -\sum q_\phi(z|x)\log p_\theta(z|x) + \sum q_\phi(z|x)\log q_\phi(z|x)\\
&\quad= -\sum q_\phi(z|x)\log\frac{p_\theta(x,z)}{p_\theta(x)} + \sum q_\phi(z|x)\log q_\phi(z|x)\\
&\quad= -\sum q_\phi(z|x)\log p_\theta(x,z) + \sum q_\phi(z|x)\log p_\theta(x) + \sum q_\phi(z|x)\log q_\phi(z|x)\\
&\quad= \sum q_\phi(z|x)\log p_\theta(x) - \sum q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\\
&\quad= \log p_\theta(x) - \sum q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}
\end{aligned}$$

Hence,

$$\log p_\theta(x) = D_{KL}(q_\phi(z|x)||p_\theta(z|x)) + \sum q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)} \qquad (39)$$

We define second term in Eq. 39 as variational lower bound $\mathcal{L}(\theta,\phi;x)$. Then,

$$\log p_\theta(x) = D_{KL}(q_\phi(z|x)||p_\theta(z|x)) + \mathcal{L}(\theta,\phi;x) \qquad (40)$$

And,

$$\begin{aligned}
\mathcal{L}(\theta,\phi;x) &= \sum q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)} \\
&= \sum q_\phi(z|x)\log\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)} \\
&= \sum q_\phi(z|x)\left[\log p_\theta(x|z) - \log\frac{p_\theta(z)}{q_\phi(z|x)}\right] \\
&= \sum q_\phi(z|x)\log p_\theta(x|z) - \sum q_\phi(z|x)\log\frac{p_\theta(z)}{q_\phi(z|x)} \\
&= -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]
\end{aligned} \tag{41}$$

References

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. doi: 10.1073/pnas.0307752101. URL `http://www.pnas.org/content/101/suppl_1/5228.abstract`.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `http://proceedings.mlr.press/v28/sutskever13.html`.