EM & Variational Bayes

Hanxiao Liu

September 9, 2014

Outline

MLE by Gradient Ascent

Goal: maximize $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \log p(\boldsymbol{X}|\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$

Gradient Ascent (GA)

- One-step view: $\boldsymbol{\theta}^{t+1} \leftarrow \nabla\mathcal{L}\left(\boldsymbol{\theta}^{t}; \boldsymbol{X}\right) + \boldsymbol{\theta}^{t}$
- Two-step view:
  1. $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{t}\right) = \mathcal{L}\left(\boldsymbol{\theta}^{t}; \boldsymbol{X}\right) + \left(\boldsymbol{\theta}^{t} - \boldsymbol{\theta}\right)\nabla\mathcal{L}\left(\boldsymbol{\theta}^{t}; \boldsymbol{X}\right) - \frac{1}{2}\left\|\boldsymbol{\theta}^{t} - \boldsymbol{\theta}\right\|_2^2$
  2. $\boldsymbol{\theta}^{t+1} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{t}\right)$

Drawbacks

1. $\nabla\mathcal{L}$ can be too complicated to work with
2. Too general to be efficient for structured problems

MLE by EM

Expectation-maximization (EM)

1. Expectation: $Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^t\right) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t}\,\mathcal{L}\left(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z}\right)$
2. Maximization: $\boldsymbol{\theta}^{t+1} \leftarrow \mathrm{argmax}_{\boldsymbol{\theta}}\, Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^t\right)$

- Replace $\underbrace{\mathcal{L}\left(\boldsymbol{\theta};\boldsymbol{X}\right)}_{\text{log-likelihood}}$ by $\underbrace{\mathcal{L}\left(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z}\right)}_{\text{complete log-likelihood}}$
- $\mathcal{L}\left(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z}\right)$ is a random function w.r.t $\boldsymbol{Z}$
  —use the expected function as a surrogate

why EM is superior

A comparison between $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^t\right)$, i.e., the local concave model

1. EM

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^t\right) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t} \mathcal{L}\left(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}\right)$$
$$= \mathcal{L}\left(\boldsymbol{\theta}; \boldsymbol{X}\right) - D_{KL}\left(p\left(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^t\right) || p\left(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}\right)\right) + C$$

2. GA

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^t\right) = \mathcal{L}\left(\boldsymbol{\theta}^t; \boldsymbol{X}\right) + \left(\boldsymbol{\theta}^t - \boldsymbol{\theta}\right) \nabla \mathcal{L}\left(\boldsymbol{\theta}^t; \boldsymbol{X}\right) - \frac{1}{2}\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}\right\|_2^2$$

Example: vMF mixture

Notations

- $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$, $\boldsymbol{\theta} = \left\{\boldsymbol{\pi} \in \Delta^{k-1}, \{(\boldsymbol{\mu}_i, \kappa_i)\}_{i=1}^k\right\}$
- $\boldsymbol{Z} = \{z_{ij} \in \{0, 1\}\}$
  - $z_{ij} = 1 \implies x_i \sim$ the j-th mixture component

Log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^n \log p(\boldsymbol{x}_i | \boldsymbol{\theta}) = \sum_{i=1}^n \underbrace{\log \sum_{j=1}^k \pi_j \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right)}_{\text{log sum coupling}}$$

Complete log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^n \log p(\boldsymbol{x}_i, \boldsymbol{z}_i | \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log\left(\pi_j \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right)\right)$$

E-step

Compute $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^t\right) \triangleq \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t} \mathcal{L}\left(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}\right)$

$$
\begin{aligned}
& Q\left(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}; \boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\kappa}^t\right) \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\pi}^t,\boldsymbol{\mu}^t,\boldsymbol{\kappa}^t} \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log\left(\pi_j \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right)\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^t \log \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right) + w_{ij}^t \log \pi_j
\end{aligned}
$$

where

$$
\begin{aligned}
w_{ij}^t &= \mathbb{E}_{z_{ij}|\boldsymbol{X},\boldsymbol{\pi}^t,\boldsymbol{\mu}^t,\boldsymbol{\kappa}^t}\left[z_{ij}\right] = p\left(z_{ij} = 1 | \boldsymbol{x}_i, \boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\kappa}^t\right) \\
&= \frac{\pi_j^t \cdot \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j^t, \kappa_j^t\right)}{\sum_{u=1}^{k} \pi_u^t \cdot \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_u^t, \kappa_u^t\right)}
\end{aligned}
$$

M-step

Maximize

$$Q\left(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}; \boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\kappa}^t\right) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^t \log \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right) + w_{ij}^t \log \pi_j$$

w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ s.t. $|\boldsymbol{\pi}|_1 = 1$ and $\|\boldsymbol{\mu}_j\|_2 = 1$, $\forall j \in [k]$

To impose constraints, maximize

$$\tilde{Q} \triangleq Q + \lambda\left(1 - \boldsymbol{\pi}^\top \mathbf{1}\right) + \sum_{j=1}^{k} \nu_j \left(1 - \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_j\right)$$

M-step

$$\tilde{Q}\left(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}; \boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\kappa}^t\right) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^t \log \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right) + w_{ij}^t \log \pi_j$$
$$+ \lambda\left(1 - \boldsymbol{\pi}^\top \mathbf{1}\right) + \sum_{j=1}^{k} \nu_j \left(1 - \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_j\right)$$

Updating π_j^t

Combining $\sum_j^k \pi_j = \sum_j^k w_{ij}^t = 1$ with

$$\partial_{\pi_j} \tilde{Q} = \frac{\sum_{i=1}^n w_{ij}^t}{\pi_j} - \lambda = 0$$

$$\Longrightarrow \qquad \pi_j^{t+1} \leftarrow \frac{\sum_{i=1}^n w_{ij}^t}{n}$$

M-step

$$\tilde{Q}\left(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\kappa}; \boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\kappa}^t\right) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^t \log \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right) + w_{ij}^t \log \pi_j$$
$$+ \lambda\left(1 - \boldsymbol{\pi}^\top \mathbf{1}\right) + \sum_{j=1}^{k} \nu_j \left(1 - \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_j\right)$$

Updating $\boldsymbol{\mu}_j^t$

$\log \mathrm{vMF}\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \kappa_j\right) = \kappa_j \boldsymbol{\mu}_j^\top \boldsymbol{x}_i + C \quad$ (w.r.t $\boldsymbol{\mu}_j$)

$$\partial_{\mu_j} \tilde{Q} = \kappa_j \sum_{i=1}^{n} w_{ij}^t \boldsymbol{x}_i - \nu_j \boldsymbol{\mu}_j = 0$$

$\implies \boldsymbol{\mu}_j^{t+1} \leftarrow \frac{\boldsymbol{r}_j}{\|\boldsymbol{r}_j\|_2}$ where $\boldsymbol{r}_j = \sum_{i=1}^{n} w_{ij}^t \boldsymbol{x}_i$

M-step

Updating κ_j^t

- $C_p\left(\kappa_j\right) = \frac{\kappa_j^{\frac{p}{2}-1}}{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa_j)}$
- the recurrence property of modified Bessel function [1]

$$\partial_{\kappa_j} \log I_{\frac{p}{2}-1}\left(\kappa_j\right) = \frac{\frac{p}{2}-1}{\kappa_j} + \frac{I_{\frac{p}{2}}(\kappa_j)}{I_{\frac{p}{2}-1}(\kappa_j)}$$

$$\partial_{\kappa_j} \tilde{Q} = \sum_{i=1}^{n} w_{ij}^t \left(-\frac{I_{\frac{p}{2}}\left(\kappa_j\right)}{I_{\frac{p}{2}-1}\left(\kappa_j\right)} + \boldsymbol{\mu}_j^\top \boldsymbol{x}_i \right) = 0$$

$$\Longrightarrow \frac{I_{\frac{p}{2}}\left(\kappa_j\right)}{I_{\frac{p}{2}-1}\left(\kappa_j\right)} = \bar{r}_j \Longrightarrow \kappa_j^{t+1} \approx \frac{\bar{r}_j p - \bar{r}_j^3}{1 - \bar{r}_j^2} \quad \textbf{[?]}$$

where $\bar{r}_j = \frac{\sum_{i=1}^{n} w_{ij}^t \boldsymbol{\mu}_j^\top \boldsymbol{x}_i}{\sum_{i=1}^{n} w_{ij}^t}$

[1] http://functions.wolfram.com/Bessel-TypeFunctions/BesselK/introductions/Bessels/05/

An alternative view of EM

EM - original definition

1. Expectation: $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^t\right) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t} \mathcal{L}\left(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}\right)$ why?
2. Maximization: $\boldsymbol{\theta}^{t+1} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^t\right)$

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}; \boldsymbol{X}\right) &= \mathbb{E}_q \ \log p\left(\boldsymbol{X}|\boldsymbol{\theta}\right) \\
&= \underbrace{\mathbb{E}_q \left[\log \frac{p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}\right)}{q\left(\boldsymbol{Z}\right)}\right]}_{\mathrm{VLB}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_q \left[\log \frac{q\left(\boldsymbol{Z}\right)}{p\left(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}\right)}\right]}_{D_{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}))}
\end{aligned}$$

EM - coordinate ascent

1. $q^{t+1} = \operatorname{argmax}_q \ \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$
2. $\boldsymbol{\theta}^{t+1} = \operatorname{argmax}_{\boldsymbol{\theta}} \ \mathrm{VLB}\left(q^{t+1}, \boldsymbol{\theta}\right)$

Show the equivalence?

Bayes Inference

Notations

- $\boldsymbol{\theta}$: hyper parameters
- $\boldsymbol{Z}$: hidden variables + random parameters

Goals

1. find a good posterior $q(\boldsymbol{Z}) \approx p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta})$
2. estimate $\boldsymbol{\theta}$ by Empirical Bayes, i.e., maximize $\mathcal{L}(\boldsymbol{\theta};\boldsymbol{X})$ w.r.t $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{X}) = \underbrace{\mathbb{E}_q\left[\log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]}_{\mathrm{VLB}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\right]}_{D_{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}))}$$

both goals can be achieved via the same procedure as EM

Variational Bayes Inference

One should have $q \to p\,(\boldsymbol{Z}; \boldsymbol{X}, \boldsymbol{\theta}^*)$ by alternating between

1. $q^{t+1} = \operatorname{argmax}_q \, \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$
2. $\boldsymbol{\theta}^{t+1} = \operatorname{argmax}_{\boldsymbol{\theta}} \, \mathrm{VLB}\left(q^{t+1}, \boldsymbol{\theta}\right)$

However, we do not want q to be too complicated

- e.g., $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^t\right) = \mathbb{E}_q \, \mathcal{L}\left(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}\right)$ can be intractable

Solution: modify the first step as

$$q^{t+1} = \operatorname{argmax}_{q \in \mathcal{Q}} \, \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$$

$\mathcal{Q}$ - some tractable distribution families

- Recall: without $\mathcal{Q}$, $q^{t+1} \equiv p\left(\boldsymbol{Z} | \boldsymbol{X}, \boldsymbol{\theta}^t\right)$

Variational Bayes Inference

Goal: solve $\operatorname{argmax}_{q\in\mathcal{Q}} \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$

usually, $\mathcal{Q} = \left\{q \mid q\left(\boldsymbol{Z}\right) = \prod_{i=1}^M q_i\left(\boldsymbol{Z}_i\right) \triangleq \prod_{i=1}^M q_i\right\}$

Coordinate ascent

$$\begin{aligned}
\mathrm{VLB}\left(q_j; q_{-j}, \boldsymbol{\theta}^t\right) &= \mathbb{E}_q\left[\log \frac{p\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}^t\right)}{q\left(\boldsymbol{Z}\right)}\right] \\
&= \mathbb{E}_q \log p\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}^t\right) - \sum_{i=1}^M \mathbb{E}_q \log q_i \\
&= \mathbb{E}_{q_j}\left(\mathbb{E}_{q_{-j}} \log p\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}^t\right)\right) - \mathbb{E}_{q_j} \log q_j + C \\
&= -D_{KL}\left(\log q_j || \mathbb{E}_{q_{-j}} \log p\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}^t\right)\right) + C
\end{aligned}$$

$$\log q_j^* = \mathbb{E}_{q_{-j}} \log p\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}^t\right)$$

Example: Bayes Mixture of Gaussians

Consider putting a prior over the means in GM [2]

- For $k = 1, 2 \dots K$, $\mu_k \sim \mathcal{N}\left(0, \tau^2\right)$
- For $i = 1, 2 \dots N$
  1. $z_i \sim \mathrm{Mult}\left(\boldsymbol{\pi}\right)$
  2. $x_i \sim \mathcal{N}\left(\mu_{z_i}, \sigma^2\right)$

$$
\begin{aligned}
p\left(\boldsymbol{z}, \boldsymbol{\mu} | \boldsymbol{X}\right) &= \frac{p\left(\boldsymbol{X} | \boldsymbol{z}, \boldsymbol{\mu}\right) p\left(\boldsymbol{z}\right) p\left(\boldsymbol{\mu}\right)}{p\left(\boldsymbol{X}\right)} \\
&= \frac{\prod_{i=1}^{N} p\left(z_i\right) p\left(x_i | z_i, \boldsymbol{\mu}\right) \prod_{k=1}^{K} p\left(\mu_k\right)}{\int \sum_{\boldsymbol{z}} \prod_{i=1}^{N} p\left(z_i\right) p\left(x_i | z_i, \boldsymbol{\mu}\right) \prod_{k=1}^{K} p\left(\mu_k\right) d\boldsymbol{\mu}}
\end{aligned}
$$

$$
q\left(\boldsymbol{z}, \boldsymbol{\mu}\right) = \prod_{i=1}^{N} q\left(z_i; \boldsymbol{\phi}_i\right) \prod_{k=1}^{K} q\left(\mu_k; \tilde{\mu}_k, \tilde{\sigma}_k^2\right)
$$

[2] https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf

Example: Bayes Mixture of Gaussians

$$
\begin{aligned}
\log q^*(z_j) &= \mathbb{E}_{q^{\setminus z_j}} \log p(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{X}) \\
&= \mathbb{E}_{q^{\setminus z_j}} \left(\sum_{i=1}^{N} \log p(z_i) + \log p(x_i | z_i, \boldsymbol{\mu}) + \sum_{k=1}^{K} \log p(\mu_k) \right) \\
&= \log p(z_j) + \mathbb{E}_{q(\mu_{z_j})} \log p\left(x_j | z_j, \mu_{z_j}\right) + C \\
&= \log \pi_{z_j} + x_j \underbrace{\mathbb{E}_{q(\mu_{z_j})}\left[\mu_{z_j}\right]}_{\tilde{\mu}_{z_j}} - \frac{1}{2} \underbrace{\mathbb{E}_{q(\mu_{z_j})}\left[\mu_{z_j}^2\right]}_{\tilde{\mu}_{z_j}^2 + \tilde{\sigma}_{z_j}^2} + C
\end{aligned}
$$

By observation $q^*(z_j) \sim \mathrm{Mult}$, we can update ϕ_j accordingly

Example: Bayes Mixture of Gaussians

$$\begin{aligned}
\log q^*(\mu_j) &= \mathbb{E}_{q\setminus\mu_j} \log p(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{X}) \\
&= \mathbb{E}_{q\setminus\mu_j} \left(\sum_{i=1}^{N} \log p(z_i) + \log p(x_i|z_i, \mu_{z_i}) + \sum_{k=1}^{K} \log p(\mu_k) \right) \\
&= \mathbb{E}_{q\setminus\mu_j} \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{z_i=k} \log \mathcal{N}(x_i|\mu_k) + \log p(\mu_j) + C \\
&= \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z_i}\left[\delta_{z_i=j}\right]}_{\phi_i^j} \log \mathcal{N}(x_i|\mu_j) + \log p(\mu_j) + C
\end{aligned}$$

Observing that $q^*(\mu_j) \sim \mathcal{N}$, $\tilde{\mu}_j$ and $\tilde{\sigma}_j^2$ can be updated accordingly

Stay tuned

Next topics

- LDA (Wanli)
- Bayes vMF