Variational Bayesian Inference Note

Keshi Dai

November 20, 2009

1 Motivation

When we use EM that uses maximum likelihood as a criterion to select the number of Gaussians, we face the problem of that as the complexity of model increases, the training likihood strictly improves, which means the larger number of Gaussians, the better fit of the training data (see Figure 1). We can see from this example, the traning log-likelihood can even become positive when some clusters has only few data, because log-likehood is the only critertion for selecting the number of clusters, when the variance in a cluster decrease, the likelihood goes up. Take an extreme example, if there is only one data point in the cluster, the log-likelyhood of training data will be infinite. This is obviously over-fitting and incorrect.

Of course, cross-validation-like strategy can fix this kind of overfitting problem, which requires a large amount of data. However, can we find a model that is not only the most likely one to generate all data we have seen, but also takes the model complexity into account? Can we define a model without restriction of the number of clusters but penalize the use of more clusters. Yes, we can! Bayesian treatment can solve this over-fitting problem by imposing certain prior distributions on parameters of the model that gives lower probability to more complex model. Now, we maximize the posterior instead of the likelihood, which can be seen as a regularization of the likelihood.

Assume we want to estimate an unobserved model parameter θ based on observations x, so the likelihood function is the probability of x given θ. According to the maximum likelihood criterion, we obtain the estimated θ by setting the derivative of Equation 1 to zero.

$$\widehat{\theta_{\mathrm{ML}}} = \arg\max_{\theta} p(x \mid \theta) \tag{1}$$

Now assume we have a prior distribution for parameter θ. Then the posterior distribution of θ is:

$$p(\theta|x) = \frac{f(x \mid \theta)g(\theta)}{p(x)} = \frac{f(x \mid \theta)g(\theta)}{\int f(x|\theta')g(\theta')d\theta'} \tag{2}$$

Hence, we can get the estimated θ by maximizing the posterior distribution:

$$\widehat{\theta_{\mathrm{MAP}}} = \arg\max_{\theta} \frac{f(x \mid \theta)g(\theta)}{f(x)} = \arg\max_{\theta} f(x \mid \theta)g(\theta) \tag{3}$$

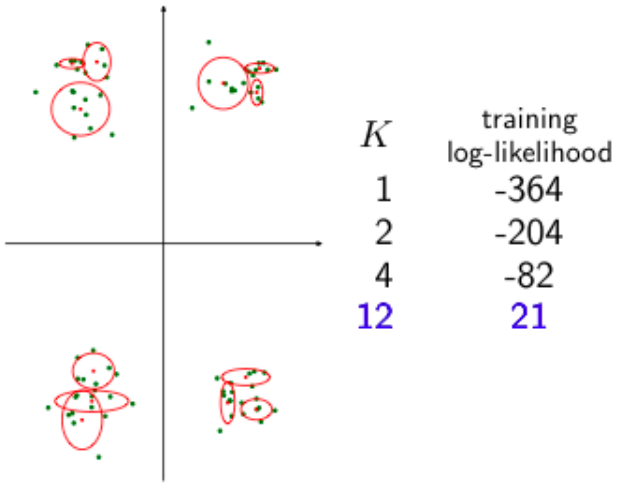

Figure 1: A motivating example

Because $f(x)$ does not depend on θ, we can ingore it in the maximizing process. Obsever Equation 1 and Equation 3, we can see that maximizing likelihood is actually a special case of maximizing posterier, whose prior is a uniform distribution, so $p(\theta)$ is constant.

There are some substantial advantages of using the Bayesian approach: (1) The sigularit problem in EM will not happen in the Bayesian treatment. (2) Over-fitting problem by selecting a large number of K is also solved. (3) The variational treatment provides a way to determine the optimal number of components in the mixture without resorting to cross validation.

2 Methodology

After imposing certain priors on the likelyhood function, the opmization problem can not be solved by setting the derivative of the posterior to zero because of the the complexity of the prior distribution. Approximation algorithms such as Markov chain Monte Carlo or variational inference are needed to help us achieve a close solution to the problem. Variational inference is an analytical approximation technique that finds a distribution closest to the true posterior distribution in terms of KL distance.

We know the log marginal probability can be decomposed as:

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{4}$$

where,

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \tag{5}$$

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \tag{6}$$

Because KL divergence is always greater than zero, $\ln p(\mathbf{X})$ is the lower bound of the log marginal probability of $\mathbf{X}$. To optimize Equation 4, we can maximize the lower bond

$\ln p(\mathbf{X})$, which is equivalent to minimizing the KL divergence. The KL divergence vanishes when $q(\mathbf{Z})$ equals to the true posterier distribution $p(\mathbf{Z}|\mathbf{X})$. However the true posterier is intractable, so we factorize $q(\mathbf{Z})$ and seek an approximation of $q(\mathbf{Z})$ from a retricted family of tractable distributions, which minimizes the KL divergence.

2.1 Factorized distributions

Let us look at a general case. Suppose we partition all elments $\mathbf{Z}$ into disjoint groups $\mathbf{Z}_i$, where $i = 1, \ldots, M$. We assume $q(\mathbf{Z})$ factorizes with respect to these groups, so that

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i) \tag{7}$$

Now we are looking for a distribution that maximize the lower bound $\mathcal{L}(q)$ from all distributions $q(\mathbf{Z})$ having the form (Equation 17). First, we subtitute Equation 17 into Equation 5, we obtain (denoting $q_j(\mathbf{Z}_j)$ by q_j)

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod q_i \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z_i} \right\} d\mathbf{Z_j} - \int q_j \ln q_j d\mathbf{Z_j} + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z_j} + \text{const}
\end{aligned} \tag{8}$$

where we define $\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ as follows:

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const} \tag{9}$$

Here $\mathbb{E}_{i \neq j}[\cdots]$ denotes an expectation with respect to the q distributions over all variables $\mathbf{z}_i$ for $i \neq j$, so that

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \tag{10}$$

In fact, the first part of $\mathcal{L}(q)$ is a negtive KL divergence between $q_j(\mathbf{Z}_j)$ and $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$. Hence, maximizing Equation 8 is equivalent to minimizing the KL divergergve, and minmum occurs when $q_j(\mathbf{Z}_j) = \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$. We obtain a general expression for the optimal solution given by:

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const} \tag{11}$$

where the constant part is from the normalization of q distribution. Then we have,

$$q_j^*(\mathbf{Z}_j) = \frac{exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})])}{\int exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]) d\mathbf{Z}_j} \tag{12}$$

2.2 Example: Mixture of Gaussians

For each observation $\mathbf{x}_n$ we have a corresponding latent variable $\mathbf{z}_n$ comprising a 1-of-K binary vector with elements z_{nk} for $k = 1, \ldots, K$. We denote the latent variables by $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$. We can write down the conditional distribution of $\mathbf{Z}$, given the mixing coefficients π, in the form

$$p(\mathbf{Z}|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{Z_{nk}} \tag{13}$$

Similarly, the conditional distribution of the observed data given the latent variables and the component parameters

$$p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\mu_k, \Lambda_k^{-1})_{nk}^{z} \tag{14}$$

According to the conjugate prior distributions, we choose a Dirichlet distribution over the mixing coefficients π,

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{15}$$

and an independent Gaussian-Wishart prior governing the mean and precison of each Gaussian component,

$$\begin{aligned} p(\mu, \Lambda) &= p(\mu|\Lambda)p(\Lambda) \\ &= \prod_{k=1}^{K} \mathcal{N}(\mu_k|\mathbf{m}_0, (\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|\mathbf{W}_0, v_0) \end{aligned} \tag{16}$$

Now we factorize the joint distribution of parameters so that

$$p(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda) \tag{17}$$

Based on the previous section, factors $q(\mathbf{Z})$ and $q(\pi, \mu, \Lambda)$ will be determined automatically by optimization of the variational distribution. According to the result given by Equation 11, we have

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda)] + \text{const} \tag{18}$$

The joint distribution of all random variables can also be also written as

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda})p(\mathbf{Z}|\pi)p(\pi)p(\mu|\mathbf{\Lambda})p(\mathbf{\Lambda}) \tag{19}$$

Substituting Equation 19 into Equation 18 results in (Note: we only keep parts that are dependent on the variable Z),

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\pi}[\ln p(\mathbf{Z}|\pi)] + \mathbb{E}_{\mu,\Lambda}[\ln p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda})] + \text{const} \tag{20}$$

Substituting two conditional distribution Equation 13 and Equation 14, we have

$$\ln q^*(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const} \tag{21}$$

where we have defined

$$\begin{aligned} \ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\mathbf{\Lambda}_k|] - \frac{D}{2}\ln(2\pi) \\ -\frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}[(\mathbf{x_n} - \mu_\mathbf{k})^\mathbf{T}\mathbf{\Lambda_k}(\mathbf{x_n} - \mu_\mathbf{k})] \end{aligned} \tag{22}$$

where D is the dimensionality of the data variable $\mathbf{x}$. Taking the exponential of both sides of Equation 21 and normalizing it, we obtain

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \tag{23}$$

Let us define three statistics of the observed data set evaluated with respect to the responsibilities, given by

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{24}$$

$$\overline{\mathbf{x}_k} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n \tag{25}$$

$$\mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^T \tag{26}$$

3 Experiment

1. Set initial number of Gaussian to 10
2. Make a random assignment of points to 10 Gaussians
3. Run k-means to initialize the parameter for EM
4. Run EM to get intial parameter for prior
5. Set the parameter α of Dirichlet distribution to a small value (0.001), which lets the observed data dictate the posterier distribution.
6. Set m_0 to the average of scores of relevant documents
7. Set β_0 to 10, v to 10, and $\mathbf{W}$ to a indentity matrix with a scalar of 10 for Gaussian-Whishart distribution.
8. Learn the number of Gaussians and other parameters through Variational Bayesian Inference, meanwhile get the log-likelihood of the fitting.
9. Repeat step 2-3 for 10 times, and select the model giving the largest log-likelihood.