Bayesian Methods in Machine Learning, Seminar: 6

Evgenii Egorov, egorov.evgenyy@ya.ru

Skoltech, 2020

Recap: EM equations

$$\text{Problem: } \max_{\theta} \log p(x;\theta) = \max_{\theta} \log \int p(x,z;\theta)dz,$$

$$\text{Variational Bound:} (H[q] = -\langle \log q \rangle_q).$$

$$\log \int p(x,z;\theta)dz = \max_{q(z)} \langle \log p(x,z;\theta) \rangle_{q(z)} + H[q],$$

$$p(z|x;\theta) = \arg\max_{q(z)} \langle \log p(x,z;\theta) \rangle_{q(z)} + H[q],$$

$$\int \log p(x,z;\theta)dz \geq \langle \log p(x,z;\theta) \rangle_{q(z)} + H[q].$$

Hence, we have EM iterations:

$$\text{E-step: } q(z) = p(z|x;\theta^{\text{old}}),$$

$$\text{M-step: } \max_{\theta} \langle \log p(x,z;\theta) \rangle_{q(z)}.$$

Problem: Mixture Separation

Consider following model:

$$p(x) = \gamma p_0(x) + (1-\gamma)p_1(x),$$
$$p_0(x) = \alpha[x=1] + (1-\alpha)[x=2],$$
$$p_1(x) = \beta[x=2] + (1-\beta)[x=3].$$

From iid sample $X = \{x_n\}_{n=1}^N$ we need to recover $\theta = \{\alpha, \beta, \gamma\}$.

- Note, that problem $\max\limits_{\theta} \log p(X)$ is not concave or log-concave.
- We will use EM-algorithm. It is still converge to the local optimum, but quick and efficiently :)

Model specification

To apply EM, we need to specify who-is-who:

$$\theta = \{\alpha, \beta, \gamma\},$$
$$P(X, Z; \theta) = \prod_{n=1}^{N} p(x_n, z_n; \theta) = \prod_{n=1}^{N} p(x_n|z_n; \theta)p(z_n),$$
$$p(x_n|z_n; \theta) = p_{z_n}(x_n; \theta),$$
$$p(z_n) = \gamma[z_n = 0] + (1 - \gamma)[z_n = 1].$$

Note, that our latent variables z_n are **local**, and θ is **global**.

Problem: E-step

Initialize θ some how reasonable and we need to estimate:

$$
\begin{aligned}
p(z_n = 1 | x_n = 1, \theta^{old}) &= ? \\
p(z_n = 1 | x_n = 2, \theta^{old}) &= ? \\
p(z_n = 1 | x_n = 3, \theta^{old}) &= ?
\end{aligned}
$$

Problem: M-step

And we need to solve:

$$\max_{\theta} \langle \log p(X, Z; \theta) \rangle_{q(Z)} = \max_{\theta} \sum_{n=1}^{N} \langle \log p(x_n, z_n; \theta) \rangle_{p(z_n | x_n; \theta^{\text{old}})}.$$

Encourage you to do it at home. At the class we consider the general case.

Derivation of Discrete Mixture Model For Exponential Family

The goal of the note is 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:

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

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

Problem: E-step

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

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

- It is computationally stable to estimate log of $p(z_n = k|x_n; \theta^{\text{old}}))$ matrix and use for denominator sum-log-exp trick.
- For large n we have a scale problem.

Problem: M-step

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

$$\theta^{\text{new}} = \arg\max_{\theta} \sum_{n=1}^{N} \langle \log p(x_n, z_n; \theta) \rangle_{q(t_n|x_n;\theta^{\text{old}})}.$$

Model for K components:

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

Solution: M-step

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

We can recall a similar result for simple MLE estimation:

$$\log p(x_1, \ldots, x_n; \lambda) = \langle \sum_{n=1}^{N} \phi(x_n), \lambda \rangle - NF(\lambda).$$

Hence, we obtain by first order optimal condition : $\frac{1}{N}\sum_{n=1}^{N} \phi(x_n) = \langle \phi(x) \rangle_{p(x;\lambda)}$.

So, we can see that our EM algorithm works just make soft-clustering and estimation of MLE inside each cluster. We can easily go further and add prior to the θ.

Problem: Multivariate Student Distribution

Recall Gamma distribution:

$$\text{Gamma}x; \alpha, \beta = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x).$$

Consider:

$$p(x|z) = \mathcal{N}(\mu, z^{-1}\Sigma), \quad p(z) = \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}).$$

And their mixture model:

$$p(x|\mu, \Sigma, \nu) = \int_{\mathbb{R}^+} p(x|\mu, z^{-1}\Sigma) p(z) dz.$$

Solution: Multivariate Student Distribution

$$p(x|\mu,\Sigma,\nu) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})}|\nu\pi\Sigma|^{-\frac{1}{2}}\left(1+\frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)^{-\frac{d+\nu}{2}}.$$

The mixture representation much **easier** for taking expectations, linear transformations, conditioning ....
For example, you can just use:

$$\mathbb{E}x = \mathbb{E}_z\mathbb{E}_{x|z}x = \mu.$$