9.2 Parameter estimation for mixtures of distributions

We now investigate parameter estimation for mixture models, which is most commonly carried out using the *expectation-maximization (EM) algorithm*. As before, we are given a set of i.i.d. observations $\mathcal{D} = \{x_i\}_{i=1}^n$, with the goal of estimating the parameters of the mixture distribution

$$p(x|\theta) = \sum_{j=1}^{m} w_j p(x|\theta_j).$$

In the equation above, we used $\theta = (w_1, w_2, \ldots, w_m, \theta_1, \theta_2, \ldots, \theta_m)$ to combine all parameters. Just to be more concrete, we shall assume that each $p(x_i|\theta_j)$ is an exponential distribution with parameter λ_j, i.e. $p(x|\theta_j) = \lambda_j e^{-\lambda_j x}$, where $\lambda_j > 0$. Finally, we shall assume that m is given and will address simultaneous estimation of θ and m later.

Let us attempt to find the maximum likelihood solution first. By plugging the formula for $p(x|\theta)$ into the likelihood function we obtain

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} w_j p(x_i|\theta_j) \right), \end{aligned} \tag{9.1}$$

which, unfortunately, is difficult to maximize using differential calculus (why?). Note that although $p(\mathcal{D}|\theta)$ has $O(m^n)$ terms, it can be calculated in $O(mn)$ time as a log-likelihood.

Before introducing the EM algorithm, let us for a moment present two hypothetical scenarios that will help us to understand the algorithm. First, suppose that information is available as to which mixing component generated which data point. That is, suppose that $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is an i.i.d. sample from some distribution $p_{XY}(x, y)$, where $y \in \mathcal{Y} = \{1, 2, \ldots, m\}$ specifies the mixing component. How would the maximization be performed then? Let us write the likelihood function as

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i, y_i|\theta) \\
&= \prod_{i=1}^{n} p(x_i|y_i, \theta) p(y_i|\theta) \\
&= \prod_{i=1}^{n} w_{y_i} p(x_i|\theta_{y_i}), \qquad (9.2)
\end{aligned}
$$

where $w_j = p_Y(j) = P(Y = j)$. The log-likelihood is

$$
\begin{aligned}
\log p(\mathcal{D}|\theta) &= \sum_{i=1}^{n} (\log w_{y_i} + \log p(x_i|\theta_{y_i})) \\
&= \sum_{j=1}^{m} n_j \log w_j + \sum_{i=1}^{n} \log p(x_i|\theta_{y_i}),
\end{aligned}
$$

where n_j is the number of data points in $\mathcal{D}$ generated by the j-th mixing component.

It is useful to observe here that when $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is known, the internal summation operator in Eq. (9.1) disappears. More importantly, it follows that Eq. (9.2) can be maximized in a relatively straightforward manner. Let us show how. To find $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ we need to solve a constrained optimization problem, which we will do by using the method of Lagrange multipliers. We shall first form the Lagrangian function as

$$
L(\mathbf{w}, \alpha) = \sum_{j=1}^{m} n_j \log w_j + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right)
$$

where α is the Lagrange multiplier. Then, by setting $\frac{\partial}{\partial w_k} L(\mathbf{w}, \alpha) = 0$ for every $k \in \mathcal{Y}$ and $\frac{\partial}{\partial \alpha} L(\mathbf{w}, \alpha) = 0$, we derive that $w_k = -\frac{n_k}{\alpha}$ and $\alpha = -n$. Thus,

$$
w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k),
$$

where $I(\cdot)$ is the indicator function. To find all θ_j, we recall that we assumed a mixture of exponential distributions. Thus, we proceed by setting

$$
\frac{\partial}{\partial \lambda_k} \sum_{i=1}^{n} \log p(x_i|\lambda_{y_i}) = 0,
$$

for each $k \in \mathcal{Y}$. We obtain that

$$\lambda_k = \frac{n_k}{\sum_{i=1}^{n} I(y_i = k) \cdot x_i},$$

which is simply the inverse mean over those data points generated by the k-th mixture component. In summary, we observe that if the mixing component designations $\mathbf{y}$ are known, the parameter estimation is greatly simplified. This was achieved by decoupling the estimation of mixing proportions and all parameters of the mixing distributions.

In the second hypothetical scenario, suppose that parameters θ are known, and that we would like to estimate the best configuration of the mixture designations $\mathbf{y}$ (one may be tempted to call them "class labels"). This task looks like clustering in which cluster memberships need to be determined based on the known set of mixing distributions and mixing probabilities. To do this we can calculate the posterior distribution of $\mathbf{y}$ as

$$\begin{aligned} p(\mathbf{y}|\mathcal{D}, \theta) &= \prod_{i=1}^{n} p(y_i|x_i, \theta) \\ &= \prod_{i=1}^{n} \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \end{aligned} \tag{9.3}$$

and subsequently find the best configuration out of m^n possibilities. Obviously, because of the i.i.d. assumption each element y_i can be estimated separately and, thus, this estimation can be completed in $O(mn)$ time. The MAP estimate for y_i can be found as

$$\hat{y}_i = \arg\max_{y_i \in \mathcal{Y}} \left\{ \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \right\}$$

for each $i \in \{1, 2, \ldots, n\}$.

In reality, neither "class labels" $\mathbf{y}$ nor the parameters θ are known. Fortunately, we have just seen that the optimization step is relatively straightforward if one of them is known. Therefore, the intuition behind the EM algorithm is to form an iterative procedure by *assuming* that either $\mathbf{y}$ or θ is known and calculate the other. For example, we can initially pick some value for θ, say $\theta^{(0)}$, and then estimate $\mathbf{y}$ by computing $p(\mathbf{y}|\mathcal{D}, \theta^{(0)})$ as in Eq. (9.3). We can refer to this estimate as $\mathbf{y}^{(0)}$. Using $\mathbf{y}^{(0)}$ we can now refine the estimate of θ to $\theta^{(1)}$ using Eq. (9.2). We can then iterate these two steps until convergence. In the case of mixture of exponential distributions, we arrive at the following algorithm:

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Calculate $y_i^{(0)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(0)} p(x_i|\lambda_k^{(0)})}{\sum_{j=1}^m w_j^{(0)} p(x_i|\lambda_j^{(0)})} \right\}$ for $\forall i \in \{1, 2, \ldots, n\}$
3. Set $t = 0$
4. Repeat until convergence
   (a) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n I(y_i^{(t)} = j)$
   (b) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n I(y_i^{(t)}=k)}{\sum_{i=1}^n I(y_i^{(t)}=k) \cdot x_i}$
   (c) $t = t + 1$
   (d) $y_i^{(t+1)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^m w_j^{(t)} p(x_i|\lambda_j^{(t)})} \right\}$
5. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

This procedure is not quite yet the EM algorithm; rather, it is a version of it referred to as *classification EM algorithm*. In the next section we will introduce the EM algorithm.

The expectation-maximization algorithm

The previous procedure was designed to iteratively estimate class memberships and parameters of the distribution. In reality, it is not necessary to compute $\mathbf{y}$; after all, we only need to estimate θ. To accomplish this, at each step t, we can use $p(\mathbf{y}|\mathcal{D}, \theta^{(t)})$ to maximize the *expected log-likelihood* of both $\mathcal{D}$ and $\mathbf{y}$

$$E_{\mathbf{Y}}[\log p(\mathcal{D}, \mathbf{y}|\theta)|\theta^{(t)}] = \sum_{\mathbf{y}} \log p(\mathcal{D}, \mathbf{y}|\theta) p(\mathbf{y}|\mathcal{D}, \theta^{(t)}), \tag{9.4}$$

which can be carried out by integrating the log-likelihood function of $\mathcal{D}$ and $\mathbf{y}$ over the posterior distribution for $\mathbf{y}$ in which the current values of the parameters $\theta^{(t)}$ are assumbed to be known. We can now formulate the expression for the parameters in step $t + 1$ as

$$\theta^{(t+1)} = \arg\max_{\theta} \left\{ E[\log p(\mathcal{D}, \mathbf{y}|\theta)|\theta^{(t)}] \right\}. \tag{9.5}$$

The formula above is all that is necessary to create the update rule for the EM algorithm. Note, however, that inside of it we always have to re-compute

$E_{\mathbf{Y}}[\log p(\mathcal{D}, \mathbf{y}|\theta)|\theta^{(t)}]$ function because the parameters $\theta^{(t)}$ have been updated from the previous step. We then can perform maximization. Hence the name "expectation-maximization", although it is perfectly valid to think of the EM algorithm as an interative maximization of expectation from Eq. (9.4), i.e. "expectation maximization".

We now proceed as follows

$$\begin{aligned} E[\log p(\mathcal{D}, \mathbf{y}|\theta)|\theta^{(t)}] &= \sum_{y_1=1}^{m} \cdots \sum_{y_n=1}^{m} \log p(\mathcal{D}, \mathbf{y}|\theta) p(\mathbf{y}|\mathcal{D}, \theta^{(t)}) \\ &= \sum_{y_1=1}^{m} \cdots \sum_{y_n=1}^{m} \sum_{i=1}^{n} \log p(x_i, y_i|\theta) \prod_{l=1}^{n} p(y_l|x_l, \theta^{(t)}) \\ &= \sum_{y_1=1}^{m} \cdots \sum_{y_n=1}^{m} \sum_{i=1}^{n} \log \left(w_{y_i} p(x_i|\theta_{y_i})\right) \prod_{l=1}^{n} p(y_l|x_l, \theta^{(t)}). \end{aligned}$$

After several simplification steps, that we omit for space reasons, the expectation of the likelihood can be written as

$$E[\log p(\mathcal{D}, \mathbf{y}|\theta)|\theta^{(t)}] = \sum_{i=1}^{n} \sum_{j=1}^{m} \log \left(w_j p(x_i|\theta_j)\right) p_{Y_i}(j|x_i, \theta^{(t)}),$$

from which we can see that $\mathbf{w}$ and $\{\theta_j\}_{j=1}^{m}$ can be separately found. In the final two steps, we will first derive the update rule for the mixing probabilities and then by assuming the mixing distributions are exponential, derive the update rules for their parameters.

To maximize $E[\log p(\mathcal{D}, \mathbf{y}|\theta)|\theta^{(t)}]$ with respect to $\mathbf{w}$, we observe that this is an instance of constrained optimization because it must hold that $\sum_{i=1}^{m} w_i = 1$. We will use the method of Lagrange multipliers; thus, for each $k \in \mathcal{Y}$ we need to solve

$$\frac{\partial}{\partial w_k} \left(\sum_{j=1}^{m} \log w_j \sum_{i=1}^{n} p_{Y_i}(j|x_i, \theta^{(t)}) + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right) \right) = 0,$$

where α is the Lagrange multiplier. It is relatively straightforward to show that

$$w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{Y_i}(k|x_i, \theta^{(t)}). \tag{9.6}$$

Similarly, to find the solution for the parameters of the mixture distributions, we obtain that

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n p_{Y_i}(k|x_i, \theta^{(t)})}{\sum_{i=1}^n x_i p_{Y_i}(k|x_i, \theta^{(t)})} \tag{9.7}$$

for $k \in \mathcal{Y}$. As previously shown, we have

$$p_{Y_i}(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^m w_j^{(t)} p(x_i|\lambda_j^{(t)})}, \tag{9.8}$$

which can be computed and stored as an $n \times m$ matrix. In summary, for the mixture of m exponential distributions, we summarize the EM algorithm by combining Eqs. (9.6-9.8) as follows:

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Set $t = 0$
3. Repeat until convergence
   (a) $p_{Y_i}(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^m w_j^{(t)} p(x_i|\lambda_j^{(t)})}$ for $\forall (i, k)$
   (b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n p_{Y_i}(k|x_i, \theta^{(t)})$
   (c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^n p_{Y_i}(k|x_i, \theta^{(t)})}{\sum_{i=1}^n x_i p_{Y_i}(k|x_i, \theta^{(t)})}$
   (d) $t = t + 1$
4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$

Similar update rules can be obtained for different probability distributions; however, a separate derivatives have to be found.

Notice the difference between the CEM and the EM algorithms.

Identifiability

When estimating the parameters of a mixture, it is possible that for some parametric families one obtains multiple solutions. In other words,

$$p(x; \theta) = \sum_{j=1}^m w_j p(x; \theta_j),$$

but can also be expressed as

$$p(x;\theta') = \sum_{j=1}^{m'} w'_j p(x;\theta'_j),$$

The parameters are identifiable if

$$\sum_{j=1}^{m} w_j p(x;\theta_j) = \sum_{j=1}^{m'} w'_j p(x;\theta'_j),$$

implies that $m = m'$ for each $j \in \{1, 2, \ldots, m\}$ there exists some $l \in \{1, 2, \ldots, m\}$ such that $w_j = w'_l$ and $\theta_j = \theta'_l$.