CS M146 - Week 9

Xinzhu Bei

xzbei@cs.ucla.edu

March 10, 2018

Overview

- Kmeans, Kmedoids, GMM
- Bayes Learning, MAP and MLE
- Naive Bayes Classifer
- HMM
- Advance topic: EM

Adaboost - weak learner

- First, we concretely define a **weak classifier**:

$$h_t \colon \mathbb{R}^d \to \{-1, +1\} \tag{2}$$

- A weak classifier must work better than chance. In the two-class setting this means it has less than 50% error and this is easy; if it would have higher than 50% error, just flip the sign. So, we want only a classifier that does not have exactly 50% error (since these classifiers would add no information).
- The error rate of a weak classifier $h_t(\mathbf{x})$ is calculated empirically over the training data:

$$\epsilon(h_t) = \frac{1}{m} \sum_{i=1}^{m} \delta(h_t(x_i) \neq y_i) < \frac{1}{2} \ . \tag{3}$$

** Borrowed from here

Adaboost - Convergence

- Key Idea: **AdaBoost minimizes an upper bound on the classification error**.
- Claim: After t steps, the error of the strong classifier is bounded above by quantity Z, as we just defined it (the product of the data weight normalization factors):

$$\mathrm{Err}(H) \leq Z = Z(\alpha, h) = Z_t(\alpha_t, h_t) \ldots Z_1(\alpha_1, h_1) \tag{28}$$

- AdaBoost is a greedy algorithm that minimizes this upper bound on the classification error by choosing the optimal h_t and α_t to minimize Z_t at each step.

$$(h, \alpha)^* = \operatorname{argmin} Z(\alpha, h) \tag{29}$$

$$(h_t, \alpha_t)^* = \operatorname{argmin} Z_t(\alpha_t, h_t) \tag{30}$$

- As Z goes to zero, the classification error goes to zero. Hence, it converges. (But, we need to account for the case when no new weak classifier has an error rate better than 0.5, upon which time we should stop.)

** Borrowed from here

K-means algorithm
a.k.a Llyod's algorithm

- Step 0: randomly assign the cluster centers $\{\mu_k\}$
- Step 1: Minimize J over $\{r_{nk}\}$ -- Assign every point to the closest cluster center

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

- Step 2: Minimize J over $\{\mu_k\}$ -- update the cluster centers

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}}$$

- Loop until it converges

GMM

- Let θ represent all parameters $\{w_k, \mu_k, \Sigma_k\}$
- Step 0: Initialize θ with some values (random or otherwise)
- Step 1: Compute $\gamma_{nk} = p(z_n = k|\mathbf{x}_n)$ using the current θ

$$p(z_n = k|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|z_n = k)p(z_n = k)}{p(\boldsymbol{x}_n)} = \frac{\overset{\boxed{N(x|\mu_k,\Sigma_k)}}{p(\boldsymbol{x}_n|z_n = k)}\overset{\boxed{\omega_k}}{p(z_n = k)}}{\sum_{k'=1}^{K} p(\boldsymbol{x}_n|z_n = k')p(z_n = k')}$$

- Step 2: Update θ using the just computed γ_{nk}

$$\omega_k = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}}, \quad \boldsymbol{\mu}_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} \boldsymbol{x}_n$$

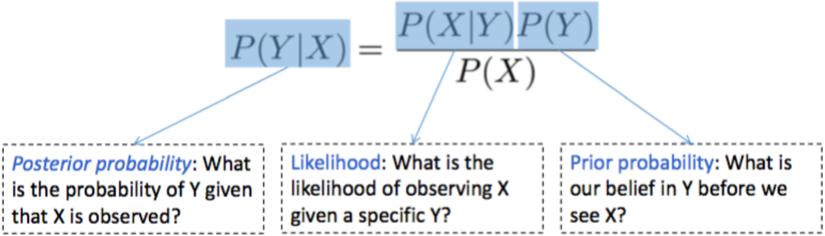
$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

- Loop until converge

Kmeans v.s. GMM

	K means	GMM
Supervision	Unsupervised	Unsupervised
Model	Deterministic	Probabilistic
Hyperparameter	k	k
Objective	$\min \sum_{n,k} \gamma_{nk} \Vert x_n - \mu_k \Vert_2^2$	$\max p(x \mid \theta) = \prod_n p(x_n \mid \theta)$ $= \prod_n \sum_k p(z_n = k) p(x_n \mid z_n = k, \theta_k)$
Parameter	γ_{nk}, μ_k $\gamma_{nk} \in \{0, 1\}$ $\mu_k = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}}$ -	w_k, μ_k, Σ_k $\gamma_{nk} = p(z_n = k \mid x_n)$ $\mu_k = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}}$ w_k, Σ_k
Testing	$k = \arg\min_j \Vert x_n - \mu_j \Vert_2^2$	$k = \arg\max_j p(Z_n = j \mid x_n)$

Bayes Theorem



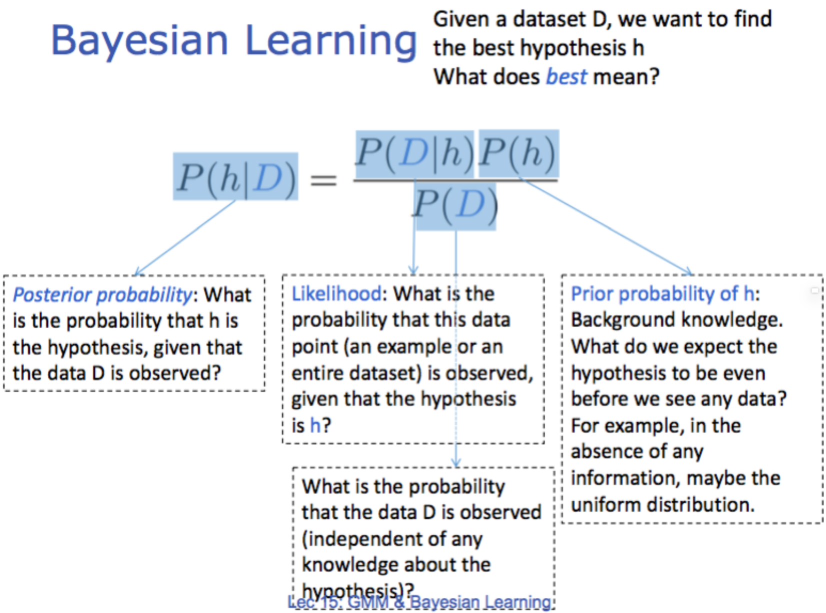
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Posterior / Likelihood £ Prior



Bayesian Learning

Given a dataset D, we want to find the best hypothesis h

What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Likelihood: What is the probability that this data point (an example or an entire dataset) is observed, given that the hypothesis is h?

What is the probability that the data D is observed (independent of any knowledge about the hypothesis)?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Lec 15: GMM & Bayesian Learning

Choosing a Hypothesis

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{1}$$

Given some data, find the most probable hypothesis

- The Maximum a Posteriori (MAP) hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h)p(h) \tag{2}$$

- If we assume that the prior is uniform, i.e. $p(h_i) = p(h_j), \forall i, j$, then we can simplify the above to get the Maximum Likelihood (ML) hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D|h) \tag{3}$$

** Note that it is computationally easier to maximize log likelihood.

Example: Logistic Regression under MAP and MLE

- Maximum Likelihood Estimator to find hypothesis of logistic regression:

$$\begin{aligned}
&\arg\max_h P(D|h) = \arg\max_{\mathbf{w}} \prod_i P(y_i|\mathbf{x}_i, \mathbf{w}) \\
&\arg\max_h \log P(D|h) \\
&\quad = \arg\max_{\mathbf{w}} \sum_i y_i \log \sigma(\mathbf{w}^T\mathbf{x}) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^T\mathbf{x})) \\
&\quad = \arg\max_{\mathbf{w}} - \sum_i \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))
\end{aligned} \tag{4}$$

- Maximum a Posteriori to find hypothesis of logistic regression:

$$\begin{aligned}
&\arg\max_h \log P(D|h) + \log P(h) \\
&\quad = \arg\max_{\mathbf{w}} - \sum_i \log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)) - \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w} \\
&\quad = \arg\min_{\mathbf{w}} \sum_i \log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}
\end{aligned} \tag{5}$$

Naive Bayes Classifier

- Decision Rule: maximizing joint distribution

$$\begin{aligned} h_{NB}(\boldsymbol{x}) &= \operatorname*{argmax}_{y} P(y)P(x_1, x_2, \cdots, x_d|y) \\ &= \operatorname*{argmax}_{y} P(y)\prod_{j} P(x_j|y) \end{aligned}$$

- For binary case, we predict the label to be $+$ if:

$$p(y=+)\prod_{j} p(x_j|y=+) > p(y=-)\prod_{j} p(x_j|y=-)$$

HMM

Properties of Transition matrix and emission matrix

EM Algorithm

Advance reading:
Bishop: Pattern Recognition and Machine Learning Chapter 9

EM Algorithm for Gaussian Mixtures

- The log of the likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right\} \tag{6}$$

- Setting the derivatives with respect to the means $\boldsymbol{\mu}_k$ to zero:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

- Multiplying by $\boldsymbol{\Sigma}^{-1}$ (which we assume to be nonsingular):

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n, \text{ where } N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

- If we set the derivative of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}_k$ to zero:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

EM Algorithm for Gaussian Mixtures

- Finally, we maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k. Here we must take account of the constraint which requires the mixing coefficients to sum to one. This can be achieved using a Lagrange multiplier and maximizing the following quantity

  $$\ln p(X|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

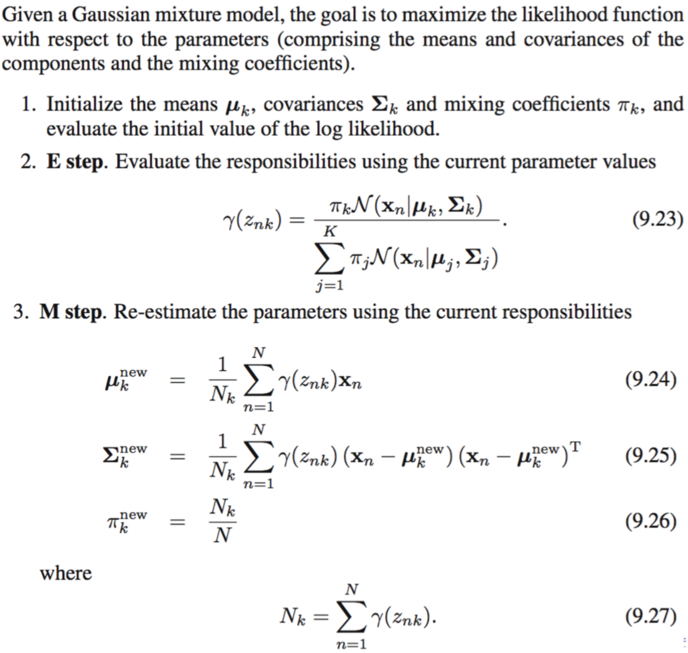
  which gives

  $$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

  If we now multiply both sides by π_k and sum over k making use of the constraint, we find $\lambda = -N$. Using this to eliminate λ and rearranging:

  $$\pi_k = \frac{N_k}{N}$$

EM for Gaussian Mixtures



Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\displaystyle\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \tag{9.25}$$

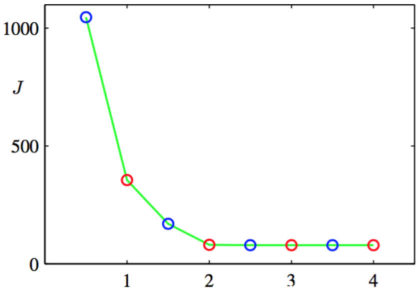
$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

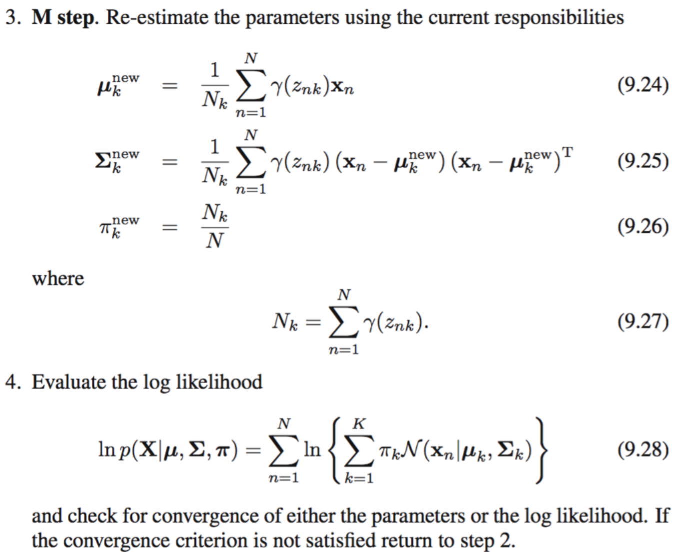
where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{9.27}$$

EM Algorithm for Gaussian Mixtures

Plot of the cost function J given by (9.1) after each E step (blue points) and M step (red points) of the K-means algorithm for the example shown in Figure 9.1. The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors.





3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \tag{9.27}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\} \tag{9.28}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

The General EM Algorithm

- We denote the set of all observed data by $\mathbf{X}$, and the set of all latent variables by $\mathbf{Z}$, the set of all model parameters is denoted by $\boldsymbol{\theta}$, the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

- We shall call $\mathbf{X}, \mathbf{Z}$ the **complete** data set, and we shall refer to the actual observed data $\mathbf{X}$ as **incomplete**.
- In the **E step**, we use the current parameter values θ^{old} to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$.

The General EM Algorithm

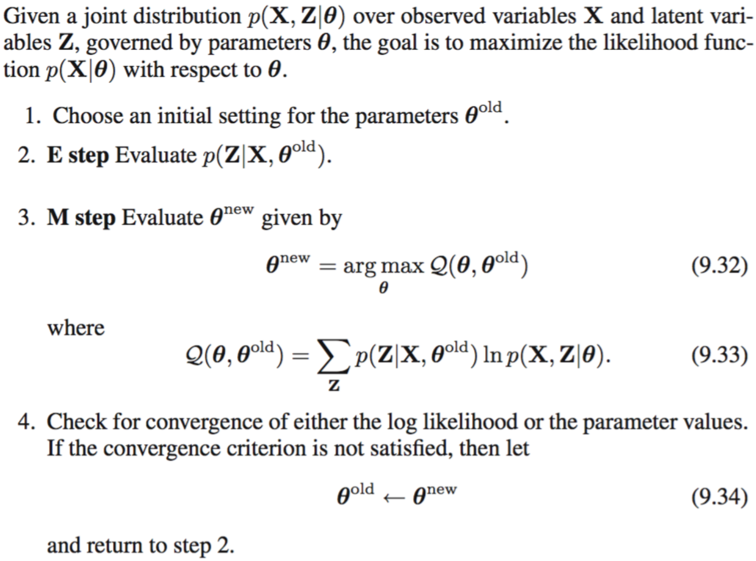
- In the **E step**, we use the current parameter values θ^{old} to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$.
- Because we cannot use the complete-data log likelihood $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, we consider instead its expected value under the posterior distribution of the latent variable, denoted by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

- In the **M step**, we determine the revised parameter estimate $\boldsymbol{\theta}^{new}$ by maximizing this function

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

The General EM Algorithm



Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

and return to step 2.

The End