Chapter 9

Approximate Bayesian Learning

Although the general approach to Bayesian learning is simple in principle, it can lead to computational difficulties when applied to complex probabilistic models. In Maximum-Likelihood learning of the parameters in hidden Markov models or Gaussian mixture models, we also encountered similar problems, and Ch. 7 showed that the Expectation Maximization algorithm provides an elegant solution.

Typically, computational difficulties in Bayesian learning arise when we want to derive posterior densities of model parameters in mixture models, where the complete model includes hidden variables that control the choice of mixture components as, for example, in a GMM or an HMM.

This chapter presents *Variational Inference* (VI) as an approach that can be used for Bayesian learning in those more complex situations. Bishop (2006) gives a detailed discussion of VI and other approximate methods for Bayesian learning, for example:

- *Numerical sampling.* Even if the exact posterior parameter distribution cannot be expressed in a closed form, it is possible to generate random samples that follow the exact distribution. The samples can be used, for example, to calculate predictive probabilities. The accuracy can be very good, at the cost of large amounts of computation.

- *Expectation Propagation* (EP). This is, like VI, an analytical method to approximate the posterior parameter distribution. However, while the VI approximation, in general, is most accurate near the peak of the distribution, the EP concentrates on describing the global properties of the distribution, such as its mean and covariance, but may conversely be less accurate near the peak.

Here we focus on *Variational Inference* mainly because it often gives explicit formulas that can be solved analytically, and because it is similar to EM.

As we will see, VI can be regarded as a generalization of the "EM trick," introduced in Ch. 7. Like EM, we get an iterative optimization procedure that converges to a locally optimal solution. Unlike EM, however, we are able to stochastically model our uncertainty in the parameters and all other unknowns.

9.1 Variational Inference – Notation

For a very general solution to this kind of problem, in this section $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M)$ denotes a combined set of groups of hidden variables and groups of other model parameters, all regarded as random variables. For the HMM, for example, we might let $\boldsymbol{Z}_1$ represent the sequence of discrete state variables that we used to denote as $\underline{S} = (S_1, \ldots, S_T)$, while $\boldsymbol{Z}_2$ might include all the elements in the transition probability matrix that we used to call A, and $\boldsymbol{Z}_3$ could include all the mean vectors of state-conditional Gaussian output density functions, etc. Thus, some variable groups can have discrete distributions, specified by probability mass functions, and other groups may have continuous distributions.

As usual, the training procedure uses a sequence $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of observed feature vectors $\boldsymbol{x}_t$ that are regarded as samples of corresponding random vectors $\boldsymbol{X}_t$. The observed vectors in the sequence may be drawn from identical or different distributions, statistically independent or dependent of other variables in the sequence. All such model details are specified by some variables in $\underline{\boldsymbol{Z}}$. Some variable group $\boldsymbol{Z}_m$ may include exactly T elements, corresponding to the T observed vectors, whereas other parameter groups may have a fixed size, regardless of the number of observed feature vectors. One variable group $\boldsymbol{Z}_i$ may have a size that corresponds to the number of components in a mixture model, or the number of internal hidden states, and another variable group may control how many such mixture components, or hidden states, that are actually needed in the model. In short, $\underline{\boldsymbol{Z}}$ collects everything that is unknown in the current situation.

Just as in the standard Bayesian approach, we formulate an explicit conditional probability model $f_{\underline{\boldsymbol{X}}|\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{x}} \mid \underline{\boldsymbol{z}})$ for the observations, given all the model parameters and hidden variables. We also need a prior model (density and/or mass) $f_{\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{z}})$ for all the unknown variables and parameters. The difficult step is to obtain a useful posterior distribution

$$f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}}) \propto f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{z}}) = f_{\underline{\boldsymbol{X}}|\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{x}} \mid \underline{\boldsymbol{z}}) f_{\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{z}}) \tag{9.1}$$

Although this expression still looks quite simple, the problem is that we would often like to obtain a separate posterior density for some of the parameter groups in $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M)$, *independent* of other groups of hidden variables. For example, when training HMM parameters with the EM approach, the trained model λ should not depend explicitly on the specific

hidden state sequence that might have generated the training data, because the model is to be applied to future test data, generated by other hidden state sequences. In the exact Bayesian posterior of Eq. (9.1), all the groups of parameters and hidden variables, gathered in $\underline{\boldsymbol{Z}}$, may still depend on each other in some complex way, and it may be computationally intractable to find the marginal distribution of, e.g., $\boldsymbol{Z}_1$, as

$$f_{\boldsymbol{Z}_1|\underline{\boldsymbol{X}}}(\boldsymbol{z}_1 \mid \underline{\boldsymbol{x}}) = \int_{\boldsymbol{z}_2} \cdots \int_{\boldsymbol{z}_M} f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M \mid \underline{\boldsymbol{x}})\, d\boldsymbol{z}_2 \ldots d\boldsymbol{z}_M \tag{9.2}$$

by integrating out all the other unwanted variables.

9.2 Variational Inference – General Solution

The goal of the variational inference approach is to find the best possible approximation of the exact posterior distribution,

$$q(\underline{\boldsymbol{z}}) \approx f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}}) \tag{9.3}$$

within the constraints imposed by the structure and mathematical form chosen for the approximate density (and/or mass) function q. The model designer is free to choose any suitable parametric mathematical form for this function.

In the following the shorthand notation $E_q\left[h(\underline{\boldsymbol{Z}})\right]$ denotes the expected value of the random (transformed) variable $h(\underline{\boldsymbol{Z}})$, calculated using the density function q, as

$$E_q\left[h(\underline{\boldsymbol{Z}})\right] = \int q(\underline{\boldsymbol{z}}) h(\underline{\boldsymbol{z}}) d\underline{\boldsymbol{z}} \tag{9.4}$$

We now derive an optimization criterion from the following expressions for the log-likelihood of the observed data:

$$\begin{aligned}
\ln f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}}) &= E_q\left[\ln f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}})\right] = \\
&= E_q\left[\ln \frac{f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}} \mid \underline{\boldsymbol{x}}) f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}})}{f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}} \mid \underline{\boldsymbol{x}})}\right] = E_q\left[\ln \frac{f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}}, \underline{\boldsymbol{x}})}{f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}} \mid \underline{\boldsymbol{x}})}\right] = \\
&= \underbrace{E_q\left[\ln \frac{f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}}, \underline{\boldsymbol{x}})}{q(\underline{\boldsymbol{Z}})}\right]}_{\mathcal{L}(q)} + \underbrace{E_q\left[\ln \frac{q(\underline{\boldsymbol{Z}})}{f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}} \mid \underline{\boldsymbol{x}})}\right]}_{\mathrm{KL}\left(q \,\|\, f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}\right)}
\end{aligned} \tag{9.5}$$

Here, the equality on the first line is valid for any q, simply because the log-likelihood, $\ln f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}})$, by definition does not depend on $\underline{\boldsymbol{Z}}$. The equality on the second line follows from Bayes' rule, and the expansion on the third line simply divides and multiplies by $q(\underline{\boldsymbol{Z}})$.

The Kullback-Leibler divergence $\mathrm{KL}\left(q \,\|\, f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}\right)$ is defined (see Sec. 9.2.1) to be a non-negative measure of the "distance" between q and $f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}$. It is

zero only if $q(\underline{\boldsymbol{z}}) = f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}})$ for all $\underline{\boldsymbol{z}}$. Thus, the remaining term $\mathcal{L}(q)$ on the third line of Eq. (9.5) is a lower bound to the log-likelihood $\ln f_{\underline{\boldsymbol{X}}}(\underline{\boldsymbol{x}})$. By adjusting the function q to find

$$\hat{q} = \underset{q}{\operatorname{argmax}}\, \mathcal{L}(q) = \underset{q}{\operatorname{argmax}}\, E_q\left[\ln \frac{f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}},\underline{\boldsymbol{x}})}{q(\underline{\boldsymbol{Z}})}\right] \tag{9.6}$$

which maximizes the lower bound on the the log-likelihood, we reach an optimal approximation in the sense that the Kullback-Leibler divergence $\mathrm{KL}\left(q \,\|\, f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}\right)$ is minimal. As we already have an explicit expression for $f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\underline{\boldsymbol{z}},\underline{\boldsymbol{x}}) = f_{\underline{\boldsymbol{X}}|\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{x}} \mid \underline{\boldsymbol{z}}) f_{\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{z}})$, and we are free to choose a mathematical form for q, the objective function $\mathcal{L}(q)$ can be expressed in a tractable form. The details of the optimization depend, of course, on the form chosen for q.

For example, if q is a density function defined by a set of hyper-parameters $\boldsymbol{\theta}$, the objective becomes a (non-linear) function $Q(\boldsymbol{\theta})$, and the optimization can be performed simply by varying those hyper-parameters using a suitable optimization algorithm. In this way, VI can be used as a technique to impose a simple approximate parametric form on complicated posterior distributions. As discussed in Sec. 9.3, we can sometimes use the expression $\mathcal{L}(q)$ to find a suitable mathematical form for q.

9.2.1 Kullback-Leibler Divergence

The Kullback-Leibler divergence, also called *relative entropy*, is a logarithmic measure of the difference between two probability distributions.

Definition 9.1 (KL divergence): *Given two probability density (or mass) functions q and p, both defined for random variables $\boldsymbol{Z}$ in the same space, the Kullback-Leibler divergence is defined as*

$$\mathrm{KL}\left(q \,\|\, p\right) = E_q\left[\ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right] \tag{9.7}$$

□

If the random variable is continuous-valued, the expectation is

$$\mathrm{KL}\left(q \,\|\, p\right) = E_q\left[\ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right] = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\, d\boldsymbol{z} \tag{9.8}$$

If the distribution is discrete, q and p are probability mass functions, and the expectation is

$$\mathrm{KL}\left(q \,\|\, p\right) = E_q\left[\ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right] = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} \tag{9.9}$$

The definition is asymmetric in the arguments, so $\mathrm{KL}\left(q \,\|\, p\right) \neq \mathrm{KL}\left(p \,\|\, q\right)$ in general.

Theorem 9.1: *The Kullback-Leibler divergence is non-negative,*

$$\mathrm{KL}\,(q \,\|\, p) \geq 0$$

with equality if and only if $q(\boldsymbol{z}) = p(\boldsymbol{z})$ *for all* $\boldsymbol{z}$. □

Proof: For any *convex* function $h(\,\cdot\,)$, *Jensen's inequality* guarantees that

$$E\,[h(g(\boldsymbol{Z}))] \geq h\,(E\,[g(\boldsymbol{Z}]))\tag{9.10}$$

where $Y = g(\boldsymbol{Z})$ is a transformed scalar variable defined by some function g. As the function $h(\,\cdot\,) = -\ln(\,\cdot\,)$ is convex, Jensen's inequality can be applied to the Kullback-Leibler divergence as

$$\begin{aligned}\mathrm{KL}\,(q \,\|\, p) = E_q\left[-\ln\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] \geq -\ln E_q\left[\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] = \\ = -\ln\int_{\boldsymbol{z}} q(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\,d\boldsymbol{z} = -\ln\int_{\boldsymbol{z}} p(\boldsymbol{z})\,d\boldsymbol{z} = 0\end{aligned}\tag{9.11}$$

The integral equals 1, simply because p is a normalized probability density function. If the distributions are discrete, the integral is replaced by a sum, and the result is the same. ■

9.3 Factorized Approximation

As already mentioned in Sec. 9.1, the complete set of hidden variables and model parameters, $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M)$, might include separate groups of variables. In order to simplify the learning procedure, or for other reasons, it may be desirable to approximate the posterior density function as a factorized product, as

$$f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}}) \approx q(\underline{\boldsymbol{z}}) = q_1(\boldsymbol{z}_1)\cdots q_M(\boldsymbol{z}_M)\tag{9.12}$$

This means that we model the different groups of variables in $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M)$ as statistically *independent*, although the exact posterior density in Eq. (9.1) may include some complex dependencies between the groups.

We will now show that the optimal density for each group can be found iteratively, by choosing

$$\ln q_i(\boldsymbol{z}_i) = E_{q_{j\neq i}}\left[\ln f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\boldsymbol{Z}_1, \ldots \boldsymbol{Z}_{i-1}, \boldsymbol{z}_i, \boldsymbol{Z}_{i+1}, \ldots, \boldsymbol{Z}_M, \underline{\boldsymbol{x}})\right] + c\tag{9.13}$$

Here, c is just a normalization constant, and $E_{q_{j\neq i}}[\,]$ means that the density functions q_j are kept fixed for all $j \neq i$, and these functions are used to calculate the expectation over all those *other* groups of variables $\boldsymbol{Z}_j$, *except* $\boldsymbol{Z}_i$. This is repeated for $i = 1, \ldots, M$ to improve each approximate density

q_i at a time, while keeping all the other functions $q_{j\neq i}$ fixed. As the different density functions are actually coupled, the complete procedure must be iterated until the result approaches a stable solution.

The criterion function $\mathcal{L}(q)$ in Eq. (9.5) cannot decrease in any step of this procedure. Therefore, the complete procedure is guaranteed to converge towards a locally optimal solution to the problem formulated in Eq. (9.6). The stepwise improvement of $\mathcal{L}(q)$ is guaranteed by the following theorem:

Theorem 9.2: *Given a joint likelihood function $f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\boldsymbol{z}_1, \boldsymbol{z}_2, \underline{\boldsymbol{x}})$ for a set of variables $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ and an observation $\underline{\boldsymbol{x}}$, we may seek a factorized approximation $q(\boldsymbol{z}_1, \boldsymbol{z}_2) = q_1(\boldsymbol{z}_1)q_2(\boldsymbol{z}_2) \approx f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\boldsymbol{z}_1, \boldsymbol{z}_2 \mid \underline{\boldsymbol{x}})$. Then, for any fixed q_2, a density function q_1, obtained as*

$$\ln q_1(\boldsymbol{z}_1) = E_{q_2}\left[\ln f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\boldsymbol{z}_1, \boldsymbol{Z}_2, \underline{\boldsymbol{x}})\right] + c, \tag{9.14}$$

maximizes

$$\mathcal{L}(q) = E_q\left[\ln \frac{f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\underline{\boldsymbol{Z}}, \underline{\boldsymbol{x}})}{q(\underline{\boldsymbol{Z}})}\right] \tag{9.15}$$

□

Proof: After taking the expectation over $\boldsymbol{Z}_2$ in Eq. (9.14), the remaining expression is just a function of $\boldsymbol{z}_1$. With proper normalization, this function can be interpreted as the logarithm of a density function, called $\tilde{p}(\boldsymbol{z}_1)$ here. The objective function $\mathcal{L}(q)$ can be expressed as

$$\begin{aligned}
\mathcal{L}(q) = \int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1) \underbrace{\int_{\boldsymbol{z}_2} q_2(\boldsymbol{z}_2) \ln f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\boldsymbol{z}_1, \boldsymbol{z}_2, \underline{\boldsymbol{x}})\, d\boldsymbol{z}_2}_{E_{q_2}\left[\ln f_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}(\boldsymbol{z}_1, \boldsymbol{Z}_2, \underline{\boldsymbol{x}})\right] = \ln \tilde{p}(\boldsymbol{z}_1) + \text{const.}} d\boldsymbol{z}_1 \\
- \int_{\boldsymbol{z}_1}\int_{\boldsymbol{z}_2} q_1(\boldsymbol{z}_1)q_2(\boldsymbol{z}_2)\left(\ln q_1(\boldsymbol{z}_1) + \ln q_2(\boldsymbol{z}_2)\right) d\boldsymbol{z}_2\, d\boldsymbol{z}_1 = \\
= \int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1) \ln \tilde{p}(\boldsymbol{z}_1)\, d\boldsymbol{z}_1 + \text{const.} \\
- \int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1) \ln q_1(\boldsymbol{z}_1)\, d\boldsymbol{z}_1 \underbrace{\int_{\boldsymbol{z}_2} q_2(\boldsymbol{z}_2)\, d\boldsymbol{z}_2}_{=1} \\
- \underbrace{\int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1)\, d\boldsymbol{z}_1}_{=1} \underbrace{\int_{\boldsymbol{z}_2} q_2(\boldsymbol{z}_2) \ln q_2(\boldsymbol{z}_2)\, d\boldsymbol{z}_2}_{=\text{const.}} = \\
= \int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1) \ln \tilde{p}(\boldsymbol{z}_1)\, d\boldsymbol{z}_1 - \int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1) \ln q_1(\boldsymbol{z}_1)\, d\boldsymbol{z}_1 + \text{const.} = \\
= \int_{\boldsymbol{z}_1} q_1(\boldsymbol{z}_1) \ln \frac{\tilde{p}(\boldsymbol{z}_1)}{q_1(\boldsymbol{z}_1)}\, d\boldsymbol{z}_1 + \text{const.} = \\
= -\,\mathrm{KL}\left(q_1 \,\|\, \tilde{p}\right) + \text{const.}
\end{aligned} \tag{9.16}$$

As the Kullback-Leibler divergence on the last line is minimized to zero, if $q_1(\boldsymbol{z}_1) = \tilde{p}(\boldsymbol{z}_1)$ for all $\boldsymbol{z}_1$, this choice maximizes $\mathcal{L}(q)$ as stated.

This proof also covers the general formulation in Eq. (9.13). We have simply renamed q_i and $q_{j\neq i}$ as q_1 and q_2 in the proof. ∎

9.4 VI Example with Solution

Example 9.1: Consider a sequence $\underline{x} = (x_1, \ldots, x_T)$ of scalar samples x_t, drawn from i.i.d. random variables X_t with the GMM density function[1]

$$f_{X_t|\Theta}(x_t \mid \theta) = 0.5\frac{1}{\sqrt{2\pi}}e^{-x_t^2/2} + 0.5\frac{1}{\sqrt{2\pi}}e^{-(x_t-\theta)^2/2}, \quad \text{for all } t \tag{9.17}$$

The mean parameter θ of the second Gaussian component is unknown and modeled as an outcome of a random variable Θ. The prior distribution for Θ is assumed to be broad and uniform,

$$f_\Theta(\theta) \to \frac{1}{c}, \qquad c \to \infty \tag{9.18}$$

Calculate the posterior density $f_{\Theta|\underline{X}}(\theta \mid \underline{x})$ for the parameter, given the observed data.

Solution:

The given mixture density is equivalent to assuming that each x_t is generated in a two-step random procedure: First, a binary sample z_t is drawn from the random variable Z_t, with $P[Z_t = 0] = P[Z_t = 1] = 0.5$, and then x_t is generated from the conditional distribution for X_t, given z_t. If $z_t = 0$, the Gaussian component with known zero mean is used, and if $z_t = 1$, the Gaussian component with unknown mean θ is used. Thus, the conditional density for X_t can be written as

$$f_{X_t|Z_t,\Theta}(x_t \mid z_t, \theta) = \left(\frac{1}{\sqrt{2\pi}}e^{-x_t^2/2}\right)^{1-z_t}\left(\frac{1}{\sqrt{2\pi}}e^{-(x_t-\theta)^2/2}\right)^{z_t} \tag{9.19}$$

The prior probability mass function for the hidden sequence $\underline{Z}$ can be written as

$$f_{\underline{Z}}(\underline{z}) = \prod_{t=1}^{T} f_{Z_t}(z_t) = \prod_{t=1}^{T} 0.5^{1-z_t}0.5^{z_t} \tag{9.20}$$

Applying the improper constant prior $f_\Theta(\theta) = 1/c$, the joint probability density and mass for the complete data, including observations $\underline{X}$, hidden

[1]This example was proposed by Wasserman (2000, 2012) to point out an interesting difficulty with exact Bayesian learning in mixture models, also discussed in Problem 9.1. The solution presented here uses approximate Bayesian learning to avoid this difficulty.

variables $\underline{Z}$, and the parameter Θ, is

$$f_{\underline{X},\underline{Z},\Theta}(\underline{x},\underline{z},\theta) = f_\Theta(\theta)\prod_{t=1}^{T} f_{X_t|Z_t,\Theta}(x_t \mid z_t,\theta) f_{Z_t}(z_t) = \\ = \frac{1}{c}\prod_{t=1}^{T}\left(\frac{0.5}{\sqrt{2\pi}}e^{-x_t^2/2}\right)^{1-z_t}\left(\frac{0.5}{\sqrt{2\pi}}e^{-(x_t-\theta)^2/2}\right)^{z_t} \quad (9.21)$$

To find a posterior density of the desired form $f_{\Theta|\underline{X}}(\theta \mid \underline{x})$, we use the factorized approximation

$$f_{\Theta,\underline{Z}|\underline{X}}(\theta,\underline{z} \mid \underline{x}) \approx q_1(\theta)q_2(\underline{z}) \quad (9.22)$$

For this purpose, we must start from the log-likelihood

$$\ln f_{\underline{X},\underline{Z},\Theta}(\underline{x},\underline{z},\theta) = \sum_{t=1}^{T} -(1-z_t)\frac{x_t^2}{2} - z_t\frac{(x_t-\theta)^2}{2} + \text{const.} \quad (9.23)$$

Here, the constant prior $1/c$, and all other constants, are collected in the "const." term. The solution in Eq. (9.13) gives two coupled equations:

$$\ln q_1(\theta) = \sum_{t=1}^{T} -E_{q_2}[Z_t]\frac{(x_t-\theta)^2}{2} + \text{const.} \quad (9.24)$$

$$\ln q_2(\underline{z}) = \sum_{t=1}^{T} -(1-z_t)\frac{x_t^2}{2} - z_t\frac{E_{q_1}[(x_t-\Theta)^2]}{2} + \text{const.} \quad (9.25)$$

As the equations are coupled, they must be solved iteratively: When calculating $E_{q_2}[Z_t]$ in the first equation, we use the approximate q_2 obtained in the previous round. To calculate the expectation over Θ in the second equation, we use the approximate q_1 from the previous iteration.

As the log-density function $\ln q_1(\theta)$ in (9.24) is a second-degree polynomial in θ, the density must be Gaussian, i.e., it is specified by two hyperparameters, the mean μ and the variance σ^2, as

$$q_1(\theta) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} \quad (9.26)$$

Using the shorthand notation $\gamma_t = E_{q_2}[Z_t]$ in (9.24), we get

$$\ln q_1(\theta) = -\frac{\theta^2\sum_t\gamma_t - 2\theta\sum_t\gamma_t x_t + \sum_t\gamma_t x_t^2}{2} + \text{const.} \quad (9.27)$$

Here, the hyperparameters μ and σ are easily identified as

$$\frac{1}{\sigma^2} = \sum_{t=1}^{T} \gamma_t \tag{9.28}$$

$$\frac{\mu}{\sigma^2} = \sum_{t=1}^{T} \gamma_t x_t \tag{9.29}$$

$$\mu = \frac{\sum_t \gamma_t x_t}{\sum_t \gamma_t} \tag{9.30}$$

These results seem intuitively reasonable: the posterior mean μ for Θ is just a weighted average of observed values x_t, with each observation weighted by the currently estimated probability that it was actually generated by the Gaussian mixture component with the unknown mean Θ. The inverse posterior variance of Θ is $\sum_t \gamma_t$, which is the effective number of observed values assigned to this mixture component. The more observations, the lower the variance.

Similarly, in Eq. (9.25) we see that $\ln q_2(\underline{z}) = \sum_t \ln q_{2,t}(z_t)$, with

$$\begin{aligned} \ln q_{2,t}(z_t) &= -(1-z_t)\frac{x_t^2}{2} - z_t \frac{(x_t-\mu)^2 + E_{q_1}\left[(\mu-\Theta)^2\right]}{2} + \text{const.} = \\ &= (1-z_t)\left(-\frac{x_t^2}{2}\right) + z_t\left(-\frac{(x_t-\mu)^2+\sigma^2}{2}\right) + \text{const.} \end{aligned} \tag{9.31}$$

Thus, the posterior probability mass for Z_t has the form

$$q_{2,t}(z_t) \propto \left(e^{-\frac{x_t^2}{2}}\right)^{1-z_t} \left(e^{-\frac{(x_t-\mu)^2+\sigma^2}{2}}\right)^{z_t} \tag{9.32}$$

As the probability mass must be normalized so that $q_{2,t}(0) + q_{2,t}(1) = 1$, we can identify the posterior distribution as

$$q_{2,t}(z_t) = (1-\gamma_t)^{1-z_t}\gamma_t^{z_t} \tag{9.33}$$

where

$$\gamma_t = E_{q_2}[Z_t] = \frac{e^{-\frac{(x_t-\mu)^2+\sigma^2}{2}}}{e^{-\frac{x_t^2}{2}} + e^{-\frac{(x_t-\mu)^2+\sigma^2}{2}}} \tag{9.34}$$

■

Again, this result is intuitively reasonable: it is quite similar to the corresponding calculation in Sec. 7.4 for the weight factors in a Gaussian mixture, using the EM algorithm. The estimated weight factor $\gamma_t = P[Z_t = 1 \mid \underline{x}]$ is the currently estimated probability that the observed x_t was generated by the Gaussian component with the unknown mean Θ. The only difference from the EM solution is caused by the fact that the Bayesian approach

does not assume a single point estimate $\Theta = \mu$, but also accounts for the remaining variance σ^2 of the parameter Θ.

To reach the final solution we apply Eqs. (9.28), (9.30), and (9.34) in sequence, and repeat this procedure until the result has converged. The only remaining issue is how to *initialize* the algorithm. The exact method for this is not necessarily critical, as the procedure is guaranteed to reach a local optimum anyway.

In this example, we know that about half of the observed samples are probably generated from each of the two mixture components. Therefore, we can initially make a hard assignment for half of the observations to the component centered at zero, by setting $\gamma_t = 0$ for those x_t values that are closest to zero, and $\gamma_t = 1$ for the other samples that are more distant from zero. When training a GMM with the EM approach, it is also common to initialize the component weight factors by a similar hard assignment of the observed samples to different mixture components.

Example 9.1

9.5 EM – Special Case of Variational Inference

This section derives the EM algorithm as a special case of the more general VI approach. We have an observed training-data sequence $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ which is regarded as an outcome of a corresponding random sequence $\underline{\boldsymbol{X}}$. We need to use a model which includes a set of hidden random variables $\underline{\boldsymbol{Z}}$, as well as a parameter vector $\boldsymbol{\theta}$, which is regarded as an outcome of a random vector $\boldsymbol{\Theta}$. For example, in the case of an HMM, the hidden variables $\underline{\boldsymbol{Z}}$ would be the state sequence, and the parameter vector $\boldsymbol{\theta}$ would include all the HMM parameters. The conditional distribution of the observations, given any outcome of the hidden variables and the parameters, is explicitly known as $f_{\underline{\boldsymbol{X}}|\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}} \mid \underline{\boldsymbol{z}}, \boldsymbol{\theta})$. Prior distributions are also known as $f_{\underline{\boldsymbol{Z}}|\boldsymbol{\Theta}}(\underline{\boldsymbol{z}} \mid \boldsymbol{\theta})$ for the hidden variables, and a (possibly non-informative, perhaps improper) density $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ for the unknown parameters. Thus, the complete log-probability function can be written as

$$\ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{z}}, \boldsymbol{\theta}) = \ln f_{\underline{\boldsymbol{X}}|\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}} \mid \underline{\boldsymbol{z}}, \boldsymbol{\theta}) f_{\underline{\boldsymbol{Z}}|\boldsymbol{\Theta}}(\underline{\boldsymbol{z}} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \tag{9.35}$$

Let us now assume that the goal of the training procedure only is to find a *point estimate* $\hat{\boldsymbol{\theta}}$ for the parameters, in analogy with the EM approach, even though the Bayesian learning actually can produce a more sophisticated model in the form of a full posterior density for $\boldsymbol{\Theta}$. As the estimated parameter values will be applied to future observations, the estimate should also be independent of any particular outcome of the hidden variables $\underline{\boldsymbol{Z}}$, which are valid only for the training data set. Accepting both these restrictions, we apply a factorized approximation to the posterior distribution,

$$f_{\boldsymbol{\Theta},\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\boldsymbol{\theta}, \underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}}) \approx q(\boldsymbol{\theta}, \underline{\boldsymbol{z}}) = q_1(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) q_2(\underline{\boldsymbol{z}}) \tag{9.36}$$

where we also force the posterior density q_1 to be very sharply peaked, such that nearly all the probability is concentrated at the single point $\hat{\boldsymbol{\theta}}$. Formally, this might be expressed as

$$q_1(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) \rightarrow \prod_k \frac{1}{c_k} \delta\left(\frac{\theta_k - \hat{\theta}_k}{c_k}\right) \tag{9.37}$$

using a Dirac delta function for each element θ_k of the parameter vector, with some arbitrary scale hyper-parameters c_k having the same physical dimension as the corresponding θ_k. The main consequence of enforcing a very sharply peaked density is that the expected value of any transformation $g(\boldsymbol{\Theta})$ of the random variable is, asymptotically, just the value at the single point,

$$E_{q_1}\left[g(\boldsymbol{\Theta})\right] = \int q_1(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \rightarrow g(\hat{\boldsymbol{\theta}}) \tag{9.38}$$

In each step of the VI iterative learning procedure, we first apply the general factorized solution in (9.13) to re-estimate the distribution of the hidden variables, using the fixed point estimate $\hat{\boldsymbol{\theta}}_{old}$ for $\boldsymbol{\Theta}$ that was obtained in the previous step:

$$\begin{aligned} \ln q_2(\underline{\boldsymbol{z}}) = E_{q_1}\left[\ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{z}}, \boldsymbol{\Theta})\right] + \text{const.} = \\ = \ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{z}}, \hat{\boldsymbol{\theta}}_{old}) + \text{const.} \end{aligned} \tag{9.39}$$

After proper normalization, we have

$$q_2(\underline{\boldsymbol{z}}) \propto f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{z}}, \hat{\boldsymbol{\theta}}_{old}) \propto f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}},\boldsymbol{\Theta}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{old}) \tag{9.40}$$

To find a new point estimate for $\boldsymbol{\Theta}$, we use the general VI solution in Eq. (9.6), and define an objective function where we are free to choose any new location point $\boldsymbol{\theta}'$ for $q_1(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$:

$$\begin{aligned} \mathcal{L}(q) &= E_q\left[\ln \frac{f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}}, \boldsymbol{\Theta})}{q_1(\boldsymbol{\Theta} \mid \boldsymbol{\theta}') q_2(\underline{\boldsymbol{Z}})}\right] = \\ &= E_{q_2}\left[E_{q_1}\left[\ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}}, \boldsymbol{\Theta}) - \ln q_1(\boldsymbol{\Theta} \mid \boldsymbol{\theta}')\right] - \ln q_2(\underline{\boldsymbol{Z}})\right] = \\ &= \underbrace{E_{q_2}\left[\ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}}, \boldsymbol{\theta}')\right]}_{Q(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}_{old})} \underbrace{- E_{q_1}\left[\ln q_1(\boldsymbol{\Theta} \mid \boldsymbol{\theta}')\right]}_{h(\boldsymbol{\Theta})} \underbrace{- E_{q_2}\left[\ln q_2(\underline{\boldsymbol{Z}})\right]}_{h(\underline{\boldsymbol{Z}})} \end{aligned} \tag{9.41}$$

Here, the entropy $h(\boldsymbol{\Theta}) = -E_{q_1}\left[\ln q_1(\boldsymbol{\Theta} \mid \boldsymbol{\theta}')\right]$ is constant, regardless of the location parameter $\boldsymbol{\theta}'$, because changing $\boldsymbol{\theta}'$ only translates the position of the peak of q_1 while its shape remains constant. The entropy $h(\underline{\boldsymbol{Z}}) = -E_{q_2}\left[\ln q_2(\underline{\boldsymbol{Z}})\right]$ does not involve $\boldsymbol{\theta}'$ at all. Thus, to maximize $\mathcal{L}(q)$ as a function of $\boldsymbol{\theta}'$, we only need to maximize the first term $Q(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}_{old})$. This objective is a function of the new point estimate $\boldsymbol{\theta}'$, which we are free to choose, and it is also a function of the previous fixed value $\hat{\boldsymbol{\theta}}_{old}$, because this

value was used to obtain q_2, which was then used to calculate the expectation $E_{q_2}[\cdot]$. Thus, the optimal choice for the new point estimate is

$$\hat{\boldsymbol{\theta}}_{new} = \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\, Q\left(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}_{old}\right) = \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\, E_{q_2}\left[\ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}}, \boldsymbol{\theta}')\right] \tag{9.42}$$

Now recall that the EM update step was derived in Chapter 7 as

$$\hat{\boldsymbol{\theta}}^{ML}_{new} = \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}_{old}) = \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\, E_{\underline{\boldsymbol{Z}}}\left[\ln P\left[\underline{\boldsymbol{Z}}, \underline{\boldsymbol{x}} \mid \boldsymbol{\theta}'\right] \mid \underline{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{old}\right] \tag{9.43}$$

where $Q(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}_{old})$ is the "EM help function", and the notation $E_{\underline{\boldsymbol{Z}}}\left[\cdot \mid \underline{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{old}\right]$ was used to emphasize that we must use $P\left[\underline{\boldsymbol{Z}} \mid \underline{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{old}\right]$ to calculate the expected value.

In the VI update equation (9.42), we have

$$q_2(\underline{\boldsymbol{z}}) = f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}},\boldsymbol{\Theta}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{old}) \tag{9.44}$$

and

$$\ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}},\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}}, \boldsymbol{\theta}') = \ln f_{\underline{\boldsymbol{X}},\underline{\boldsymbol{Z}}|\boldsymbol{\Theta}}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}} \mid \boldsymbol{\theta}') f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}') \tag{9.45}$$

Thus, the maximization step in (9.42) is exactly the same as the corresponding EM update step, if the prior parameter density is non-informative, i.e., if the density $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}')$ is constant.[2] Thus, the objective function $Q(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}_{old})$, here derived from the VI approach, can be exactly identical to the EM help function defined in the EM procedure. It is interesting to see that both approaches can lead to exactly the same computational algorithm.

The difference here is, of course, that the Bayesian VI approach starts from a model where the parameters in $\boldsymbol{\Theta}$ are considered as random variables, whereas the EM approach only assumes that the parameters have some fixed unknown values. The VI approach is also much more general, as it can also handle several groups of parameters, and the "EM-like" point approximation may be used for all parameters, or only for a subset. Thus, the EM procedure can be seen as a special case of the VI approach.

[2] Actually, a non-constant prior function $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}')$ could also have been applied in the EM procedure, as

$$\hat{\boldsymbol{\theta}}^{MAP}_{new} = \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\left(E_{\underline{\boldsymbol{Z}}}\left[\ln P\left[\underline{\boldsymbol{Z}}, \underline{\boldsymbol{x}} \mid \boldsymbol{\theta}'\right] \mid \underline{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{old}\right] + \ln f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}')\right)$$

resulting in MAP rather than Maximum-Likelihood estimates.

Summary

This chapter introduced *Variational Inference* (VI) for approximate Bayesian learning. The VI approach is a general method for finding iterative algorithms to estimate an approximate posterior distribution for unknown parameters and hidden model variables $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M)$, given a set of training observations $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$.

- The result of VI is a density function $q_{\underline{\boldsymbol{Z}}}$, which is a good approximation to the exact posterior density function $f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}$,

$$q_{\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{z}}) \approx f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{x}})$$

- The mathematical form of $q_{\underline{\boldsymbol{Z}}}$ can be chosen by the experimenter. Parametric and factorized forms

$$q_{\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{z}}) = \prod_{m=1}^{M} q_m(\boldsymbol{z}_m; \boldsymbol{\theta}_m)$$

  are common.

- The approximation is guaranteed to be optimal in the sense that the Kullback-Leibler divergence $\mathrm{KL}\left(q_{\underline{\boldsymbol{Z}}} \,\|\, f_{\underline{\boldsymbol{Z}}|\underline{\boldsymbol{X}}}\right)$ is minimized, within the constraints imposed by the chosen mathematical form for the approximation.

- The VI approach can lead to computationally efficient methods even when the probabilistic model is highly complex.

- The *Expectation Maximization* (EM) algorithm can be seen as a special case of the much more general VI approach.

- Other approximate methods for Bayesian learning exist, including *sampling* methods and *Expectation Propagation* (EP).