Maximum Posterior Estimation for a Mixture Model with Real and Categorical Features

We have a set of N examples with real features x_n, and categorical features y_n. The hidden variable z_n will the denote the protype membership. The likelihood for the observations given the parameters θ

Observations

- $\{x_n^r\}$: Real valued feature r for example n
- $\{y_n^c\}$: Categorical valued feature c for example n

Latent States

- $\{z_n\}$: Prototype membership for example n

Model Parameters θ

- $\{\pi_k\}$: Prior probability an example belonging to prototype k.
- $\{\mu_k^r\}$: Observation mean for feature r in prototype k.
- $\{\lambda_k^r\}$: Observation precision (1/variance) for feature r in prototype k.
- $\{\rho_k^c\}$ Categorical probabilities for feature c in in prototype k

Likelihood

$$
\begin{aligned}
p(x, y|\theta) &= \sum_z p(x, y|z, \theta) \\
&= \prod_n \sum_{z_n} \prod_k [p(x_n|\theta_k) p(y_n|\theta_k) p(z_n)]^{z_{n,k}}
\end{aligned}
$$

$$
\begin{aligned}
p(x_n|\theta_k) &= \prod_r N(x_n^r|\mu_k^r, \lambda_k^r) \\
&= \prod_r (\lambda_k^r/2\pi)^{1/2} \exp[-\tfrac{1}{2}\lambda_k^r(x-\mu_k^r)^2] \\
p(y_n|\theta_k) &= \prod_c \mathrm{Categ}(y_n^c|\rho_k^c) \\
&= \prod_c \prod_d (\rho_{k,d}^c)^{y_{n,d}^c} \\
p(z_n) &= \prod_k \pi_k^{z_{n,k}}
\end{aligned}
$$

Priors

$$
\begin{aligned}
p(\theta) &= p(\pi)\prod_k \prod_r p(\mu_k^r, \lambda_k^r) \prod_c p(\rho_k^c) \\
\pi &\sim \mathrm{Dir}(\pi_0) \\
\lambda_k^r &\sim \mathrm{Gamma}(a_k^r, b_k^r) \\
\mu_k^r &\sim \mathrm{N}(m_k^r, \beta_k^r \lambda_k^r) \\
\rho_k^c &\sim \mathrm{Dir}(\alpha_k^c)
\end{aligned}
$$

Conjugate Exponential Form

Likelihood

$$
\begin{aligned}
p(x_n|\eta_k) &= \prod_r p(x_n^r|\eta_k^r) \\
&= \prod_r h(x_n^r) g(\eta_k^r) \exp[\eta_k^r \cdot u(x_n^r)] \\
p(y_n|\eta_k) &= \prod_c p(y_n^c|\eta_k^c) \\
&= \prod_c h(y_n^c) g(\eta_k^c) \exp[\eta_k^c \cdot u(y_n^c)]
\end{aligned}
$$

Prior

$$
\begin{aligned}
p(\eta_k|\nu_k, \chi_k) &= \prod_r p(\eta_k^r|\nu_k^r, \chi_k^r) \prod_c p(\eta_k^c|\nu_k^c, \chi_k^c) \\
&= \prod_r f(\nu_k^r, \chi_k^r) g(\eta_k^r)^{\nu_k^r} \exp[\eta_k^r \cdot \chi_k^r] \\
&\quad \prod_c f(\nu_k^c, \chi_k^c) g(\eta_k^c)^{\nu_k^c} \exp[\eta_k^c \cdot \chi_k^c]
\end{aligned}
$$

Real Features

- $h(x_n) = 1$
- $g(\eta_k^r) = (\eta_{k,1}^r/2\pi)^{1/2} \exp[-(\eta_{k,2}^r)^2/2\eta_{n1}^r]$
- $\eta_k^r = \{\lambda_k^r, \lambda_k^r \mu_k^r\}$
- $u(x_n) = \{-\frac{1}{2}x_n^2, x_n\}$
- $\nu_k^r = \beta_k^r = 2a_k^r - 1$
- $\chi_k^r = \{-\frac{1}{2}(\beta_k^r (m_k^r)^2 + 2b_k^r), \beta_k^r m_k^r\}$

Categorical Features

- $h(y_n) = 1$
- $g(\eta_k^c) = 1$

- $\eta_k^c = \{\log \rho_{k,d}^c\}$
- $u(y_n) = \{y_{n,d}\}$
- $\nu_k^c = 1$
- $\chi_k^c = \{\alpha_{k,d}^c - 1\}$

Maximum Posterior Estimation

For maximum likelihood esitmation, we optimize define a log likelihood L, defined as:

$$L^{\text{ml}} = \log p(x, y|\eta)$$

to find

$$\eta^{\text{ml}} = \arg\max_{\eta} \log p(x, y|\eta) = \arg\max_{\eta} L^{\text{ml}}$$

In maximum posterior estimation, we optimize

$$\begin{aligned} L^{\text{map}} &= \log p(x, y, \eta) \\ &= \log p(x, y|\eta) + \log p(\eta) \\ &= L^{\text{ml}} + L^{\text{prior}} \end{aligned}$$

to obtain

$$\begin{aligned} \eta^{\text{map}} &= \arg\max_{\eta} \log p(\eta|x, y) \\ &= \arg\max_{\eta} \log \left[\frac{p(x, y|\eta)p(\eta)}{p(x, y)} \right] \\ &= \arg\max_{\eta} \log[p(x, y|\eta)p(\eta)] = \arg\max_{\eta} L^{\text{map}} \end{aligned}$$

In order to maximize L^{map} with respect to the parameters η we have to solve the set of equations:

$$\frac{\partial L^{\text{map}}}{\partial \eta_{k,i}^{\{r,c\}}} = \frac{\partial \left(L^{\text{ml}} + L^{\text{prior}} \right)}{\partial \eta_{k,i}^{\{r,c\}}} = 0$$

For each real feature r we must solve for two variables $i = 1, 2$. For each categorical feature c we must solve for D^c variables $d = 1, \ldots, D^c$.

Real Features

The partial derivatives of L^{ml} expand to:

$$\begin{aligned} \frac{\partial L^{\text{ml}}}{\partial \eta_{k,i}^r} &= \sum_n \frac{p(x_n|\eta_k^r)p(y_n|\eta_k^c)\pi_k}{\sum_l p(x_n|\eta_l^r)p(y_n|\eta_l^c)\pi_l} \left[\frac{1}{g} \frac{\partial g(\eta_k^r)}{\partial \eta_{k,i}^r} + u_i(x_n) \right] \\ &= \sum_n \gamma_{nk} \left[\frac{1}{g} \frac{\partial g(\eta_k^r)}{\partial \eta_{k,i}^r} + u_i(x_n) \right] \end{aligned}$$

with the responsibilities γ_{nk} defined by:

$$\gamma_{nk} = \frac{p(x_n|\eta_k^r)p(y_n|\eta_k^c)\pi_k}{\sum_l p(x_n|\eta_l^r)p(y_n|\eta_l^c)\pi_l}$$

The derivatives of L^{prior} are given by:

$$\frac{\partial L^{\text{prior}}}{\partial \eta_{k,i}^r} = \frac{\nu_k^r}{g}\frac{\partial g}{\partial \eta_{k,i}^r} + \chi_{k,i}^r$$

Adding both terms together, the condition $\partial L^{\text{map}}/\partial \eta_{k,i}^r = 0$ becomes:

$$0 = \frac{\nu_k^r + N_k}{g}\frac{\partial g}{\partial \eta_{k,i}^r} + \chi_{k,i}^r + \sum_n \gamma_{nk} u_i(x_n^r) \qquad\qquad N_k = \sum_n \gamma_{nk}$$

This can be interpreted as a weighted average of the prior for the sufficient statistics $\tilde{\chi}$ and averaged sufficient statistics of the data:

$$\frac{1}{g}\frac{\partial g(\eta_k^r)}{\partial \eta_{k,i}^r} = -\frac{1}{\nu_k^r + N_k}\left[\nu_k^r \tilde{\chi}_{k,i}^r + N_k \left\langle u_i(x_n^r)\right\rangle_{\gamma_{nk}}\right]$$

with

$$\tilde{\chi}_{k,i}^r = \frac{\chi_{k,i}^r}{\nu_k^r} \qquad\qquad \left\langle u_i(x_n^r)\right\rangle_{\gamma_{nk}} = \frac{1}{N_k}\sum_n \gamma_{nk} u_i(x_n^r)$$

If we now substitute the expressions for η_k^r, g and χ_k^r given above, we obtain:

$$(\lambda_k^r)^{-1} + (\mu_k^r)^2 = \frac{1}{\nu_k^r + N_k}\left[\beta_k^r (m_k^r)^2 + 2b_k^r + N_k \left\langle (x_n^r)^2\right\rangle_{\gamma_{nk}}\right]$$

and

$$\mu_k^r = \frac{1}{\nu_k^r + N_k}\left[\beta_k^r m_k^r + N_k \left\langle x_n^r\right\rangle_{\gamma_{nk}}\right]$$

If we now wish to consider the case where we have a prior only on λ and not on μ, we set $a > 1/2$, and $\beta = 0$. We then substitute $\nu_k \mapsto \beta_k$ in the equation for μ_k and $\nu_k \mapsto 2a_k - 1$ in the equation for λ, to obtain:

$$\begin{aligned}
\mu_k^r &= \langle x_n^r\rangle_{\gamma_{nk}} \\
(\sigma_k^r)^2 = (\lambda_k^r)^{-1} &= \frac{1}{2a_k - 1 + N_k}\left[b_k^r + N_k \left\langle (x_n^r)^2\right\rangle_{\gamma_{nk}}\right] - \langle x_n^r\rangle_{\gamma_{nk}}^2
\end{aligned}$$

Categorical Features

For the categorical variables we must introduce a Lagrange multiplier to enforce the constraint $\sum_d \rho_{k,d}^c = \sum_d \exp[\eta_{k,d}^c] = 1$.

$$\begin{aligned}
0 &= \frac{\partial}{\partial \eta_{k,d}^c}\left[L^{\text{map}} + \lambda\left(1 - \sum_e \exp[\eta_{k,e}^c]\right)\right] \\
&= \frac{\partial L^{\text{map}}}{\partial \eta_{k,d}^c} - \lambda \exp[\eta_{k,d}^c]
\end{aligned}$$

since $g(\eta_k^c) = 1$ and $\nu_k^c = 1$, the expression for the derivative of L^{map} reduces to:

$$\frac{\partial L^{\text{map}}}{\partial \eta_{k,d}^c} = \chi_{k,d}^c + \sum_n \gamma_{nk} u_i(y_n^c)$$

The solution to the constrained equation is given, up to a normalisation λ by

$$\rho_{k,d}^c = \exp[\eta_{k,d}^c] = \frac{1}{\lambda}\left(\chi_{k,d}^c + \sum_n \gamma_{nk} y_{n,d}\right)$$

Whereas λ can be found by noting that:

$$1 = \sum_d \exp[\eta_{k,d}^c] = \frac{1}{\lambda}\sum_d\left(\chi_{k,d}^c + \sum_n \gamma_{nk} y_{n,d}\right)$$

So

$$\lambda = \sum_d\left(\chi_{k,d}^c + \sum_n \gamma_{nk} y_{n,d}\right)$$

Note that seting $\chi_{k,d}^c = 0$ reduces the updates to the maximum likelihood case.

Algorithm

Repeat until L converges:

- Calculate γ_{nk} using parameters θ^i. Probabilities for missing features are set to 1.
- Calculate updates θ^{i+1} from γ_{nk}, x_n and y_n.