Seminar 6, Derivation of Discrete Mixture Model For Exponential Family

Evgenii Egorov, egorov.evgenyy@ya.ru

September 21, 2020

The goal of the note to establish the connection between the MLE estimator for a non-mixture model and MLE estimator for each component of the discrete mixture model. Model for K components discrete mixture:

$$\begin{aligned}
&X = \{x_n\}_{n=1}^N,\\
&\theta = \{\lambda_1, \dots, \lambda_k, \pi_1, \dots, \pi_K\},\\
&p(x_n|t_n = k; \theta) = \exp\left(\langle \phi(x_n), \lambda_k \rangle - F(\lambda_k)\right),\\
&F(\lambda_k) = \log \int \exp\left(\langle \phi(x_n), \lambda_k \rangle\right)\, dx,\\
&p(t_n = k) = \pi_k, \forall n.
\end{aligned}$$

Note, that t_n are local variables and θ is global.

1 E-step

E-step is trivial as usual for discrete mixture model:

$$p(t_n = k|x_n; \theta^{\text{old}}) = \frac{\pi_k p(x_n|t_n = k)}{\sum\limits_{k'} \pi_{k'} p(x_n|t_n = k')} = \frac{\pi_k \exp\left(\langle \phi(x_n), \lambda_k \rangle - F(\lambda_k)\right)}{\sum\limits_{k'} \pi_{k'} \exp\left(\langle \phi(x_n), \lambda_{k'} \rangle - F(\lambda_{k'})\right)}.$$

Note, that is more computational stable to estimate the $\log p(t_n = k|x_n)$ matrix and use for denominator sum-log-exp trick.

2 M-step

M-step is more interesting and has the wonderful connection with simple MLE. Let me denote the result of the E-step: $q(t_n = k|x_n; \theta^{\text{old}}) = q_{nk}$.

$$\mathcal{F} = \sum_{n=1}^{N} \langle \log p(x_n, t_n|\theta) \rangle_{q(t_n|x_n)} = \sum_{n=1}^{N} \langle \log p(x_n|t_n) + \log p(t_n) \rangle_{q(t_n|x_n)} =$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \left[\langle \phi(x_n), \lambda_k \rangle - F(\lambda_k) + \log \pi_k \right] = \sum_{k=1}^{K} \left\{ \left\langle \sum_{n=1}^{N} q_{nk} \phi(x_n), \lambda_k \right\rangle + \left(\sum_{n=1}^{N} q_{nk} \right) (\log \pi_k - F(\lambda_k)) \right\}.$$

Derivation for π_k is common for any discrete mixture model:

$$\nabla_{\pi_k} \left(\mathcal{F} + \lambda(1 - \sum_{k=1}^{K} \pi_k) \right) = 0, \left(\sum_{n=1}^{N} q_{nk} \right) \pi_k^{-1} - \lambda = 0, \pi_k = \frac{1}{\lambda} \sum_{n=1}^{N} q_{nk}.$$

$$1 = \sum_{k=1}^{K} \pi_k = \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} = \frac{N}{\lambda}, \lambda = N.$$

Hence, we get simple result:

$$\boxed{\pi_k = \frac{1}{N} \sum_{n=1}^{N} q_{nk}} = \frac{1}{N} \sum_{n=1}^{N} p(t_n = k|x_n; \theta^{\text{old}}).$$

Finally, we get to the interesting part:

$$\nabla_{\lambda_k} \mathcal{F} = \nabla_{\lambda_k} \left\{ \left\langle \sum_{n=1}^{N} q_{nk} \phi(x_n), \lambda_k \right\rangle + \left(\sum_{n=1}^{N} q_{nk} \right) (\log \pi_k - F(\lambda_k)) \right\} = 0,$$

$$\sum_{n=1}^{N} q_{nk} \phi(x_n) - \left(\sum_{n=1}^{N} q_{nk} \right) \nabla F(\lambda_k) = 0, \boxed{\sum_{n=1}^{N} \frac{q_{nk}}{\sum_{n=1}^{N} q_{nk}} \phi(x_n) = \nabla F(\lambda_k)}.$$

Moreover,

$$\nabla F(\lambda_k) = \nabla_{\lambda_k} \log \int \exp\left(\langle \phi(x), \lambda_k \rangle \right) \, dx = \langle \phi(x) \rangle_{p(x;\lambda_k)}.$$

Hence, we obtain matching expectations of statistics:

$$\boxed{\sum_{n=1}^{N} \frac{q_{nk}}{\sum_{n=1}^{N} q_{nk}} \phi(x_n) = \langle \phi(x) \rangle_{p(x;\lambda_k)}}.$$

So, we can see that the EM algorithm works by just making soft-clustering and estimation of MLE inside each. For the 1 component mixture we recover the simple MLE estimation:

$$\boxed{\sum_{n=1}^{N} \frac{1}{N} \phi(x_n) = \langle \phi(x) \rangle_{p(x;\lambda)}}.$$