Machine Learning for Data Science (CS 4786)

Lecture 16: EM Algorithm and Mixture Models

1 EM Algorithm Recap

E-step:

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

M-step:

$$\theta^{(i)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(c_t) \log(P(x_t, c_t|\theta))$$

1.1 EM for Mixture Models

For any mixture model with π as mixture distribution, and any arbitrary parameterization of likelihood of data given cluster assignment, one can write down a more detailed form for EM algorithm.

E-step On iteration i, for each data point $t \in [n]$, set

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

Note that

$$\begin{aligned}
Q_t^{(i)}(c_t) &= P(c_t|x_t, \theta^{(i-1)}) \\
&\propto p(x_t|c_t, \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\
&\propto p(x_t|c_t, \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\
&= \frac{p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^{K} p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}
\end{aligned}$$

So all we need to fill out the $n \times K$ sized Q matrix is to have a current guess at π and the ability to compute $p(x_t|c_t, \theta^{(i-1)})$ up to multiplicative factor.

$$
\begin{aligned}
\theta &= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log P(x_t, c_t = k | \theta) \\
&= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log P(x_t | c_t = k, \theta) \times P(c_t = k | \theta) \\
&= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta) \times \pi[k]\right) \\
&= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta)\right) + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(\pi[k]\right)
\end{aligned}
$$

Using $\Theta^{\setminus \pi}$ to denote the set of parameters excluding π,

$$
\begin{aligned}
&= \operatorname*{argmax}_{\theta \in \Theta^{\setminus \pi}, \pi} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta)\right) + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(\pi_k\right) \right) \\
&= \left(\operatorname*{argmax}_{\theta \in \Theta^{\setminus \pi}} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta)\right) \right), \operatorname*{argmax}_{\pi} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(\pi_k\right) \right) \right)
\end{aligned}
$$

Notice that the term in red is exactly the optimization we solved for in GMM example. We know this already! The solution is:

$$
\pi_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}
$$

and this is the same for any mixture model.

On the other hand, the optimization problem,

$$
\operatorname*{argmax}_{\theta \in \Theta^{\setminus \pi}} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta)\right) \right)
$$

is simply a weighted version of MLE when our observation includes c_t's the hidden or latent variables. In the M-step, this is the only portion that changes the mixture distribution solution has same form always.

2 Mixture of Multinomials

Each $\theta \in \Theta$ consist of mixture distribution π which is a distribution over the choices of the K clusters or types, $p_1, \ldots, p_K$ are K distributions over the d items. The latent variables are $c_1, \ldots, c_n$ the cluster assignments for the n points indicating that the t^{th} data point was drawn using distribution p_{c_t}. $x_1, \ldots, x_n$ are the n observations.

Story: You own a grocery store and multiple customers walk in to your store and buy stuff. You want group customers into K group based on distribution over the d products/choices in your store. Think of customers as being independently drawn and they each belong to one of K groups. We will first start with a simple scenario and build up to a more general one. To start with, say each day a customer walks in to your store and buys $m = 1$ product. The generative story then is that we first draw customer type $c_t \sim \pi$ from a mixture distribution π, next associated with type c_t, there is a distribution p_{c_t} over products the customer would buy. We draw $x_t \in [d]$ the product the customer bought as $x_t \sim p_{c_t}$. That is

$$p(x_t|c_t = k, \theta) = p_{c_t}[x_t]$$

Next we can move to a slightly more complex scenario where the customer on every round buys (fixed) $m > 1$ products by drawing x_t as m samples from the multinomial distribution. That is,

$$p(x_t|c_t = k, \theta) = \frac{m!}{x_t[1]! \cdot \ldots \cdot x_t[d]!} p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]}$$

where $x_t[j]$ indicates the amount of product j bought by the customer t.

2.1 Mixture of Multinomials (Primer $m = 1$)

E-step On iteration i, for each data point $t \in [n]$, set

$$\begin{aligned} Q_t^{(i)}(c_t) &= \frac{p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^K p(x_t|c_t, \theta^{(i-1)}) \cdot P(c_t|\theta^{(i-1)})} \\ &= \frac{p_{c_t}^{(i-1)}[x_t] \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^K p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]} \end{aligned}$$

M-step As we already saw, we set

$$\pi_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

Now as for the remaining parameters, we want to maximize

$$\underset{p_1,\ldots,p_K}{\operatorname{argmax}} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log\left(p_k[x_t]\right) \right)$$

Define $L(p_1, \ldots, p_K) = \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log\left(p_k[x_t]\right)$. We want to optimize $L(p_1, \ldots, p_K)$ w.r.t. $p_1, \ldots, p_k$ s.t. each p_k is a valid probability distribution over $\{1, \ldots, d\}$. As an example, to find the optimal p_k, we want to optimize over p_k subject to the constraint $\sum_{j=1}^d p_k[j] = 1$ (ie. its a distribution), we do so by introducing Lagrange variables. That is we find $p_k[j]$'s by taking derivative and equating to 0 the following Lagrangian objective,

$$L(p_1, \ldots, p_K) + \lambda_k(1 - \sum_{j=1}^d p_k[j])$$

Taking derivative and equating to 0, we want to find p_k s.t.,

$$\sum_{t=1}^{n} Q_t^{(i)}(k)\frac{1}{p_k[x_t]} - \lambda_k = 0$$

In other words, for every $j \in [d]$,

$$\sum_{t:x_t=j} Q_t^{(i)}(k)\frac{1}{p_k[j]} - \lambda_k = 0$$

Hence we conclude that

$$p_k[j] \propto \sum_{t:x_t=j} Q_t^{(i)}(k)$$

Hence,

$$p_k[j] = \frac{\sum_{t:x_t=j} Q_t^{(i)}(k)}{\sum_{t=1}^{n} Q_t^{(i)}(k)}$$

Thus for the M-step when we are dealing with the mixture model with exactly $m = 1$ purchase on every round, we get that, for every $k \in [K]$ and every $j \in [d]$,

$$p_k[j] = \frac{\sum_{t:x_t=j} Q_t^{(i)}(k)}{\sum_{t=1}^{n} Q_t^{(i)}(k)}$$

2.2 Mixture of Multinomials ($m > 1$)

E-step On iteration i, for each data point $t \in [n]$, set

$$\begin{aligned} Q_t^{(i)}(c_t) &= \frac{p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}{\sum_{k=1}^{K} p(x_t|k, \theta^{(i-1)}) \cdot P(k|\theta^{(i-1)})} \\ &= \frac{p_{c_t}[1]^{x_t[1]} \cdot \ldots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^{K} p_{c_t}[1]^{x_t[1]} \cdot \ldots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi^{(i-1)}[k]} \end{aligned}$$

M-step For mixture distribution, as usual,

$$\pi_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}$$

Now as for the remaining parameters, we want to maximize

$$\begin{aligned} &\underset{p_1,\ldots,p_K}{\operatorname{argmax}} \left(\sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\log\left(P(x_t|c_t = k, \theta)\right)\right) \\ &= \underset{p_1,\ldots,p_K}{\operatorname{argmax}} \left(\sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\log\left(p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]}\right)\right) \\ &= \underset{p_1,\ldots,p_K}{\operatorname{argmax}} \left(\sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\sum_{j=1}^{d} x_t[j]\log\left(p_k[j]\right)\right) \end{aligned}$$

Again to solve this, define $L(p_1, \ldots, p_K) = \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \sum_{j=1}^{d} x_t[j] \log(p_k[j])$. We want to optimize $L(p_1, \ldots, p_K)$ w.r.t. $p_1, \ldots, p_k$ s.t. each p_k is a valid probability distribution over $\{1, \ldots, d\}$. As an example, to find the optimal p_k, we want to optimize over p_k subject to the constraint $\sum_{j=1}^{d} p_k[j] = 1$ (ie. its a distribution), we do so by introducing Lagrange variables. That is we find $p_k[j]$'s by taking derivative and equating to 0 the following Lagrangian objective,

$$L(p_1, \ldots, p_K) + \lambda_k(1 - \sum_{j=1}^{d} p_k[j])$$

Taking derivative and equating to 0, we want to find p_k s.t.,

$$\sum_{t=1}^{n} Q_t^{(i)}(k) \sum_{j=1}^{d} x_t[j] \frac{1}{p_k[j]} - \lambda_k = 0$$

In other words, for every $j \in [d]$,

$$\sum_{t=1}^{n} Q_t^{(i)}(k) \frac{x_t[j]}{p_k[j]} - \lambda_k = 0$$

Hence we conclude that

$$p_k[j] \propto \sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]$$

Hence,

$$p_k[j] = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]}{\sum_{j=1}^{d} \sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]}{\sum_{t=1}^{n} Q_t^{(i)}(k) \left(\sum_{j=1}^{d} x_t[j]\right)} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]}{m \sum_{t=1}^{n} Q_t^{(i)}(k)}$$