KTH Electrical Engineering

Solutions to Exam in Pattern Recognition EN2202

Date: Friday, Oct 21, 2011, 14:00 – 19:00

Place: E32, E33.

Allowed: Beta (or corresponding), calculator with empty memory. No notes!

Grades: A: 31p; B: 27p; C: 23p; D: 20p; E: 17; of max 25p + 10p project bonus.

Language: Swedish or English.

Solutions: To be published on the course web page.

Results: Friday, Nov 11.

Review: At KTH-S3/ STEX, Osquldas v. 10.

Contact: Gustav Henter, 070 526 0185

Good Luck!

Please do the **Course Evaluation**! See the course web page.

1 Consider a signal *power spectrum classifier*, where the signal source can belong to one of two categories, here called $S = 1$ and $S = 2$. The two source categories are known to occur with equal probabilities. The classifier is designed to disregard the absolute power levels and classify the spectrum based only on the *relative power distribution* across different non-overlapping frequency bands.

For this purpose, the feature extractor measures three absolute signal power values $(Y_1, Y_2, Y_3)^T$ in the low-, mid-, and high-frequency bands, and then calculates a feature vector

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} Y_1/(Y_1 + Y_2 + Y_3) \\ Y_2/(Y_1 + Y_2 + Y_3) \\ Y_3/(Y_1 + Y_2 + Y_3) \end{pmatrix}$$

Here, feature element X_k is the *relative* power in band k, normalized in relation to the total power. Thus, the feature elements are always non-negative, and the sum of feature elements is always fixed as $\sum_{k=1}^{3} X_k = 1$.

Therefore, it is reasonable to assume that the feature vector has a *Dirichlet* distribution, with conditional density functions specified by three parameters $\boldsymbol{a}_i = (a_{1i}, a_{2i}, a_{3i})^T$, depending on the source category $S = i$:

$$f_{\boldsymbol{X}|S}(x_1, x_2, x_3 \mid i) = \frac{\Gamma(a_{0i})}{\Gamma(a_{1i})\Gamma(a_{2i})\Gamma(a_{3i})} x_1^{a_{1i}-1} x_2^{a_{2i}-1} x_3^{a_{3i}-1}, \quad \text{where } a_{0i} = a_{1i} + a_{2i} + a_{3i},$$

with mean and covariance

$$\begin{aligned} E\left[X_k \mid S = i\right] &= a_{ki}/a_{0i} \\ \operatorname{var}\left[X_k \mid S = i\right] &= \frac{a_{ki}(a_{0i} - a_{ki})}{a_{0i}^2(a_{0i} + 1)} \\ \operatorname{cov}\left[X_k, X_l \mid S = i\right] &= \frac{-a_{ki}a_{li}}{a_{0i}^2(a_{0i} + 1)}, \quad k \neq l. \end{aligned}$$

The gamma function $\Gamma(\,\cdot\,)$ is known and available numerically. In particular, $\Gamma(n) = (n-1)!$ for any positive integer n.

(a) Design an optimal classifier that can guess the source category with minimum error probability. Simplify the classifier to show that optimal performance can be obtained with a single discriminant function, linear in the transformed feature values $\ln x_k$, for any values of the distribution parameters a_{ki} as defined above. (4p)

Solution: As both source alternatives are equally probable, we use the *maximum likelihood* (ML) decision rule. Using the ML rule, we can define a single discriminant function simply as

$$\begin{aligned} g(\boldsymbol{x}) &= \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|1) - \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|2) = \\ &= \ln\Gamma(a_{01}) - \ln\Gamma(a_{02}) + \sum_{k=1}^{3} \left(\ln\Gamma(a_{k2}) - \ln\Gamma(a_{k1}) + (a_{k1} - a_{k2})\ln x_k\right). \end{aligned}$$

This discriminant function is obviously just a linear combination of the transformed feature elements $\ln x_k$, as desired. The optimal classifier decides $S = 1$, whenever $g(\boldsymbol{x}) > 0$, and vice versa.

(b) We now assume that all the distribution parameters are exactly known, as

$$S = 1 : \begin{cases} a_{11} = 4 \\ a_{21} = 5 \\ a_{31} = 6 \end{cases} \qquad S = 2 : \begin{cases} a_{12} = 6 \\ a_{22} = 5 \\ a_{32} = 4 \end{cases}$$

What is the optimal decision if the observed feature vector is $\boldsymbol{x} = (0.3, 0.3, 0.4)^T$)? (1p)

Solution: In this special case the constant terms in the discriminant function sum to zero, because of the symmetry, and $a_{21} - a_{22} = 0$, so the remaining discriminant function with the given parameter values is simply

$$g(\boldsymbol{x}) = (a_{11} - a_{12}) \ln x_1 + (a_{31} - a_{32}) \ln x_3 = -2 \ln 0.3 + 2 \ln 0.4 > 0.$$

Thus, the classifier must decide $d(\boldsymbol{x}) = 1$ for this observation.

2 Determine for each of the following statements whether it is *true* or *false*.
No motivation is required, but you should be certain about your choice. A correct answer gives +1 point, no answer gives 0 points, but an incorrect answer gives −1 point! A negative sum will be counted as 0. The final result can be any integer from 0 to the maximum of (5p).

(a) You observe a sequence of *binary* data symbols $\underline{x} = (x_1, \ldots, x_T)$, where each symbol is $x_t \in \{0, 1\}$, and want to use the observed sequence as training data for an HMM with N states. Then the highest possible log-likelihood for the training sequence will be reached with $N = 2$ states in the HMM, and the log-likelihood cannot get any higher if we initialize the HMM with $N > 2$ states.
Solution: FALSE. The HMM can have any number of states, for example, with different probabilities of repeating the same symbol. More states will usually give higher log-likelihood for the training data (although this may be a result of over-fitting).

(b) A 2-dimensional feature vector $\boldsymbol{X} = (X_1, X_2)^T$ with a probability density function modelled as a Gaussian mixture model (GMM) as

$$f_{\boldsymbol{X}}(x_1, x_2) \propto e^{-(x_1+2)^2-(x_2-2)^2} + e^{-(x_1-2)^2-(x_2+2)^2}$$

has its expected value at the origin, i.e., $E[X_1] = 0$ and $E[X_2] = 0$.
Solution: TRUE. This is a 2-component GMM with equal weights and mean vectors $(-2, +2)^T$ and $(+2, -2)$, i.e., symmetric around the origin, so the overall mean must be at the origin.

(c) For a classification task with N_s source categories and N_d decision alternatives, by modeling the distribution of K-dimensional observed feature vectors using a Gaussian mixture model (GMM) with M components, an optimal classifier can be designed with exactly N_d different discriminant functions.
Solution: TRUE. The number of discriminant functions is determined by N_d and nothing else.

(d) Regarding an observed output sequence $\underline{x} = (x_1, \ldots, x_T)$ from a hidden Markov model λ as an outcome of the random sequence $\underline{X} = (X_1, \ldots, X_T)$, we can use the results of the forward algorithm to calculate the conditional joint probability density

$$P(X_t = x_t \cap X_{t+1} = x_{t+1} \mid X_1 = x_1 \cap \ldots \cap X_{t-1} = x_{t-1}, \lambda)$$

for any $2 \leq t \leq T - 1$.

Solution: TRUE. The forward algorithm can calculate a sequence of scale factors defined as $c_t = P(x_t | x_1, \ldots, x_{t-1}, \lambda)$ for all t. Therefore, using Bayes' rule, we can calculate

$$\begin{aligned} P((x_t, x_{t+1}) \mid (x_1, \ldots, x_{t-1}), \lambda) = \\ = P(x_{t+1} \mid (x_1, \ldots, x_{t-1}, x_t), \lambda)\, P(x_t \mid (x_1, \ldots, x_{t-1}, \lambda) = c_{t+1} c_t. \end{aligned}$$

(e) A hidden Markov model with the following initial state probabilities and state transition probabilities produces an *ergodic* and *stationary* random sequence.

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.99 & 0.01 \\ 0.04 & 0.96 \end{pmatrix};$$

Solution: TRUE. Ergodic because it is irreducible and aperiodic. Stationary, because $A^T q = q$.

3 You can observe some elements of the output sequence $\underline{x} = (x_1, \ldots, x_t, \ldots)$ from a discrete hidden Markov source, but you do not know the corresponding internal state sequence $\underline{S} = (S_1, \ldots, S_t, \ldots)$ in the source. The initial state probability vector is

$$q = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}, \text{with elements } P(S_1 = i).$$

The state transition probability matrix is

$$A = \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix}, \text{with elements } a_{ij} = P(S_{t+1} = j \mid S_t = i).$$

The output probability matrix is

$$B = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{pmatrix}, \text{with elements } b_{ik} = P(X_t = k \mid S_t = i).$$

(a) Calculate $P(X_7 = 1 \mid S_6 = 2)$. (2p)
Solution:

$$\begin{aligned} P(X_7 = 1 \mid S_6 = 2) &= \sum_{i=1}^{2} P(X_7 = 1 \cap S_7 = i \mid S_6 = 2) = \\ &= \sum_{i=1}^{2} P(S_7 = i \mid S_6 = 2) P(X_7 = 1 \mid S_6 = 2 \cap S_7 = i) = \\ &= \sum_{i=1}^{2} a_{2i} b_{i1} = 0.1 \cdot 0.1 + 0.9 \cdot 0.7 = 0.64 \end{aligned}$$

(b) Calculate $P(S_7 = 2 \mid X_6 = 1 \cap S_6 = 2 \cap X_7 = 1 \cap S_8 = 2 \cap X_8 = 3)$. (3p)
Solution: Given S_6 and S_8, S_7 is statistically independent of X_6 and X_8, so we have

$$\begin{aligned} &P(S_7 = 2 \mid X_6 = 1 \cap S_6 = 2 \cap X_7 = 1 \cap S_8 = 2 \cap X_8 = 3) = \\ &= P(S_7 = 2 \mid S_6 = 2 \cap X_7 = 1 \cap S_8 = 2) = \\ &= \frac{P(S_7 = 2 \cap X_7 = 1 \cap S_8 = 2 \mid S_6 = 2)}{P(X_7 = 1 \cap S_8 = 2 \mid S_6 = 2)} \end{aligned}$$

Now,

$$\begin{aligned} P(X_7 = 1 \cap S_8 = 2 \mid S_6 = 2) &= \sum_{i=1}^{2} P(S_7 = i \cap X_7 = 1 \cap S_8 = 2 \mid S_6 = 2) = \\ &= \sum_{i=1}^{2} a_{2i} b_{i1} a_{i2} \end{aligned}$$

so the desired probability is

$$\begin{aligned} \frac{P(S_7 = 2 \cap X_7 = 1 \cap S_8 = 2 \mid S_6 = 2)}{P(X_7 = 1 \cap S_8 = 2 \mid S_6 = 2)} = = \frac{a_{22} b_{21} a_{22}}{a_{21} b_{11} a_{12} + a_{22} b_{21} a_{22}} = \\ = \frac{0.567}{0.004 + 0.567} \approx 0.993 \end{aligned}$$

4 (Climate Change) You have observed a recorded time series $\underline{x} = (x_1, \ldots, x_T)$ of average yearly temperatures over the T most recent years. To quantify the trend in the data we assume that each observed temperature value x_t is an outcome of a random variable X_t, generated by the following probabilistic model:

$$X_t = x_0 + w \cdot (t-1) + R(t).$$

Here x_0 is assumed to be exactly known as the average temperature before year $t = 1$, w is an unknown parameter defining the mean temperature change per year during the recent years $1 \leq t \leq T$, and $R(t)$ is a Gaussian random variable describing the unpredictable temperature deviations from the general trend. We assume that the mean and variance are exactly known as $E[R(t)] = 0$ and $\text{var}[R(t)] = \sigma_R^2$ for all t, and that deviations $R(t_1)$ and $R(t_2)$ are statistically independent for any $t_1 \neq t_2$.

We regard w as an outcome of a random variable W, because we want to apply Bayesian learning for this parameter, using the observed sequence $\underline{x}$.

(a) Let us now assume a Gaussian prior probability density for the unknown parameter, i.e.,

$$f_W(w) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(w-\mu_0)^2}{2\sigma_0^2}}$$

To make this prior non-informative, we can later assign $\mu_0 = 0$ and apply the asymptotic limit $\sigma_0 \to \infty$ for the hyperparameters. Show that the Gaussian density is a *conjugate prior* in this case, i.e., show that the posterior density also has a Gaussian form,

$$f_{W|\underline{X}}(w \mid \underline{x}) = \frac{1}{\sigma_T\sqrt{2\pi}} e^{-\frac{(w-\mu_T)^2}{2\sigma_T^2}}$$

(2p)
Hint: To show that the posterior density is Gaussian, it is not necessary to find precise expressions for the updated hyperparameters μ_T and σ_T^2.
Solution: The posterior density for the parameter has the form

$$f_{W|\underline{X}}(w \mid \underline{x}) \propto f_W(w) f_{\underline{X}|W}(\underline{x} \mid w) \propto e^{-\frac{(w-\mu_0)^2}{2\sigma_0^2}} \prod_{t=1}^{T} e^{-\frac{(x_t - x_0 - w(t-1))^2}{2\sigma_R^2}}$$

This is obviously just an exponential function of a second-degree polynomial in w, which is exactly the form of a Gaussian density function. Thus, the posterior density has a Gaussian form.

(b) Determine the posterior hyperparameters μ_T and σ_T^2 for the parameter W, expressed in terms of the observed data sequence $\underline{x}$. (2p)
Solution: We now only need to identify the mean and variance parameters of the Gaussian posterior density. It must have the following form:

$$f_{W|\underline{X}}(w \mid \underline{x}) \propto e^{-\frac{(w-\mu_T)^2}{2\sigma_T^2}} = e^{-\frac{w^2}{2\sigma_T^2} + \frac{2w\mu_T}{2\sigma_T^2} - \frac{\mu_T^2}{2\sigma_T^2}}$$

Starting again from the the form of the Gaussian posterior we just derived in the (a) problem above, we identify the w^2 and w terms in the exponent as

$$f_{W|\underline{X}}(w \mid \underline{x}) \propto e^{-w^2\left(\frac{1}{2\sigma_0^2}+\frac{1}{2\sigma_R^2}\sum_{t=1}^{T}(t-1)^2\right)+2w\left(\frac{\mu_0}{2\sigma_0^2}+\sum_{t=1}^{T}\frac{(x_t-x_0)(t-1)}{2\sigma_R^2}\right)+\ldots}$$

In these two expressions for $f_{W|\underline{X}}(w \mid \underline{x})$, we can now identify the coefficients for w^2 and w as

$$\begin{aligned}\frac{1}{\sigma_T^2} &= \frac{1}{\sigma_0^2}+\frac{1}{\sigma_R^2}\sum_{t=1}^{T}(t-1)^2 \\ \frac{\mu_T}{\sigma_T^2} &= \frac{\mu_0}{\sigma_0^2}+\sum_{t=1}^{T}\frac{(x_t-x_0)(t-1)}{\sigma_R^2}\end{aligned}$$

yielding the solution

$$\begin{aligned}\sigma_T^2 &= \left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma_R^2}\sum_{t=1}^{T}(t-1)^2\right)^{-1} \\ \mu_T &= \sigma_T^2\left(\frac{\mu_0}{\sigma_0^2}+\sum_{t=1}^{T}\frac{(x_t-x_0)(t-1)}{\sigma_R^2}\right)\end{aligned}$$

If we now want to apply a non-informative prior density, by approaching the asymptotic limit $\sigma_0 \to \infty$, the posterior parameters depend only on the observed data:

$$\begin{aligned}\sigma_T^2 &\to \frac{\sigma_R^2}{\sum_{t=1}^{T}(t-1)^2} \\ \mu_T &\to \frac{\sum_{t=1}^{T}(x_t-x_0)(t-1)}{\sum_{t=1}^{T}(t-1)^2}\end{aligned}$$

(c) Determine the mean $E\left[X_{T+1} \mid \underline{x}\right]$ and variance $\mathrm{var}\left[X_{T+1} \mid \underline{x}\right]$ for the average temperature of the next future year, given the previous observations. Express the result in terms of the given parameters and the posterior hyperparameters. (1p)
Solution: We just apply the original data model $X_t = x_0 + W \cdot (t-1) + R(t)$ at $t = T+1$, and use the previously determined posterior hyperparameters $E\left[W \mid \underline{x}\right] = \mu_T$ and $\mathrm{var}\left[W \mid \underline{x}\right] = \sigma_T^2$, to obtain

$$\begin{aligned}E\left[X_{T+1} \mid \underline{x}\right] &= x_0 + E\left[W \mid \underline{x}\right] \cdot (T+1-1) + E\left[R(T+1)\right] = x_0 + \mu_T \cdot T + 0 \\ \mathrm{var}\left[X_{T+1} \mid \underline{x}\right] &= \mathrm{var}\left[W \mid \underline{x}\right] \cdot (T+1-1)^2 + \mathrm{var}\left[R(T+1)\right] = \sigma_T^2 \cdot T^2 + \sigma_R^2\end{aligned}$$

5 As part of the training of a signal spectrum classifier, we collect as training data a sequence of feature vectors $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, where each observed vector $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tK})^T$ defines the distribution of relative signal power values across frequency bands $k = 1, \ldots, K$. Each observed vector $\boldsymbol{x}_t$ in the sequence is regarded as an outcome of a corresponding random vector $\boldsymbol{X}_t$.

We assume, for simplicity, that feature vectors at different t are statistically *independent* of each other, i.e., the source has *no memory*. However, as the sequence includes spectra of several different types (as in speech or music), we must allow different probability density functions for different types of spectra.

Therefore, we assume that the probability density of any feature vector depends on a hidden random binary indicator vector $\boldsymbol{Z}_t = (Z_{t1}, \ldots, Z_{tM})^T$ with one and only one binary element $Z_{tm} = 1$, and with all other elements $Z_{ti} = 0, i \neq m$, if the spectrum is of type m. Since we have assumed that the source has no memory, the indicator vectors at different t must be statistically independent of each other.

Using this model and notation, the conditional density for the complete observed sequence, given the hidden indicator vectors, can be expressed as

$$f_{\underline{\boldsymbol{X}}|\underline{\boldsymbol{Z}}}(\underline{\boldsymbol{x}} \mid \underline{\boldsymbol{z}}) = \prod_{t=1}^{T} \prod_{m=1}^{M} b_m(\boldsymbol{x}_t)^{z_{tm}}$$

Now we assume that all M component density functions $b_m(\,\cdot\,)$ are of the Dirichlet type (see Problem 1) with parameters exactly known from previous training data sets. All we need to do is to determine the probability mass distribution for the hidden indicator vectors,

$$P[Z_{tm} = 1] = w_m.$$

Apply the well-known Expectation Maximization (EM) procedure to determine a step-wise update equation for the unknown parameter vector $\boldsymbol{w} = (w_1, \ldots, w_M)$, using the observed sequence. Each step in the EM procedure should maximize the help function

$$Q(\boldsymbol{w}', \boldsymbol{w}) = E_{\underline{\boldsymbol{Z}}} \left[\ln P(\underline{\boldsymbol{Z}}, \underline{\boldsymbol{x}} \mid \boldsymbol{w}') \mid \underline{\boldsymbol{x}}, \boldsymbol{w} \right].$$

(5p)

Solution: This is just a general mixture model, with component densities of the Dirichlet type. The model implies that each vector $\boldsymbol{x}_t$ in the sequence is the result of two random steps: First, a value of the hidden binary switch vector $\boldsymbol{z}_t$ is generated at random to point to, e.g., the m-th component, if $z_{tm} = 1$. Next, the observable vector $\boldsymbol{x}_t$ is drawn at random from the selected distribution with density $b_m(\,\cdot\,)$.

The given probability mass distribution of the binary switch vector elements, assuming $\boldsymbol{w}$ had been known,

$$P(Z_{tm} = 1 \cap Z_{t,k\neq m} = 0 \mid \boldsymbol{w}) = w_m$$

can be expressed equivalently as

$$P(\boldsymbol{Z}_t = \boldsymbol{z}_t \mid \boldsymbol{w}) = \prod_{m=1}^{M} w_m^{z_{tm}}$$

for any t. Thus, the combined probability, and log-probability, that we must use in the EM algorithm, is

$$P(\underline{\boldsymbol{Z}} = \underline{\boldsymbol{z}} \cap \underline{\boldsymbol{X}} = \underline{\boldsymbol{x}} \mid \boldsymbol{w}') = \prod_{t=1}^{T} \prod_{m=1}^{M} b_m(\boldsymbol{x}_t)^{z_{tm}} {w'_m}^{z_{tm}}$$

$$\ln P(\underline{\boldsymbol{Z}} = \underline{\boldsymbol{z}} \cap \underline{\boldsymbol{X}} = \underline{\boldsymbol{x}} \mid \boldsymbol{w}') = \sum_{t=1}^{T} \sum_{m=1}^{M} z_{tm} \ln b_m(\boldsymbol{x}_t) + z_{tm} \ln w'_m = \sum_{t=1}^{T} \sum_{m=1}^{M} \ln b_m(\boldsymbol{x}_t) + \ln w'_m$$

To calculate the expected value in $Q(\boldsymbol{w}', \boldsymbol{w})$, we must also find the conditional probability mass for all Z_{tm}, given the observed sequence, using the previous estimate parameter vector $\boldsymbol{w}$:

$$\gamma_{tm} = P[Z_{tm} = 1 \mid \underline{\boldsymbol{x}}, \boldsymbol{w}]$$

Using the statistical independence across different t, we have

$$P(\underline{\boldsymbol{Z}} = \underline{\boldsymbol{z}} \cap \underline{\boldsymbol{X}} = \underline{\boldsymbol{x}} \mid \boldsymbol{w}) = \prod_{t=1}^{T} \prod_{m=1}^{M} b_m(\boldsymbol{x}_t)^{z_{tm}} {w_m}^{z_{tm}}$$

$$P(\boldsymbol{Z}_t = \boldsymbol{z}_t \cap \boldsymbol{X}_t = \boldsymbol{x}_t \mid \boldsymbol{w}) = \prod_{m=1}^{M} b_m(\boldsymbol{x}_t)^{z_{tm}} {w_m}^{z_{tm}}$$

$$P(Z_{tm} = 1 \cap \boldsymbol{X}_t = \boldsymbol{x}_t \mid \boldsymbol{w}) = w_m b_m(\boldsymbol{x}_t)$$

$$\gamma_{tm} = P(Z_{tm} = 1 \mid \boldsymbol{X}_t = \boldsymbol{x}_t, \boldsymbol{w}) = \frac{w_m b_m(\boldsymbol{x}_t)}{\sum_{i=1}^{M} w_i b_i(\boldsymbol{x}_t)}$$

Now, finally, the EM help function can be written as

$$Q(\boldsymbol{w}', \boldsymbol{w}) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tm} \left(\ln b_m(\boldsymbol{x}_t) + \ln w'_m \right)$$

We must now select $\boldsymbol{w}'$ to maximize this help function, with the constraint that $\sum_{i=1}^{M} w'_i = 1$. For this purpose we apply a Lagrange multiplier and formulate an extended criterion function

$$F(\boldsymbol{w}', \boldsymbol{w}) = \lambda(1 - \sum_{i=1}^{M} w'_i) + \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tm} \left(\ln b_m(\boldsymbol{x}_t) + \ln w'_m \right)$$

A necessary condition for maximum is

$$0 = \frac{\partial F}{\partial w'_m} = -\lambda + \sum_{t=1}^{T} \gamma_{tm} \frac{1}{w'_m}, \quad \text{for all } m$$

with solution

$$w'_m = \frac{1}{\lambda} \sum_{t=1}^{T} \gamma_{tm}$$

The required constraint is obviously fulfilled if we choose

$$\lambda = \sum_{t=1}^{T} \sum_{i=1}^{M} \gamma_{ti} = T$$

Thus, the EM update rule is

$$w_m^{new} \leftarrow w'_m = \frac{1}{T} \sum_{t=1}^{T} \gamma_{tm} = \frac{1}{T} \sum_{t=1}^{T} \frac{w_m b_m(\boldsymbol{x}_t)}{\sum_{i=1}^{M} w_i b_i(\boldsymbol{x}_t)}$$

As expected, this result is exactly analogous to the similar estimation procedure for the weight parameters in a GMM.