EM ALGORITHM

BY PUN WAI TONG

How people do an optimization? Let a function $f : V \to \mathbb{R}$ be an objective function. One way to maximize f is to expand the domain of V to $V \times H$ and transform $f : V \to \mathbb{R}$ to $F : V \times H \to \mathbb{R}$ such that $F(v, h) \leq f(v)$ for all $v \in V$ and $h \in H$. Moreover, given $h \in H$, the upper bound of $F(\cdot, h)$ is assumed to be known and is denoted as $g(v) := \max_{h \in H} F(v, h)$. Therefore, the lower bound of f can be estimated as

$$\max_{v \in V} g(v) \leq \max_{v \in V} f(v).$$

Suppose there exists $(v_{optimal}) = argmax\,(g(\cdot))$ exists. Heuristically speaking, $f(v_{optimal})$ should not be too far from $\max_{v \in V} f(v)$ if an appropriate F is chosen. This idea or its variant has been seen very often in optimization area, e.g. EM algorithm and Lagrangian dual problem and is very useful in other mathematics fields, e.g. the proof of concentration inequality in a random graph.

What is a EM algorithm? EM algorithm applies the above idea to maximize the likelihood function. Let T be a training set and S be a hidden space. Let P be a probability density on $T \times S$ and the probability density function P is tuned by a set of parameters which is denoted as θ. Note θ can be viewed as lying in a set V. And by total law of probability, we define $p(\cdot; \theta) : T \to [0, 1]$ as

$$p(t; \theta) := \sum_{s \in S} p(\cdot, s; \theta). \tag{-1.1}$$

The function $p(\cdot; \theta)$ is a probability density function on T. The log-likelihood function ℓ is formulated as

$$\ell : V \to \mathbb{R}$$
$$\ell(\theta) = \sum_{t \in T} log\left(p\left(t; \theta\right)\right).$$

The log-likelihood function ℓ is our objective function for maximization and can be viewed as f the above idea. Since a log function is a concave function, i.e.

$$\sum_{i=1}^{n} w_i log t_i \leq log\left(\sum_{i=1}^{n} w_i t_i\right) \tag{-1.2}$$

for any $t_i > 0$ and $w_i \geq 0$ and $\sum_{i=1}^{n} w_i$=1. By using Eq. (-1.2) [also known as the concave version of Jensen's Inequality] and Eq.(-1.1), the likelihood function becomes

---

Date: 09/20/17.

$$\begin{aligned}\ell(\theta) &= \sum_{t\in T} log\left(\sum_{s\in S} p(t,s;\theta)\right)\\ &= \sum_{t\in T} log\left(\sum_{s\in S} q(s)\frac{p(t,s;\theta)}{q(s)}\right)\\ &\geq \sum_{t\in T}\sum_{s\in S} q(s) log\left(\frac{p(t,s;\theta)}{q(s)}\right) \qquad \text{(-1.3)}\end{aligned}$$

where q is any arbitrary probability density function on S and q can be viewed as lying in H. Let $L(\cdot,\cdot)$ be a function mapping from $V \times H$ to $\mathbb{R}$ as

$$L(\theta, q) := \sum_{t\in T}\sum_{s\in S} q(s) log\left(\frac{p(t,s;\theta)}{q(s)}\right)$$

and the function L can be viewed as F in the above idea. Hence,

$$\max_{\theta} \ell(\theta) \geq \max_{\theta}\max_{q} L(\theta, q).$$

The goal of the EM algorithm is to find $\theta_{optimal}$ to maximize $max_q L(\theta, q)$ and the algorithm is as follows:

(1) Initialize $\theta^{(0)}$ and set $t = 0$
(2) until nothing change very much,
  (a) E-Step: $q^{(t)} = argmax_q L(\theta^{(t)}, \cdot)$
  (b) M-Step: $\theta^{(t+1)} = argmax_\theta L(\cdot, q^{(t)})$
(3) Return a final estimator of θ.

In the E-Step, by using Eq.(-1.3), we have

$$\ell(\theta^{(t)}) \geq \sum_{t\in T}\sum_{s\in S} q(s) log\left(\frac{p(t,s;\theta^{(t)})}{q(s)}\right) = L(\theta^{(t)}, q). \qquad \text{(-1.4)}$$

The inequality in Eq. (-1.4) comes from the concave version of Jensen's Inequality and can be attained if and only if for each t, there is a constant c_t such that $\frac{p(t,s;\theta^{(t)})}{q^{(t)}(s)} = c_t$ for all $s \in S$. If we sum both side over $s \in S$ in the following equation, we learn

$$\begin{aligned} p(t,s;\theta^{(t)}) &= c_t q^{(t)}(s)\\ p(t;\theta^{(t)}) = \sum_{s\in S} p(t,s;\theta^{(t)}) &= c_t \sum_{s\in S} q^{(t)}(s) = c_t\end{aligned}$$

and hence,

$$q^{(t)}(s) = \frac{p(t,s;\theta^{(t)})}{p(t;\theta^{(t)})} = \frac{p(t,s;\theta^{(t)})}{\sum_{s\in S} p(t,s;\theta^{(t)})} = p(s|t;\theta^{(t)}) \qquad \text{(-1.5)}$$

which is known as posterior distribution of s given t and the parameters in $\theta^{(t)}$ under p.

In the M-Step, by using Eq. (-1.5),

$$\theta^{(t+1)} = argmax_\theta L(\cdot, q^{(t)}) = argmax_\theta \sum_{t\in T}\sum_{s\in S} p(t|s;\theta^{(t)}) log\left(\frac{p(t,s;\theta)}{p(t|s;\theta^{(t)})}\right).$$

Then gradient descent method, Newtonian method or solving critial points explicitly is applied to find $\theta^{(t+1)}$ in M-Step.

Why EM algorithm can find at least a local maximum point? It is because of the monotonic property of $L(\theta^{(t)}, q^{(t)})$ in the EM algorithm. Indeed, it can be verified that $L(\theta^{(t)}, q^{(t)}) \leq L(\theta^{(t)}, q^{(t+1)}) \leq L(\theta^{(t+1)}, q^{(t+1)})$ where the first inequality comes from E-Step and the second inequality comes from M-Step.

Same as most optimization scheme, EM algorithm usually falls at local optimal point. In order to get a better optimal point, one suggestion is to repeat EM algorithm multiple times by using different initialization of θ.

0. Bone Yard

Will answer several things in future? (1) Why cannot use stochastic gradient descent directly? (2) Study how to use EM alg on a Guassian mixture model?

E-mail address: punwai.tong@gmail.com