Example of the variational approximation for the course Machine Learning: Advanced Probabilistic Methods (2015), P.Marttinen

Suppose that we have N independent observations $\mathbf{x} = (x_1, \ldots, x_N)$ from a two-component mixture of univariate Gaussian distributions

$$p(x_n|\theta) = (1-\tau)N(x_n|0,1) + \tau N(x_n|\theta,1), \tag{1}$$

that is, with probability $1-\tau$ the observation x_n is generated from the first component $N(x_n|0,1)$, and with probability τ from the second component $N(x_n|\theta,1)$. The model (1) has two unknown parameters, (τ, θ), the mixture coefficient and the mean of the second component.

Our goal is to carry out a fully Bayesian analysis using the mean-field variational Bayes approximation. We place the following priors on the unknown parameters

$$\begin{aligned} \tau &\sim Beta(\alpha_0, \alpha_0) \\ \theta &\sim N(0, \beta_0^{-1}). \end{aligned}$$

We formulate the model using latent variables $\mathbf{z} =(z_1, \ldots, z_N)$ which explicitly specify the component responsible for generating observation x_n. In detail,

$$z_n = (z_{n1}, z_{n2})^T = \begin{cases} (1,0)^T, & (x_n \text{ is from } N(x_n|0,1)) \\ (0,1)^T, & (x_n \text{ is from } N(x_n|\theta,1)) \end{cases},$$

and place a prior on the latent variables

$$p(\mathbf{z}|\tau) = \prod_{n=1}^{N} \tau^{z_{n2}}(1-\tau)^{z_{n1}}.$$

The likelihood in the latent variable model is given by

$$p(\mathbf{x}|\mathbf{z},\theta) = \prod_{n=1}^{N} N(x_n|0,1)^{z_{n1}} N(x_n|\theta,1)^{z_{n2}}.$$

The joint distribution of all observed ($\mathbf{x}$) and unobserved variables ($\mathbf{z}, \tau, \theta$) factorizes as follows

$$p(\mathbf{x}, \mathbf{z}, \tau, \theta) = p(\tau)p(\theta)p(\mathbf{z}|\tau)p(\mathbf{x}|\mathbf{z},\theta)$$

and the log of the joint distribution can correspondingly be written as

$$\log p(\mathbf{x}, \mathbf{z}, \tau, \theta) = \log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z},\theta).$$

We approximate the posterior distribution $p(\mathbf{z}, \tau, \theta|\mathbf{x})$ using the factorized variational distribution $q(\mathbf{z})q(\tau)q(\theta)$.

Update of factor $q(\mathbf{z})$

To compute the updated distribution $q^*(\mathbf{z})$, we first compute the expectation of the log of the joint distribution over all other unknowns in the model

$$\begin{aligned} \log q^*(\mathbf{z}) &= E_{\tau,\theta}[\log p(\mathbf{x}, \mathbf{z}, \tau, \theta)] \\ &= E_\tau[\log p(\mathbf{z}|\tau)] + E_\theta[\log p(\mathbf{x}|\mathbf{z},\theta)] + \text{const (not dependent on } \mathbf{z}) \\ &= E_\tau\left\{\sum_{n=1}^{N}[z_{n2}\log\tau + z_{n1}\log(1-\tau)]\right\} + E_\theta\left\{\sum_{n=1}^{N}[z_{n1}\log N(x_n|0,1) + z_{n2}\log N(x_n|\theta,1)]\right\} + \text{const} \\ &= \sum_{n=1}^{N}\{z_{n2}E_\tau[\log\tau] + z_{n1}E_\tau[\log(1-\tau)]\} + \sum_{n=1}^{N}\{z_{n1}\log N(x_n|0,1) + z_{n2}E_\theta[\log N(x_n|\theta,1)]\} + \text{const} \\ &= \sum_{n=1}^{N} z_{n1}\left\{E_\tau[\log(1-\tau)] - \frac{1}{2}\log(2\pi) - \frac{1}{2}x_n^2\right\} + \sum_{n=1}^{N} z_{n2}\left\{E_\tau[\log(\tau)] - \frac{1}{2}\log(2\pi) - \frac{1}{2}E_\theta\left[(x_n-\theta)^2\right]\right\} + \text{const} \\ &= \sum_{n=1}^{N}\{z_{n1}\log\rho_{n1} + z_{n2}\log\rho_{n2}\} + \text{const}, \end{aligned} \tag{2}$$

where we have defined variables ρ_{n1} and ρ_{n2} for all n as follows

$$\log \rho_{n1} = E_\tau \left[\log(1-\tau)\right] - \frac{1}{2}\log(2\pi) - \frac{1}{2}x_n^2 \quad \text{and} \tag{3}$$

$$\log \rho_{n2} = E_\tau \left[\log(\tau)\right] - \frac{1}{2}\log(2\pi) - \frac{1}{2}E_\theta\left[(x_n-\theta)^2\right]. \tag{4}$$

By exponentiating both sides of equation (2), we get

$$q^*(\mathbf{z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{2}\rho_{nk}^{z_{nk}},$$

which we can normalize to make a proper distribution

$$q^*(\mathbf{z}) = \prod_{n=1}^{N}\prod_{k=1}^{2}r_{nk}^{z_{nk}},$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{2}\rho_{nj}}. \tag{5}$$

Note that to compute the updated *responsibilities* r_{nk}, we need $E_\tau\left[\log(1-\tau)\right]$, $E_\tau\left[\log(\tau)\right]$, and $E_\theta\left[(x_n-\theta)^2\right]$, where the expectations are computed over the distributions $q(\tau)$ and $q(\theta)$, which will be derived next.

Update of factor $q(\tau)$

$$\begin{aligned}\log q^*(\tau) &= E_{\mathbf{z},\theta}[\log p(\mathbf{x},\mathbf{z},\tau,\theta)]\\ &= \log p(\tau) + E_{\mathbf{z}}\left[\log p(\mathbf{z}|\tau)\right] + \text{const (not dependent on } \tau)\\ &= \ldots (\textit{left as an exercise})\end{aligned}$$

We exponentiate and recognize the exponentiated form as,

$$q^*(\tau) = Beta(\tau|N_2+\alpha_0, N_1+\alpha_0),$$

i.e., τ has a $Beta(a,b)$ with parameters $a = N_2+\alpha_0$ and $b = N_1+\alpha_0$, where $N_k = \sum_{n=1}^{N} r_{nk}$ for $k=1,2$. Using this distribution, we get the following formulas for the terms required when updating $q(\mathbf{z})$

$$E_\tau\left[\log(\tau)\right] = \psi(N_2+\alpha_0) - \psi(N_1+N_2+2\alpha_0) \tag{6}$$

$$E_\tau[\log(1-\tau)] = \psi(N_1+\alpha_0) - \psi(N_1+N_2+2\alpha_0), \tag{7}$$

where ψ is the digamma function. Formulas (6) and (7) follow from the basic properties of the beta distribution (see e.g. Wikipedia) and by noticing that if $\tau \sim Beta(a,b)$, then $1-\tau \sim Beta(b,a)$.

Update of factor $q(\theta)$

$$\log q^*(\theta) = \ldots (\textit{left as an exercise}) \tag{8}$$

Again, we exponentiate both sides of (8) and recognize this as

$$q^*(\theta) = N\left(\theta|m_2, \beta_2^{-1}\right), \tag{9}$$

with

$$\beta_2 = \beta_0 + N_2 \quad \text{and} \quad m_2 = \beta_2^{-1}N_2\overline{x}_2,$$

where we have defined

$$\overline{x}_2 = \frac{1}{N_2}\sum_{n=1}^{N} r_{n2}x_n.$$

We can use the distribution (9) to compute the formula for $E_\theta\left[(x_n-\theta)^2\right]$, needed when updating $q(\mathbf{z})$:

$$\begin{aligned}E_\theta\left[(x_n-\theta)^2\right] &= E_\theta\left[(x_n-m_2+m_2-\theta)^2\right]\\ &= (x_n-m_2)^2 + 2(x_n-m_2)E\left[m_2-\theta\right] + E\left[(m_2-\theta)^2\right]\\ &= (x_n-m_2)^2 + 0 + \beta_2^{-1}.\end{aligned} \tag{10}$$

The last equality in (10) followed from the fact that when $\theta \sim N(m_2, \beta_2^{-1})$, then $m_2-\theta \sim N(0, \beta_2^{-1})$.

The **overall VB algorithm** is obtained by cycling through updating

1. the responsibilities r_{nk} using formulas (3), (4), and (5)
2. the terms (10) needed when computing the responsibilities
3. the terms (6) and (7) needed when computing the responsibilities

Code to run the EM-algorithm: *simple_ vb.m*