Machine Learning

Zhenrong
May 2, 2021

1 Introduction

1.1 Probability Theory

1.1.1 Bayesian and Frequentist View

In the interpretation of frequentist, the model parameter w is considered as a fixed value which was estimated by different kind of estimators. In Bayesian view, the model parameter w is an uncertain value, represented by a prior probability $p(w)$ or a posterior probability $p(w|D)$. Before the data set has been observed, the uncertainty of w is represented by a prior probability $p(w)$. Once the data set has been observed, Bayesian approach adjust the probability of w by incorporating the evidence provided by the observed data, namely turning the a prior probability $p(w)$ into the a posterior probability $p(w|D)$. The key theorem in Bayesian approach is Bayes's theorem, which convert a prior probability into a posterior probability.

1.1.2 Three Types of Bayesian Approach

- Fully Bayesian: Marginalize with respect to hyper-parameters as well as parameters.

  For a curve fitting problem modeled by

  $$p(\boldsymbol{w} \mid \alpha)=N\left(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1} \boldsymbol{I}\right), p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta)=N\left(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}\right), \tag{1}$$

  the result is given by

  $$p(t \mid \boldsymbol{t})=\iiint p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \boldsymbol{t}, \alpha, \beta) p(\alpha, \beta \mid \boldsymbol{t}) d \boldsymbol{w} d \alpha d \beta. \tag{2}$$

- Empirical Bayes: First consider hyper-parameters α and β as a fixed value and then calculate them by maximizing the marginal likelihood (evidence function). With the value of $\alpha^\star$ and $\beta^\star$, the result is given by

  $$p(t \mid \boldsymbol{t}) \approx p\left(t \mid \boldsymbol{t}, \alpha^{*}, \beta^{*}\right)=\int p\left(t \mid \boldsymbol{w}, \beta^{*}\right) p\left(\boldsymbol{w} \mid \boldsymbol{t}, \alpha^{*}, \beta^{*}\right) d \boldsymbol{w}. \tag{3}$$

- MAP: A point estimation of model parameters. The value of w is obtained by maximizing the following equation,

$$p(w \mid D) = \frac{p(D \mid w)p(w)}{p(D)} \propto p(D \mid w)p(w). \tag{4}$$

1.1.3 Frequentist Approach versus Bayesian Approach

In a frequentist treatment, we choose specific values for the parameters by optimizing some criterion, such as the likelihood function. However, the over-fitting problem, which can be understood as a general property of maximum likelihood, often occurs due to the limitation of the size of data set. Introducing a penalty terms into the optimization criterion can be consider as one of the solution to this problem. L1 (Lasso regression) and L2 (ridge regression) regularizer are commonly imposed on maximum likelihood criterion. Nonetheless, cross-validation can also be adopted to overcome the over-fitting problem.

By adopting a Bayesian approach, the over-fitting problem can be avoided. In the framework of evidence approximation, the effective number of parameters in a Bayesian model adapts automatically to the size of the data set. In a particular case, the estimated parameters of the regularized least squares with L2 norm as a regularizer is equivalent to the result of the MAP solution if the data is a Gaussian distribution and the prior distribution of the parameters is also set to be Gaussian. However, this is not the case of a fully Bayesian approach.

One of the problem for a Bayesian approach is the difficulty of calculation of the marginalization. One of the alternative approach is sampling, for example Markov chain Monte Carlo. In addition, deterministic approximation such as variational Bayes and expectation propagation can also be used as an substitution.

Another problem is the choice of a prior distribution. A conjugate distribution is a popular choice of a prior distribution, which leads to posterior distributions have the same functional form as the prior, and therefore leads to a simplified Bayesian analysis. However, the conjugate distribution may turns out to be inappropriate for particular applications.

1.2 Decision Theory

1.2.1 Inference and Decision

While the probability theory provides us a mathematical framework to measure the uncertainty, the decision theory help us to make optimal decision involving uncertainty.

In the inference stage, we determine the joint probability distribution, conditional distribution or the posterior distribution. However, in a practical application, after the prediction, we also need to take a specific action based on the prediction we made, and this aspect falls into the category of decision theory.

1.2.2 Three Distinct Approaches for Solving Decision Problems

- Discriminant function: Find a function $f(x)$, called a discriminant function, which maps each input x directly onto a class label.
- Discriminative models: First solve the inference problem of determining the posterior class probabilities $p(C_k|x)$, then subsequently use decision theory to assign each x to one of the classes.
- Generative models: Model the joint distribution $p(x, C_k)$ directly and then normalize to obtain the posterior probabilities. Having found the posterior probabilities, we usu the decision theory to determine the class of x.

2 Linear Regression

The linear model (with respect to parameters) for regression is one that involve linear combinations of fixed nonlinear (linear) functions of the input variables, of the form

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \tag{5}$$

where $\mathbf{w} = (w_0, \dots, w_{M-1})^{\mathrm{T}}$ and $\phi = (\phi_0, \dots, \phi_{M-1})^{\mathrm{T}}$.

2.1 Maximum Likelihood and Least Squares

We assume that the target variable t is given by a deterministic function $y(\mathrm{x}, \mathrm{w})$ with additive Gaussian noise so that

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \tag{6}$$

where ϵ is a zero mean Gaussian random variable. We can maximize the Least Square and the likelihood function with respect to w, which gives the results as

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}. \tag{7}$$

2.2 Bayesian Linear Regression

We define the likelihood function $p(\mathbf{t} \mid \mathbf{w})$ as

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}\left(\mathbf{x}_n\right), \beta^{-1}\right), \tag{8}$$

where noise precision parameter β is a known constant. The corresponding conjugate prior is therefore given by a Gaussian distribution of the form

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0\right). \tag{9}$$

The posterior distribution can by obtained by exploit the result of marginal and conditional Gaussians,

$$p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N\right), \tag{10}$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta\ ^{\mathrm{T}} \mathbf{t}\right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \Phi^{\mathrm{T}}\ . \end{aligned} \tag{11}$$

Then the predictive distribution is given by

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \mathbf{w}, \beta) p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) \mathrm{d}\mathbf{w}. \tag{12}$$

3 Linear Classification

3.1 Discriminant Functions

By considering a K-class discriminant comprising K linear functions of the form

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}} \mathbf{x} + w_{k0}, \tag{13}$$

and then assigning a point x to class $\mathcal{C}_k$ if $y_k(\mathrm{x}) > y_j(\mathrm{x})$ for all $j \neq k$. The decision boundary between class $\mathcal{C}_k$ and class $\mathcal{C}_j$ is therefore given by $y_k(\mathrm{x}) = y_j(\mathrm{x})$ and hence corresponds to a $(D-1)$ -dimensional hyperplane defined by

$$\left(\mathbf{w}_k - \mathbf{w}_j\right)^{\mathrm{T}} \mathbf{x} + \left(w_{k0} - w_{j0}\right) = 0. \tag{14}$$

The approaches to learning the parameters of linear discriminant functions are

- Least Squares,
- Fisher’s Linear Discriminant,
- Perceptron Algorithm.

3.2 Probabilistic Generative Models

We first model the class-conditional densities $p\left(\mathbf{x} \mid \mathcal{C}_k\right)$, as well as the class priors $p\left(\mathcal{C}_k\right)$, and then use these to compute posterior probabilities $p\left(\mathcal{C}_k \mid \mathbf{x}\right)$ through Bayes’ theorem.

For the case of $K > 2$ classes, we have

$$\begin{aligned} p\left(\mathcal{C}_k \mid \mathbf{x}\right) &= \frac{p\left(\mathbf{x} \mid \mathcal{C}_k\right) p\left(\mathcal{C}_k\right)}{\sum_j p\left(\mathbf{x} \mid \mathcal{C}_j\right) p\left(\mathcal{C}_j\right)}, \\ &= \frac{\exp\left(a_k\right)}{\sum_j \exp\left(a_j\right)}. \end{aligned} \tag{15}$$

which is known as the normalized exponential and can be regarded as a multiclass generalization of the logistic sigmoid. Here the quantities a_k are defined by

$$a_k = \ln p\left(\mathbf{x} \mid \mathcal{C}_k\right) p\left(\mathcal{C}_k\right). \tag{16}$$

The normalized exponential is also known as the softmax function, as it represents a smoothed version of the ’max’ function because, if $a_k \gg a_j$ for all $j \neq k$, then

$$p\left(\mathcal{C}_k \mid \mathbf{x}\right) \simeq 1, \text{ and } p\left(\mathcal{C}_j \mid \mathbf{x}\right) \simeq 0. \tag{17}$$

If we assume that the class-conditional densities are Gaussian and all the classes share the same covariance matrix. Thus the density for class $\mathcal{C}_k$ is given by

$$p\left(\mathbf{x} \mid \mathcal{C}_k\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\ |^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)^{\mathrm{T}}\ {}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)\right\}. \tag{18}$$

For the general case of K classes we have,

$$p\left(\mathcal{C}_1 \mid \mathbf{x}\right) = \sigma\left(a_k(\mathrm{x})\right), \tag{19}$$

$$a_k(\mathrm{x}) = \mathbf{w}_k^{\mathrm{T}} \mathrm{x} + w_{k0}, \tag{20}$$

where we have defined

$$\begin{aligned} \mathbf{w}_k &= \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k, \\ w_{k0} &= -\frac{1}{2} \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln p\left(\mathcal{C}_k\right). \end{aligned} \tag{21}$$

Like what we have illustrated in section 3.1, the decision boundary is given by equation (14).

3.2.1 MLE Solution

Consider the case of two classes, each having a Gaussian class-conditional density with a shared covariance matrix, and suppose we have a data set $\{\mathbf{x}_n, t_n\}$ where $n = 1, \ldots, N$. Here $t_n = 1$ denotes class $\mathcal{C}_1$ and $t_n = 0$ denotes class $\mathcal{C}_2$. We denote the prior class probability $p(\mathcal{C}_1) = \pi$, so that $p(\mathcal{C}_2) = 1 - \pi$. Thus the likelihood function is given by

$$p(\mathbf{t} \mid \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left[\pi \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n} \left[(1-\pi)\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1-t_n}, \tag{22}$$

where $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$.

3.3 Probabilistic Discriminative Models

3.3.1 Logistic Regression

In the discussion of generative approaches in Section 3.2, we saw that under general assumptions, the posterior probability of class $\mathcal{C}_1$ can be written as a logistic sigmoid acting on a linear function of the feature vector ϕ so that

$$p(\mathcal{C}_1 \mid \boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}\right), \tag{23}$$

with $p(\mathcal{C}_2 \mid \phi) = 1 - p(\mathcal{C}_1 \mid \phi)$.

For a data set $\{\phi_n, t_n\}$, where $t_n \in \{0, 1\}$ and $\phi_n = \phi(\mathrm{x}_n)$, with $n = 1, \ldots, N$, the likelihood function can be written as

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n}, \tag{24}$$

where $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ and $y_n = p(\mathcal{C}_1 \mid \phi_n)$. As usual, we can define an error function by taking the negative logarithm of the likelihood, which gives the cross-entropy error function in the form

$$E(\mathbf{w}) = -\ln p(\mathbf{t} \mid \mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}, \tag{25}$$

where $y_n = \sigma(a_n)$ and $a_n = \mathrm{w}^{\mathrm{T}} \phi_n$. Taking the gradient of the error function with respect to w, we obtain

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \phi_n. \tag{26}$$

3.4 Mixture Models and EM

3.4.1 Mixtures of Gaussians

The Gaussian mixture distribution can be written as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{27}$$

Consider a K-dimensional binary random variable z having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0 . The values of z_k therefore satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. Because z uses a 1-of-K representation, we can write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{28}$$

The goal of the EM algorithm is to find maximum likelihood solutions for models having latent variables. We denote the set of all observed data by X, and similarly the set of all latent variables by Z. The set of all model parameters is denoted by $\boldsymbol{\theta}$, and so the distribution is given by

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}). \tag{29}$$

Suppose for each observation in X, we were told the corresponding value of the latent variable Z. We would call $\{\mathbf{X}, \mathbf{Z}\}$ the complete data set, and we shall refer to the actual observed data X as incomplete. The distribution for the complete data set simply takes the form $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$ which takes the form

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_{nk}}. \tag{30}$$

3.4.2 Maximum Likelihood

From (27) the log of the likelihood function is given by

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}. \tag{31}$$

A key observation is that the summation over the latent variables appears inside the logarithm. Even if the joint distribution $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$ belongs to the exponential family, the marginal distribution $p(\mathbf{X} \mid \boldsymbol{\theta})$ typically does not as a result of this summation. The presence of the sum prevents the logarithm from acting directly on the joint distribution, resulting in complicated expressions for the maximum likelihood solution.

3.4.3 EM Algorithm

In practice, however, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, but only the incomplete data $\mathbf{X}$. Our knowledge of the values of the latent variables in Z is given only by the posterior distribution $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$.

In the E step, we use the current parameter values θ^{old} to find the posterior distribution of the latent variables given by $p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right)$. We then use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value θ. This expectation, denoted $\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$, is given by

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \sum_{\mathbf{Z}} p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}). \tag{32}$$

In the M step, we determine the revised parameter estimate θ^{new} by maximizing this function

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}\left(\theta, \theta^{\text{old}}\right). \tag{33}$$

Note that in the definition of $\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$, the logarithm acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$, and so the corresponding M-step maximization will be tractable.

3.4.4 The EM Algorithm in General

For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q|p), \tag{34}$$

where we have defined

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}, \\ \mathrm{KL}(q|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. \end{aligned} \tag{35}$$

From (35), we see that $\mathrm{KL}(q|p)$ is the Kullback-Leibler divergence and satisfies $\mathrm{KL}(q|p) \geqslant 0$, with equality if, and only if, $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$. It therefore follows from (34) that $\mathcal{L}(q, \boldsymbol{\theta}) \leqslant \ln p(\mathbf{X} \mid \boldsymbol{\theta})$, in other words that $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$.

The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions. In the E step, the lower bound $\mathcal{L}\left(q, \boldsymbol{\theta}^{\text{old}}\right)$ is maximized with respect to $q(\mathbf{Z})$ while holding $\boldsymbol{\theta}^{\text{old}}$ fixed by setting the $q(\mathbf{Z})$ equal to the posterior distribution $p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right)$. In this case, the lower bound will equal the log likelihood since the Kullback-Leibler divergence is zero.

In the subsequent M step, the distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$. If we substitute $q(\mathbf{Z}) = p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right)$ into (35), we see that, after the E step, the lower bound takes the form

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}} p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \\ &= \mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) + \text{const},\end{aligned} \tag{36}$$

where the constant is simply the negative entropy of the q distribution and is therefore independent of θ.

Consider the fact that the $\ln p(\mathbf{X} \mid \boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta})$, EM algorithm becomes a majorization-minimization algorithm, as shown in Figure 1.

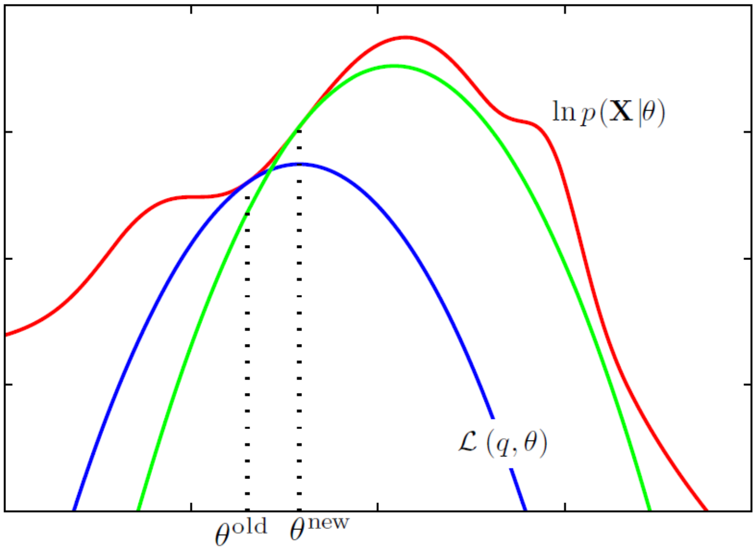

Figure 1: The EM algorithm alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values.

3.5 Variational Inference (VI)

A central task in the application of probabilistic models is the evaluation of the posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$ of the latent variables $\mathbf{Z}$ given the observed data variables X. However, if $p(\mathbf{Z} \mid \mathbf{X})$ is too complex to represent and therefore to inference, we can use VI to approximate a simpler distribution $q(\mathbf{Z} \mid \theta)$. If q and p are similar under some metrics, then we can consider q as a approximated result of p. The idea of VI is to transform an inference process into a optimization problem.

In our discussion of EM, we can decompose the log marginal probability using

$$\ln p(\mathbf{X}) = \mathcal{L}(\theta) + \mathrm{KL}(q|p). \tag{37}$$

As before, we can maximize the lower bound $\mathcal{L}(\theta)$ by optimization, which is equivalent to minimizing the KL divergence. If we allow any possible choice for $q(\mathbf{Z} \mid \theta)$, then the maximum of the lower bound occurs when the KL divergence vanishes, which occurs when $q(\mathbf{Z} \mid \theta)$ equals the posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$.

3.5.1 Example: A construction of $q(\mathbf{Z} \mid \theta)$

Suppose we partition the elements of $\mathbf{Z}$ into disjoint groups that we denote by $\mathbf{Z}_i$ where $i = 1, \ldots, M$. We then assume that the q distribution factorizes with respect to these groups, so that

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i\left(\mathbf{Z}_i\right). \tag{38}$$

Take GMM for example, since we know that p is a Gaussian distribution, we can assume that q is also Gaussian. Note that if p and q are quite different distribution, then the approximation can be considered invalid.

According to equation (38), the $q(\mathbf{Z} \mid \theta)$ can be approximated by

$$p(\mathbf{Z} \mid \mathbf{X}) \approx q(\mathbf{Z}|\theta) = \prod_{k=1}^{K} q\left(z_k|\theta_k\right). \tag{39}$$