# — *Solution notes* —

**Sixth week practical exercises in Machine learning 1 – 2024 – Paper 1**

## 1 Mixture models (October)

Consider a data distribution whose underlying generating process is a mixture of Poisson distributions, but we do not know the parameters of the mixture model. In this question, you are asked to derive the updated equations for the general Poisson mixture model. The Poisson distribution is:

$$P(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)$$

where $x = 0, 1, 2, ...$ (non-negative integers), $\lambda > 0$ is the 'rate' of the data; the expected value of $x$ is $\lambda$. A mixture representation assumes the following:

$$P(x_n) = \sum_{k=1}^{K} \pi_k P(x_n|\lambda_k)$$

where $P(x_n|\lambda_k)$ is a Poisson distribution with rate $\lambda_k$ and $x_n$ is a single data observation. To answer the following questions assume we are given a dataset $\{x_1, x_2, \ldots, x_N\}$. Make sure that the constraint $\sum_k \pi_k = 1$ is satisfied (i.e. think of the log-likelihood or log-joint as $f$ (an objective to maximize) and $\sum_k \pi_k - 1 = 0$ as $g = 0$ (a constraint that must hold)).

($a$) Write down the likelihood (as usual) for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$ and $\{\lambda_k\}$.

---

*Answer:* The likelihood of using our mixture representation can be written as:

$$L \overset{\text{IID}}{=} \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} exp(-\lambda_k)$$

---

($b$) Write down the log-likelihood (as usual) for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.

---

*Answer:*

The log-likelihood:

$$\log(L) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} exp(-\lambda_k)$$

---

($c$) Let us consider the contribution of each of the Poisson components as their *responsibility* in the generating process, denoting them $r_{nk}$ (i.e. the contribution of the $k$th- poisson distribution for the $n$th datapoint).

This responsibilities are actually the posterior probabilities given by the following expression $p(C_k|\mathbf{x}_n) = r_{nk}$

Find the expression for the responsibilities $r_{nk}$.

---

*Answer:*

By developing the Bayes theorem for the posterior:

$$r_{nk} = \frac{\pi_k P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)}$$

---

($d$)  Find the expression for $\lambda_k$ that maximizes the log-likelihood.

---

*Answer:*  This is solved by setting $\frac{\partial \log L}{\partial \lambda_k} = 0$, write it in terms of $r_{nk}$ and then solve for $\lambda_k$. Let us first compute $\frac{\partial \log L}{\partial \lambda_k}$:

$$
\begin{aligned}
\frac{\partial \log L}{\partial \lambda_k} &= \sum_{n=1}^{N} \frac{1}{\sum_l \pi_l P(x_n|\lambda_l)} \left( \pi_k \frac{x_n}{\lambda_k} P(x_n|\lambda_k) - \pi_k P(x_n|\lambda_k) \right) \\
&= \sum_{n=1}^{N} \frac{\pi_k P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)} \left( \frac{x_n}{\lambda_k} - 1 \right),
\end{aligned}
\tag{1}
$$

where we compute $\partial P(x_n|\lambda_k)/\partial \lambda_k = P(x_n|\lambda_k)(x_n\lambda_k^{-1} - 1)$ using the fact that $P(x_n|\lambda_k)$ has a product of two terms ($\lambda_k^{x_n}$ and $\exp(-\lambda_k)$) that depend on $\lambda$ so we used the product rule of differentiation to obtain ().

Next we identify all instances of $r_{nk} = \frac{\pi_k P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)}$:

$$\frac{\partial \log L}{\partial \lambda_k} = \sum_{n=1}^{N} r_{nk} \left( \frac{x_n}{\lambda_k} - 1 \right).$$

Now we solve $\frac{\partial \log L}{\partial \lambda_k} = 0$ for $\lambda_k$

$$
\begin{aligned}
&\sum_{n=1}^{N} r_{nk} \left( \frac{x_n}{\lambda_k} - 1 \right) = 0 \\
\Leftrightarrow \quad & \sum_n r_{nk} \frac{x_n}{\lambda_k} = \sum_n r_{nk} \\
\Leftrightarrow \quad & \lambda_k = \frac{\sum_n r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}.
\end{aligned}
$$

So $\lambda_k = \sum_n r_{nk} x_n / N_k$ with $N_k = \sum_{n=1}^{N} r_{nk}$.

---

($e$)  Find the expression for $\pi_k$ that maximizes the log-likelihood.

---

*Answer:*  Note that the solution for $\pi_k$ should satisfy the constraint $\sum_k \pi_k = 1$. Hence we obtain the solution via the method of Lagrange multipliers. So first we write down the Lagrangian, then we compute the stationary points w.r.t $\pi_k$ and Lagrange multiplier $\beta$.

The Langrangian is given as follows:

$$\tilde{L} = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k) + \beta(\sum_k \pi_k - 1).$$

Now we find the stationary point for $\pi_k$ (i.e. solve $\frac{\partial \tilde{L}}{\partial \pi_k}$). The derivative of the Langrangian w.r.t $\pi_k$ is given by

$$\frac{\partial \tilde{L}}{\partial \pi_k} = \sum_n \frac{P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)} + \beta$$

$$= \sum_n \frac{r_{nk}}{\pi_k} + \beta.$$

Let us set it to zero and solve for $\pi_k$:

$$\sum_n \frac{r_{nk}}{\pi_k} + \beta = 0$$

$$\Leftrightarrow \qquad \pi_k = -\sum_n \frac{r_{nk}}{\beta}. \qquad (2)$$

The stationary point for $\beta$ is obtained by solving

$$\frac{\partial \tilde{L}}{\partial \beta} = 0$$

$$\Leftrightarrow \qquad \sum_k \pi_k = 1,$$

in which we substitute the result of (2) to obtain:

$$\sum_k -\sum_n \frac{r_{nk}}{\beta} = 1,$$

$$\Leftrightarrow \qquad \beta = -\sum_n 1 = -N,$$

where we note that $\sum_k r_{nk} = 1$. Now that we have an expression for $\beta$ we get our final result from (2) as $\pi_k = \frac{1}{N} \sum_{n=1}^{N} r_{nk}$.

---

(f) Now assume priors for $\pi_k$ and $\lambda_k$, $p(\lambda_k|a,b) = \mathcal{G}(\lambda_k|a,b)$ (a Gamma prior) and $p(\pi_1, \ldots, \pi_k) = \mathcal{D}(\pi_1, \ldots, \pi_k|\alpha/K, \ldots, \alpha/K)$ (a Dirchlet distribution) respectively. These distributions are defined in the appendix of Bishop. Write down the log-joint distribution:

$$\log p(x_1, \ldots, x_N, \{\pi_k\}, \{\lambda_k\}|a, b, \alpha, K).$$

---

*Answer:*

$$\hat{L} = \log L + \sum_k \log \mathcal{G}(\lambda_k|a,b) + \log \mathcal{D}(\{\pi_k\}|\alpha, K)$$

3

whatever is not depending in $\lambda_k$ can be considered as constant C

$$\hat{L} = \log L + \sum_k (a-1)\log \lambda_k - b\lambda_k + \sum_k (\alpha/K - 1)\log \pi_k + C$$

---

$(g)$ Find the expression for $\lambda_k$ that maximizes the log-joint.

---

*Answer:* Same recipe as before but now applied to $\hat{L}$:

$$\frac{\partial \hat{L}}{\partial \lambda_k} = \sum_n r_{nk}(x_n \lambda_k^{-1} - 1) + (a-1)\lambda_k^{-1} - b = 0$$

$$\Leftrightarrow \lambda_k^{-1}(\sum_n r_{nk}x_n + a - 1) - (\sum_n r_{nk} + b) = 0$$

$$\Leftrightarrow \lambda_k = \frac{\sum_n r_{nk}x_n + a - 1}{N_k + b}.$$

---

$(h)$ Find the expression for $\pi_k$ that maximizes the log-joint.

---

*Answer:* In the following $\hat{L}$ denotes the Lagrangian (previous $\hat{L} + \beta(\sum_k \pi_k - 1)$). Find stationary point w.r.t $\pi_k$:

$$\frac{\partial \hat{L}}{\partial \pi_k} = \sum_n \frac{P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)} + \beta + (\alpha/K - 1)\pi_k^{-1} = 0$$

$$\Leftrightarrow \sum_n r_{nk} + \pi_k \beta + \alpha/K - 1 = 0$$

$$\Leftrightarrow \pi_k = -\frac{\sum_n r_{nk} + \alpha/K - 1}{\beta}$$

Stationary point for $\beta$:

$$\sum_k \pi_k = 1$$

$$\Leftrightarrow -\frac{\sum_k \sum_n r_{nk} + \alpha/K - 1}{\beta} = 1$$

$$\Leftrightarrow \beta = -(N + \alpha - K).$$

Hence $\pi_k = \frac{\sum_n r_{nk}+\alpha/K-1}{N+\alpha-K}$.

---

$(i)$ Write down an iterative algorithm using the above update equations (similar to the ones derived in class for the Mixture of Gaussians); include initialization and convergence check steps.

---

*Answer:*

($i$)   Randomly assign data to $K$ clusters (i.e. set hard values for $r_{nk}$). Compute $\pi_k$ and $\lambda_k$ from the initial assignments.

($ii$)   Update (soft) $r_{nk}$ (E-step).

    (A)   Update $\pi_k$ using MLE or MAP estimates from above (M-step for $\pi_k$).

    (B)   Update $\lambda_k$ using MLE or MAP estimates from above (M-step for $\lambda_k$).

    (C)   Compute $L$ (or $\hat{L}$) and repeat E and M steps until $\Delta L < \epsilon$ (or $\Delta \hat{L} < \epsilon$).

---