PRML chapter10

1. Variational Inference
   - aim : evaluate $p(\mathbf{Z}|\mathbf{X})$
   - optimize q(Z), means maximize L(q) or minimize KL$(q\|p)$
   - decompose the log marginal probability using

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p),$$

   under

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$KL(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

   - the parameters are now stochastic variables
   - consider a restricted family of distributions q(Z) and then seek the member of this family for which the KL divergence is minimized

   (a) Factorized distributions
   - Suppose that q(Z) can be decomposed as below

$$q(Z) = \prod_{i=1}^{M} q_i\left(Z_i\right)$$

   - maximize L(q)

$$\mathcal{L}(q) = \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} = \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{Z}_j + \mathrm{const}$$

$$= \int q_j \ln \widetilde{p}\left(\mathbf{X}, \mathbf{Z}_j\right) \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{Z}_j + \mathrm{const}$$

   - it is Kullback-Leibler divergence so we could gain a general expression for the optimal solution

$$\ln q_j^{\star}\left(\mathbf{Z}_j\right) = E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + const.$$

   (b) Properties of factorized approximations
   - consider the problem of approx- imating a general distribution by a factorized distribution.
   - show the difference between KL$(p\|q)$ and KL$(q\|p)$
   - The former avoids the region that p(Z) is low, otherwise the latter tries to cover the region that p(Z) is not zero.

(c) Example: The univariate Gaussian

- illustrate the factorized variational approximation using a Gaussian distribution over a single variable x
- likelihood is

$$p(D|\mu,\tau)=\left(\frac{\tau}{2\pi}\right)^{\frac{N}{2}}\exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- prior distribution is

$$p(\mu|\tau)=N(\mu|\mu_0,(\lambda_0\tau)^{-1})p(\tau)=Gam(\tau|a_0,b_0)$$

- we gain

$$\ln q_\mu^\star(\mu)=E_\tau[\ln p(\mathcal{D}|\mu,\tau)+\ln p(\mu|\tau)]+const=-\frac{E[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2+\sum_{n=1}^{N}(x_n-\mu)^2\right\}+const.$$

$$\ln q_\tau^\star(\tau)=E_\mu[\ln p(\mathcal{D}|\mu,\tau)+\ln p(\mu|\tau)]+\ln p(\tau)+\text{const}$$

$$=(a_0-1)\ln\tau-b_0\tau+\frac{N}{2}\ln\tau-\frac{\tau}{2}E_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2+\lambda_0(\mu-\mu_0)^2\right]+\text{const}$$

2. Illustration: Variational Mixture of Gaussians
   - later come back this session
3. Variational Linear Regression
   - $\alpha's$ prior distribution is gamma dustribution and

$$p(\mathbf{t}|\mathbf{w})=\prod_{n=1}^{N}\mathcal{N}\left(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n,\beta^{-1}\right),p(\mathbf{w}|\alpha)=\mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$$

(a) Variational distribution

- posterior distribution is expressed by the factorized epression

$$q(\mathbf{w},\alpha)=q(\mathbf{w})q(\alpha)$$

- of course we soon gain

$$\begin{aligned}\ln q^\star(\alpha)&=\ln p(\alpha)+E_{\mathbf{w}}[\ln p(\mathbf{w}|\alpha)]+const\\&=(a_0-1)\ln\alpha-b_0\alpha+\tfrac{M}{2}\ln\alpha-\tfrac{\alpha}{2}E\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]+\text{const.}\end{aligned}$$

- and

$$\ln q^\star(\mathbf{w})=\ln p(\mathbf{t}|\mathbf{w})+E_\alpha[\ln p(\mathbf{w}|\alpha)]+\text{const}$$

$$=-\frac{\beta}{2}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n-t_n\right\}^2-\frac{1}{2}E[\alpha]\mathbf{w}^{\mathrm{T}}\mathbf{w}+\text{const}$$

$$=-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(E[\alpha]\mathbf{I}+\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\mathbf{w}+\beta\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}+\text{const.}$$

- The evaluation of the variational posterior distribution begins by initializing the pa- rameters of one of the distributions q(w) or q(α), and then alternately re-estimates these factors in turn until a suitable convergence criterion is satisfied

(b) Predictive distribution

- predictive distribution is calculated easily by

$$p(t|\mathbf{x},\mathbf{t}) = \int p(t|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathbf{t})\mathrm{d}\mathbf{w} \simeq \int p(t|\mathbf{x},\mathbf{w})q(\mathbf{w})\mathrm{d}\mathbf{w}$$

(c) lower bound

- another important quantity

$$\mathcal{L}(q) = E[\ln p(\mathbf{w},\alpha,\mathbf{t})] - E[\ln q(\mathbf{w},\alpha)]$$
$$= E_{\mathbf{w}}[\ln p(\mathbf{t}|\mathbf{w})] + E_{\mathbf{w},\alpha}[\ln p(\mathbf{w}|\alpha)] + E_{\alpha}[\ln p(\alpha)] - E_{\alpha}[\ln q(\mathbf{w})]_{\mathbf{w}} - E[\ln q(\alpha)]$$

4. Exponential Family Distributions
   - make a further distinction between latent variables and parameters
   - let joint distribution be

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n,\mathbf{z}_n)\, g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\}$$

   - prior distribution of η is

$$p(\boldsymbol{\eta}|\nu_0,\mathbf{v}_0) = f(\nu_0,\chi_0)\, g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o\boldsymbol{\eta}^{\mathrm{T}}\chi_0\right\}$$

   - we gain solution as

$$\ln q^{\star}(\boldsymbol{\eta}) = \ln p(\boldsymbol{\eta}|\nu_0,\boldsymbol{\chi}_0) + E_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \mathrm{const}$$
$$\ln q^{\star}(\mathbf{Z}) = E_{\boldsymbol{\eta}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \mathrm{const}$$

   - they are dependent on each other so use EM algorithms

5. Local Variational Methods
   - approximate the convex function f (x) by a linear function y(x, λ)

$$f(x) = \max_{\lambda}\{\lambda x - \lambda + \lambda\ln(-\lambda)\}$$

   - more gerenaly

$$f(x) = \max_{\lambda}\{\lambda x - g(\lambda)\}$$
$$g(\lambda) = \max_{x}\{\lambda x - f(x)\}$$

6. Variational Logistic Regression

(a) Variational posterior distribution

- In Baysian logistic regression model,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} = \int \left[\prod_{n=1}^{N} p(t_n|\mathbf{w})\right] p(\mathbf{w})\mathrm{d}\mathbf{w}$$

- in last session, we talked about the variational lower bound on the logistic sigmoid function.

- use it for $p(t|\mathbf{w})$, then

$$p(t|\mathbf{w}) = e^{at}\sigma(-a) \geq e^{at}\sigma(\xi)\exp\left\{-(a+\xi)/2 - \lambda(\xi)\left(a^2-\xi^2\right)\right\}$$

- let ξ be a variational parameter,

$$p(\mathbf{t},\mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w})h(\mathbf{w},\boldsymbol{\xi})p(\mathbf{w})$$

where

$$h(\mathbf{w},\boldsymbol{\xi}) = \prod_{n=1}^{N}\sigma\left(\xi_n\right)\exp\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n - \left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \xi_n\right)/2 - \lambda\left(\xi_n\right)\left(\left[\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\right]^2 - \xi_n^2\right)\right\}$$

- Since log function is monotonically increasing

$$\ln\{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\} \geq \ln p(\mathbf{w}) + \sum_{n=1}^{N}\left\{\ln\sigma\left(\xi_n\right) + \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n - \left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \xi_n\right)/2 - \lambda\left(\xi_n\right)\left(\left[\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\right]^2 - \xi_n^2\right)\right.$$

- substitute $p(\mathbf{w})$, right side of the inequality becomes

$$-\frac{1}{2}\left(\mathbf{w}-\mathbf{m}_0\right)^{\mathrm{T}}\mathbf{S}_0^{-1}\left(\mathbf{w}-\mathbf{m}_0\right) + \sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n\left(t_n - 1/2\right) - \lambda\left(\xi_n\right)\mathbf{w}^{\mathrm{T}}\left(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\right)\mathbf{w}\right\} + \text{const.}$$

- this is quadratic function of $\mathbf{w}$,

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

- we shall use it shortly to evaluate the predictive distribution for new data points

(b) Optimizing the variational parameters

- how to estimate the variational parameter ξ
- Two ways to achieve this goal, one is to use EM algorithm the other is to integrate analytically and perform a direct maximization

7. Expectation Propagation

- we have a joint distribution

$$p(\mathcal{D},\boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$$

, and want to approximate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ by

$$q(\boldsymbol{\theta}) = \frac{1}{Z}\prod_i \widetilde{f}_i(\boldsymbol{\theta})$$

- we also want to approximate the model evidence $p(\mathcal{D})$
- First, we initialixe the approximating factor $\widetilde{f}_i(\theta)$
- and initialize posterior approximation $q(\boldsymbol{\theta}) \propto \prod_i \widetilde{f}_i(\boldsymbol{\theta})$
- Then, choose a factor $\widetilde{f}_j(\boldsymbol{\theta})$ that we want to refine
- remove it from the posterior distribution

$$q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\widetilde{f}_j(\boldsymbol{\theta})}$$

- calculate $q^{new}(\boldsymbol{\theta})$ by equaling to $q^{\backslash j}(\boldsymbol{\theta})f_j(\boldsymbol{\theta})$

- normalization constant is

$$Z_j = \int q^{\backslash j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

- evaluate the new factor

$$\widetilde{f}_j(\boldsymbol{\theta}) = Z_j \frac{q^{new}(\boldsymbol{\theta})}{q^{\downarrow j}(\boldsymbol{\theta})}$$

- evaluate the model evidence

$$p(\mathcal{D}) \simeq \int \prod_i \widetilde{f}_i(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$