1. Write down the EM algorithm (i.e. the E-step and M-step) to get Maximum Likelihood Estimates in a Poisson Mixture Model (this is useful to cluster features where counts are measured). You should assume that features are independent in each mixture component. Be as complete as possible in showing your derivation.

For reference, the probability mass function of the Poisson distribution is:

\[ p(x; \lambda) = \frac{\lambda^x}{x!} \exp\{-\lambda\} \]

and for a \(D\)-dimensional multivariate Poisson with independent features:

\[ p(x; \lambda) = \prod_{d=1}^{D} \frac{\lambda^{x_d}}{x_d!} \exp\{-\lambda_d\} \]

Solution:

Note that the MLE for the Poisson mixture model with assignments \(y\) are given by maximizers of

\[ L(\lambda, \theta; X, y) = \log p(x, y; \lambda, \theta) = \sum_n \left[ \log \theta_{y_n} + \sum_{d=1}^{D} x_{nd} \log \lambda_{y_n}^d - \log x_{nd}! - \lambda_{y_n}^d \right] \]

For \(\lambda_d^k\) (Poisson parameter for cluster \(k\)) we have

\[ \frac{\partial L}{\partial \lambda_d^k} = \frac{1}{\lambda_d^k} \sum_n I\{y_n = k\} x_{nd} - \sum_n I\{y_n = k\} \]

which implies \(\hat{\lambda}_d^k = \frac{\sum_n I\{y_n = k\} x_{nd}}{\sum_n I\{y_n = k\}}\), i.e., the mean of feature \(d\) for examples assigned to cluster \(k\). We should expect that in the M-step of the EM algorithm we would have a weighted mean as the MLE.
For \( \theta \) (mixture proportions) we need to solve:

\[
\max_{\theta_1, \ldots, \theta_K} \sum_n \log \{ \theta_{y_n} \}
\]

\[\text{s.t. } \sum_k \theta_k = 1\]

We construct the lagrangian function \( L(\theta) = \sum_n \log \{ \theta_{y_n} \} - \alpha (\sum_k \theta_k - 1) \) and find

\[
\frac{\partial L}{\partial \theta_k} = \frac{1}{\theta_k} \sum_n I\{y_n = k\} - \alpha
\]

to get \( \hat{\theta}_k = \frac{1}{\alpha} \sum_n I\{y_n = k\} \). To get \( \alpha \) we solve the equality constraint on \( \alpha \):

\[
\sum_k \hat{\theta}_k = 1 \Rightarrow \\
\sum_k \frac{1}{\alpha} \sum_n I\{y_n = k\} = 1 \Rightarrow \\
\alpha = N
\]

So, \( \hat{\theta}_k = \frac{1}{N} \sum_n I\{y_n = k\} \), i.e., the proportion of examples assigned to cluster \( k \).

The E-step works exactly as in the GMM case, we set

\[
z_{nk} = q(y_n = k) = p(y_n = k|x_n; \lambda, \theta) = \frac{p(x_n|y_n = k; \lambda, \theta)\theta_k}{\sum_l p(x_n|y_n = l; \lambda, \theta)\theta_l}
\]

(2)

For the M-step, we solve

\[
\max_{\lambda, \theta} \sum_n \sum_k z_{nk} \left[ \log \{ \theta_k \} + \sum_{d=1}^D x_{nd} \log \{ \lambda_d^k \} - \log \{ x_{nd} \} - \lambda_d^k \right]
\]

\[\text{s.t. } \sum_k \theta_k = 1\]

(3)

For \( \lambda_d^k \) we have
\[
\frac{\partial}{\partial \lambda_d^k} = \frac{1}{\lambda_d^k} \sum_n z_{nk} x_{nd} - \sum_n z_{nk}
\]
which implies \( \lambda_d^k = \frac{\sum_n z_{nk} x_{nd}}{\sum_n z_{nk}} \), or the weighted mean as we expected.

For \( \theta \) we construct the Lagrangian as before
\[
L(\theta) = \sum_n \sum_k z_{nk} \log\{\theta_k\} - \alpha(\sum_k \theta_k - 1)
\]
Taking the derivative:
\[
\frac{\partial L}{\partial \theta_k} = \frac{1}{\theta_k} \sum_n z_{nk} - \alpha
\]
which implies \( \theta_k = \frac{1}{\alpha} \sum_n z_{nk} \). As before solving the equality constraint we get
\[
\sum_k \theta_k = 1 \Rightarrow \\
\sum_k \frac{1}{\alpha} \sum_n z_{nk} = 1 \Rightarrow \\
\alpha = \sum_k \sum_n z_{nk} \Rightarrow \\
\alpha = N
\]
and thus \( \hat{\theta}_k = \frac{1}{N} \sum_n z_{nk} \).