Expectation-Maximisation

Wilker Aziz

Consider a collection of (discrete) observations $\mathcal{D} = \{x^{(1)}, \ldots, x^{(|\mathcal{D}|)}\}$. Suppose a model of the observations is expressed in terms of a (discrete) hidden variable z where all dependencies between observed and hidden variables are expressed by parameters θ.[1]

$$P(\mathcal{D}|\theta) = \prod_{s=1}^{|\mathcal{D}|} \underbrace{P(x^{(s)}|\theta)}_{\text{incomplete-data likelihood}} = \prod_{s=1}^{|\mathcal{D}|} \sum_{z^{(s)}} \underbrace{P(x^{(s)}, z^{(s)}|\theta)}_{\text{complete-data likelihood}} \tag{1}$$

In maximum likelihood learning, we seek to find parameters θ_{ML} that maximise the incomplete-data likelihood, or equivalently, its logarithm (since log is a monotone function),

$$\mathcal{L}(\theta|\mathcal{D}) \equiv \log P(\mathcal{D}|\theta) = \sum_{s=1}^{|\mathcal{D}|} \log P(x^{(s)}|\theta) = \sum_{s=1}^{|\mathcal{D}|} \log \left(\sum_{z^{(s)}} P(x^{(s)}, z^{(s)}|\theta) \right) \tag{2}$$

thus

$$\theta_{\text{ML}} \equiv \arg\max_{\theta} \mathcal{L}(\theta|\mathcal{D}) \ . \tag{3}$$

To avoid clutter, we sometimes derive steps using the contribution $\mathcal{L}(\theta|x)$ to Equation (2) of a single sample x, that is

$$\mathcal{L}(\theta|\mathcal{D}) \equiv \sum_{x \in \mathcal{D}} L(\theta|x) \ . \tag{4}$$

In practice, the marginalisation required to express the incomplete-data likelihood may be intractable. We can simplify the problem by introducing auxiliary distributions over hidden variables, i.e. $Q_x(z) \equiv Q(z|x, \psi)$ for all $x \in \mathcal{D}$. It turns out that, for as long as the support of these distributions include the support of the true posterior $P(z|x, \theta)$, any distribution will do.

[1] Note that x and z may well themselves be collections of observed/hidden variables, e.g., $x = (x_1, \ldots, x_m)$ and $z = (z_1, \ldots, z_n)$.

$$
\begin{aligned}
\mathcal{L}(\theta|\mathcal{D}) &= \sum_{x\in\mathcal{D}} \log \sum_z P(x,z|\theta) && \text{(5a)}\\
&= \sum_{x\in\mathcal{D}} \log \sum_z Q_x(z) \frac{P(x,z|\theta)}{Q_x(z)} && \text{(5b)}\\
&\geq \sum_{x\in\mathcal{D}} \sum_z Q_x(z) \log \frac{P(x,z|\theta)}{Q_x(z)} && \text{(5c)}\\
&= \sum_{x\in\mathcal{D}} \left(\sum_z \log Q_x(z) \log P(x,z|\theta) - \sum_z \log Q_x(z) \log Q_x(z) \right) && \text{(5d)}\\
&= \sum_{x\in\mathcal{D}} \left(\mathbb{E}_{Q_x(Z)}[\log P(x,z|\theta)] + H(Q_x) \right) && \text{(5e)}\\
&\equiv \sum_{x\in\mathcal{D}} F(Q_x,\theta|x) && \text{(5f)}\\
&\equiv F(Q,\theta|\mathcal{D}) && \text{(5g)}
\end{aligned}
$$

In Equation (5c), we make use of Jensen's inequality to obtain a lowerbound on the log-likelihood. Note that the lowerbound is a function of the parameters of the generative model θ and of the auxiliary distributions $\{Q_x : \forall x \in \mathcal{D}\}$.

1 EM

Expectation-Maximisation (EM) (Dempster et al., 1977) can be viewed as a coordinate ascent algorithm that iteratively optimises $F(Q,\theta|\mathcal{D})$ (Neal and Hinton, 1998). In the E-step, EM maximises $F(Q,\theta|\mathcal{D})$ with respect to Q holding θ fixed. In the M-step, EM maximises $F(Q,\theta|\mathcal{D})$ with respect to θ holding Q fixed.

E-step

$$Q^{(t+1)} = \arg\max_Q F(Q,\theta^{(t)}|\mathcal{D}) \tag{6}$$

M-step

$$\theta^{(t+1)} = \arg\max_\theta F(Q^{(t+1)},\theta|\mathcal{D}) \tag{7}$$

The E-step attains an exact solution,

$$Q(z|x,\psi) = P(z|x,\theta^{(t)}), \quad \forall x \in \mathcal{D} \tag{8}$$

at which the bound F is tight (proof sketch: substitute Equation (8) into Equation (5) and show that the bound becomes an equality).

The M-step can be achieved simply by setting derivatives of Equation (7) with respect to θ to zero and solving for θ.

Because the bound is tight, at the beginning of each M-step we have that $F(Q, \theta|\mathcal{D}) = \mathcal{L}(\theta|\mathcal{D})$. Since the M-step climbs $L(\theta|\mathcal{D})$, the likelihood is guaranteed not to decrease after each combined EM step.

1.1 M-step in detail

Note that solutions to the M-step must define valid probability distributions, i.e. the optimisation problem is constrained.

$$\begin{aligned}
\theta^{(t+1)} = \arg\max_{\theta} \quad & F(Q, \theta|\mathcal{D}) \\
\text{where} \quad & P(x, z|\theta) = \theta_{z,x} \\
\text{s.t.} \quad & \sum_{x \in \mathcal{X}} P(x, z|\theta) = 1, \quad z \in \mathcal{Z} \\
& 0 \leq P(x, z|\theta) \leq 1, \quad (z, x) \in \mathcal{Z} \times \mathcal{X}
\end{aligned} \tag{9}$$

We can approach this constrained optimisation by introducing a Lagrange multiplier $\lambda \in \mathbb{R}$ and solving the following unconstrained problem (for simplicity we work with a single data sample).

$$\begin{aligned}
& \arg\max_{\theta,\lambda} \quad F(Q_x, \theta|x) - \lambda \left(\sum_{z,x} \theta_{z,x} - 1 \right) \\
& \equiv \arg\max_{\theta,\lambda} \quad h(\theta, \lambda)
\end{aligned} \tag{10}$$

First, set partial derivatives with respect to $\theta_{z,x}$ to zero,

$$\frac{\partial h(\theta, \lambda)}{\partial \theta_{z,x}} = \frac{\partial}{\partial \theta_{z,x}} \sum_{z'} Q_x(z') \log P(x, z'|\theta) - \lambda \sum_{z',x'} \frac{\partial}{\partial \theta_{z,x}} \theta_{z',x'} \tag{11a}$$

$$= \sum_{z'} Q_x(z') \frac{\partial}{\partial \theta_{z,x}} \log \theta_{z',x} - \lambda \sum_{z',x'} \mathbb{1}_{z,x}(z', x') \tag{11b}$$

$$= \sum_{z'} Q_x(z') \frac{1}{\theta_{z,x}} \mathbb{1}_z(z') - \lambda \sum_{z',x'} \mathbb{1}_{z,x}(z', x') \tag{11c}$$

$$= \frac{Q_x(z)}{\theta_{z,x}} - \lambda = 0 \ , \tag{11d}$$

which implies

$$\lambda = \frac{Q_x(z)}{\theta_{z,x}} \quad \text{and} \quad \theta_{z,x} = \frac{Q_x(z)}{\lambda} \ . \tag{12}$$

Then, set partial derivatives with respect to λ to zero,

$$\frac{\partial h(\theta, \lambda)}{\partial \lambda} = 0 - \sum_{z,x} \theta_{z,x} - 1 = 0 \ , \tag{13}$$

which implies

$$\sum_{z,x} \theta_{z,x} = \sum_{z,x} \frac{Q_x(z)}{\lambda} = 1 \quad \text{and} \quad \lambda = \sum_{z,x} Q_x(z) \ . \tag{14}$$

Finally, together, Equations (12) and (14) imply that,

$$\theta_{z,x} = \frac{Q_x(z)}{\sum_{z',x'} Q_{x'}(z')} \tag{15}$$

Obviously, this solution is only viable for models where the true posterior $P(z|x,\theta)$ can be computed efficiently, which happens when the marginalisation in Equation (1) is tractable.

2 Categorical distributions

Suppose the joint distribution $P(x, z|\theta)$ factors as a product of conditionally independent categorical distributions.

$$P(x, z|\theta) = \prod_{t=1}^{k} \prod_{c,o} \theta_{t,c,o}^{\#(t:c \to o|x,z)} \ , \tag{16}$$

where $t \in \{1, \ldots, k\}$ indexes one of k factors (here conditional distributions), $c \in \mathcal{C}$ indexes conditions, $o \in \mathcal{O}$ indexes outcomes, and $\#(t : c \to o|x, z)$ indicates how many times the event (c, o) of type t is observed in (x, z).

It is not too difficult to see that the solution to the M-step is

$$\theta_{t,c,o} = \frac{\mathbb{E}_Q[\#(t : c \to o|X, Z)]}{\sum_{o'} \mathbb{E}_Q[\#(t : c \to o'|X, Z)]} \ . \tag{17}$$

The key is to have more Lagrangian multipliers (one per categorical distribution) in the relaxed problem.

$$\arg\max_{\theta,\lambda} F(Q, \theta|\mathcal{D}) - \sum_{t,d} \lambda_{t,d} \left(\sum_{o} \theta_{t,d,o} - 1 \right) \tag{18}$$

Also note that when taking partial derivatives with respect to a parameter $\theta_{t,c,o}$ the following identities hold.

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{t,c,o}} \mathbb{E}_Q[\log P(X, Z|\theta)] &= \frac{\partial}{\partial \theta_{t,c,o}} \mathbb{E}_Q \left[\sum_{t'} \sum_{c',o'} \log \theta_{t',c',o'}^{\#(t':c' \to o'|X,Z)} \right] && \text{(19a)} \\
&= \frac{\partial}{\partial \theta_{t,c,o}} \mathbb{E}_Q \left[\sum_{t'} \sum_{c',o'} \#(t' : c' \to o'|X, Z) \log \theta_{t',c',o'} \right] && \text{(19b)} \\
&= \mathbb{E}_Q \left[\sum_{t'} \sum_{c',o'} \#(t' : c' \to o'|X, Z) \frac{\partial}{\partial \theta_{t',c',o'}} \log \theta_{t',c',o'} \right] && \text{(19c)} \\
&= \mathbb{E}_Q \left[\frac{1}{\theta_{t,c,o}} \#(t : c \to o|X, Z) \right] && \text{(19d)} \\
&= \frac{1}{\theta_{t,c,o}} E_Q \left[\#(t : c \to o|X, Z) \right] && \text{(19e)}
\end{aligned}
$$

The complete proof is left as exercise.

3 Logistic distributions

Suppose that for some factor t, the conditional distribution $P_t(O|C = c)$ is a logistic distribution

$$
P_t(O = o|C = c, w) = \frac{\exp(w^\top g(c, o))}{\sum_{o' \in \mathcal{O}} \exp(w^\top g(c, o))} , \tag{20}
$$

where $w \in \mathbb{R}^d$ is a vector of real-valued weights and $g : \mathcal{C} \times \mathcal{O} \to \mathbb{R}^d$ is a feature function.

In this case the solution to the M-step is simpler, since the optimisation problem is unconstrained in w (proof sketch: Equation (20) is a valid probability distribution for all $w \in \mathbb{R}^d$). We can approach the problem in Equation (7) by setting partial derivatives with respect to w_j to zero (without the need for Lagrangian multipliers).

To make the derivation simpler, let us make a change of variable,

$$
\theta_{t,c,o}(w) \equiv \frac{\exp(w^\top g(c, o))}{\sum_{o' \in \mathcal{O}} \exp(w^\top g(c, o))} , \tag{21}
$$

where $\theta_t(w)$ can be seen as a function from $\mathcal{C} \times \mathcal{O}$ to $\mathbb{R}$, and $P_t(O = o|C = c, \theta_t(w)) \equiv \theta_{t,c,o}(w)$. We can obtain partial derivatives with respect to w_j via the chain rule for derivatives, i.e. $\frac{\partial}{\partial w_j} P_t(o|c, \theta_t(w)) = \frac{\partial}{\partial \theta} P_t(o|c, \theta_t(w)) \times \frac{\partial}{\partial w_j} \theta_{t,c,o}(w)$.

$$\frac{\partial}{\partial w_j}\mathbb{E}_Q[\log P_t(X,Z|\theta_t(w))] \tag{22a}$$

$$= \frac{\partial}{\partial w_j}\mathbb{E}_Q\left[\sum_{t'}\sum_{c',o'}\log\theta_{t',c',o'}(w)^{\#(t':c'\to o'|X,Z)}\right] \tag{22b}$$

$$= \frac{\partial}{\partial w_j}\mathbb{E}_Q\left[\sum_{t'}\sum_{c',o'}\#(t':c'\to o'|X,Z)\log\theta_{t',c',o'}(w)\right] \tag{22c}$$

$$= \mathbb{E}_Q\left[\sum_{t'}\sum_{c',o'}\#(t':c'\to o'|X,Z)\frac{\partial}{\partial w_j}\log\theta_{t',c',o'}(w)\right] \tag{22d}$$

$$= \mathbb{E}_Q\left[\#(t:c\to o|X,Z)\frac{\partial}{\partial w_j}\left(\log\exp(w^\top g(c,o)) - \log\sum_{o'}\exp(w^\top g(c,o'))\right)\right] \tag{22e}$$

$$= \mathbb{E}_Q\left[\#(t:c\to o|X,Z)\frac{\partial}{\partial w_j}\left(w^\top g(c,o) - \log\sum_{o'}\exp(w^\top g(c,o'))\right)\right] \tag{22f}$$

$$= \mathbb{E}_Q\left[\#(t:c\to o|X,Z)\left(g_j(c,o) - \frac{\frac{\partial}{\partial w_j}\sum_{o'}\exp(w^\top g(c,o'))}{\sum_{o'}\exp(w^\top g(c,o'))}\right)\right] \tag{22g}$$

$$= \mathbb{E}_Q\left[\#(t:c\to o|X,Z)\left(g_j(c,o) - \frac{\sum_{o'}\exp(w^\top g(c,o'))g_j(c,o')}{\sum_{o'}\exp(w^\top g(c,o'))}\right)\right] \tag{22h}$$

$$= \mathbb{E}_Q\left[\#(t:c\to o|X,Z)\left(g_j(c,o) - \sum_{o'}\theta_{t,c,o'}(w)g_j(c,o')\right)\right] \tag{22i}$$

$$= \left(g_j(c,o) - \sum_{o'}\theta_{t,c,o'}(w)g_j(c,o')\right)E_Q\left[\#(t:c\to o|X,Z)\right] \tag{22j}$$

Finally, for small enough γ, the update

$$\theta^{(t+1)}_{t,c,o} = \theta^{(t)}_{t,c,o} + \gamma\frac{\partial}{\partial w_j}\mathbb{E}_Q[\log P_t(X,Z|\theta_t(w))] \tag{23}$$

is guaranteed not to the decrease the log-likelihood.

4 Remarks

- Neal and Hinton (1998) present EM as a coordinate ascent;

- Salakhutdinov et al. (2003) present Expectation-Conjugate-Gradient (ECG), a gradient-based algorithm for direct likelihood optimisation, they also present conditions under which ECG outperforms EM;
- Berg-Kirkpatrick et al. (2010) present unsupervised learning (both EM and ECG) for logistic CPDs to the NLP community.

References

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California. Association for Computational Linguistics.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Neal, R. M. and Hinton, G. E. (1998). *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*, pages 355–368. Springer Netherlands, Dordrecht.

Salakhutdinov, R., Roweis, S., and Ghahramani, Z. (2003). Optimization with em and expectation-conjugate-gradient. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 672–679. AAAI Press.