Expectation Maximisation Algorithm

Boris Flach, Czech Technical University in Prague

Vaclav Hlavac, Czech Technical University in Prague

Synonyms

– EM-Algorithm

Related Concepts

– Maximum likelihood estimation
– Bayesian inference
– Unsupervised learning

Definition

The Expectation Maximisation Algorithm iteratively maximises the likelihood of a training sample with respect to unknown parameters of a probability model under the condition of missing information. The training sample is assumed to represent a set of independent realisations of a random variable defined on the underlying probability space.

Background

One of the main paradigms of statistical pattern recognition and Bayesian inference is to model the relation between the observable features $x \in \mathcal{X}$ of an object and its hidden state $y \in \mathcal{Y}$ by a joint probability measure $p(x, y)$. This probability measure is, however, often known only up to some parameters $\theta \in \Theta$. It is thus necessary to estimate these parameters from a training sample, which is assumed to represent a sequence of independent realisations of a random variable. If, ideally, these are realisations of pairs (x, y), then the corresponding estimation methods are addressed as *supervised* learning. It is, however, quite common that some of those variables describing the hidden state are latent. These latent variables are never observed in the training data. Therefore, it is necessary to marginalise over them in order to estimate the unknown parameters θ. Corresponding estimation methods are known as *unsupervised* learning. Moreover, especially in computer vision, the observation x *and* the hidden state y both may have a complex structure. The latter can be e.g., a segmentation, a depth map or a similar object. Consequently, it is often not feasible to provide the complete information y for the realisations in the sample. This means to estimate the parameters in the situation of missing information. The EM-algorithm is a method searching for maximum likelihood estimates of the unknown parameters under such conditions.

Theory

All the situations described in the previous section can be treated in a uniform way by assuming the training sample as a set of independent realisations of a random variable.

Let Ω be a finite sample space, $\mathcal{F}$ be its power set and $p_\theta \colon \mathcal{F} \to \mathbb{R}_+$ be a probability measure defined up to unknown parameters $\theta \in \Theta$. Let $X \colon \Omega \to \mathcal{X}$ be a random variable and $T = (x_1, x_2, \ldots, x_n)$ be a sequence of independent realisations of X (see e.g. [1,2] for a formal definition of independent realisations). The Maximum Likelihood estimator provides estimates of the unknown parameters θ by maximising the probability of T

$$\theta^* = \operatorname*{argmax}_\theta \prod_{i=1}^{n} p_\theta(\Omega_i), \tag{1}$$

where Ω_i denotes the pre-image $\{\omega \in \Omega \mid X(\omega) = x_i\}$. If the logarithm is taken, the task reads equivalently

$$\theta^* = \operatorname*{argmax}_\theta L(x_1, \ldots, x_n, \theta) = \operatorname*{argmax}_\theta \sum_{i=1}^{n} \log \sum_{\omega \in \Omega_i} p_\theta(\omega). \tag{2}$$

Remark 1. It is often assumed that Ω is a Cartesian product $\Omega = \mathcal{X} \times \mathcal{Y}$ and that X simply projects onto the first component $X(x, y) = x$. Then the probability $p_\theta(\Omega_i) = \sum_{y \in \mathcal{Y}} p_\theta(x_i, y)$ is nothing but the marginalisation over all possible y. This special case will be considered in an example below. □

The optimisation task (2) is often complicated and hardly solvable by standard optimisation methods – either because the objective function is not concave or because θ represents a set of parameters of different natures. Suppose, however, that the task of parameter estimation is feasible if complete information, i.e. a set of realisations of $\omega \in \Omega$ is available. This applies in particular if the corresponding simpler objective function $\sum_i \log p_\theta(\omega_i)$ is concave with respect to θ or if the task decomposes into simpler, independent optimisation tasks with respect to individual components of a parameter collection.

The key idea of the Expectation Maximisation algorithm is to exploit this circumstance and to solve the optimisation task (2) by iterating the following two feasible tasks:

(1) given a current estimate of θ, determine the missing information, i.e., $p_\theta(\omega|\Omega_i)$ for each element $x_i \in T$ and
(2) given the complete information, solve the corresponding estimation task, resulting in an improved estimate of θ.

To further substantiate this idea of "iterative splitting" of the task (2), it is convenient to introduce non-negative auxiliary variables $\alpha_i(\omega)$, $\omega \in \Omega_i$, for each element x_i of the learning sample T such that they fulfil

$$\sum_{\omega \in \Omega_i} \alpha_i(\omega) = 1, \quad \forall i = 1, 2, \ldots, n. \tag{3}$$

These variables α_i can be seen as (so far arbitrary) posterior probabilities $p(\omega|\Omega_i)$ for $\omega \in \Omega_i$, given a realisation x_i. The log-likelihood of a realisation x_i can be written by their use as

$$\log p_\theta(\Omega_i) = \sum_{\omega\in\Omega_i} \alpha_i(\omega) \log p_\theta(\Omega_i) = \\ = \sum_{\omega\in\Omega_i} \alpha_i(\omega) \log p_\theta(\omega) - \sum_{\omega\in\Omega_i} \alpha_i(\omega) \log \frac{p_\theta(\omega)}{p_\theta(\Omega_i)}, \quad (4)$$

where the first equality follows directly from (3). The log-likelihood of the training sample can be therefore expressed equivalently as

$$L(x_1,\ldots,x_n,\theta) = \sum_{i=1}^{n} \sum_{\omega\in\Omega_i} \alpha_i(\omega) \log p_\theta(\Omega_i) = \\ = \sum_{i=1}^{n} \sum_{\omega\in\Omega_i} \alpha_i(\omega) \log p_\theta(\omega) - \sum_{i=1}^{n} \sum_{\omega\in\Omega_i} \alpha_i(\omega) \log p_\theta(\omega|\Omega_i). \quad (5)$$

The expression as a whole does not depend on the specific choice of the auxiliary variables α, whereas the minuend and subtrahend do. Moreover, note that the minuend is nothing but the likelihood of a sample of complete data, if the α are interpreted as the missing information, i.e., posterior probabilities for $\omega \in \Omega_i$ given the observation x_i.

Starting with some reasonable choice for the initial $\theta^{(0)}$ the likelihood is iteratively increased by alternating the following two steps. The (E)xpectation step calculates new α such that whatever new θ will be chosen subsequently, the subtrahend will not increase. The (M)aximisation step relies on this and maximises the minuend only, avoiding to deal with the subtrahend.

$$\text{E-step} \qquad \alpha_i^{(t)}(\omega) = p_{\theta^{(t)}}\left(\omega|\Omega_i\right) \quad (6)$$

$$\text{M-step} \qquad \theta^{(t+1)} = \operatorname*{argmax}_\theta \sum_{i=1}^{n} \sum_{\omega\in\Omega_i} \alpha_i^{(t)}(\omega) \log p_\theta(\omega). \quad (7)$$

From the conceptual point of view the E-step can be seen as inference – it calculates the missing data, i.e., the posterior probabilities $p_{\theta^{(t)}}\left(\omega|\Omega_i\right)$ for each element x_i in the training sample. The M-step utilises these posterior probabilities for a supervised learning step. The names themselves stem from a rather formal view: the E-step calculates the α and therefore the objective function in (7) which has the form of an expectation of $\log p_\theta(\omega)$. The computation of this objective function is sometimes considered to be a part of the E-step. The name for the M-step is obvious.

It is easy to see that the likelihood is monotonically increasing: The choice (6) for α guarantees that the subtrahend in (5) can only decrease whatever new θ will be chosen in the subsequent M-step. This follows from the inequality

$$\sum_{\omega\in\Omega_i} p_\theta(\omega|\Omega_i) \log p_{\theta'}(\omega|\Omega_i) \leqslant \sum_{\omega\in\Omega_i} p_\theta(\omega|\Omega_i) \log p_\theta(\omega|\Omega_i) \;\; \forall\theta' \neq \theta. \quad (8)$$

Since the M-step chooses the new θ so as to maximise the minuend, the likelihood can only increase (or stay constant). Another convenient way to prove monotonicity of the EM algorithm can be found in [3,4]. These tutorials consider the EM algorithm as the maximisation of a lower bound of the likelihood.

It remains unclear whether the global optimum of the likelihood is reached in a fix-point of the algorithm. Moreover, it happens quite often that the M-step is infeasible for complex models p_θ. Then a weaker form of the EM algorithm is used by choosing $\theta^{(t+1)}$ so as to guarantee an increase of the objective function of the M-step.

The derivation of the concept of the EM algorithm was given here for a discrete probability space and discrete random variables. It can be however generalised for uncountable probability spaces and random variables X with continuous probability density.

Example The EM-algorithm is often considered for the following special case. The sampling space Ω is a Cartesian product $\Omega = \mathcal{X} \times \mathcal{Y}$ and the random variable X simply projects onto the first component $X(x, y) = x$. The parameters $\theta \in \Theta$ of the probability $p_\theta(x, y)$ are to be estimated given a sequence of independent realisations of x. In this special case, the log-likelihood has the form

$$L = \sum_{i=1}^{n} \log \sum_{y \in \mathcal{Y}} p_\theta(x_i, y). \tag{9}$$

Its decomposition (5) is

$$L = \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \alpha_i(y) \log p_\theta(x_i, y) - \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \alpha_i(y) \log p_\theta(y|x_i). \tag{10}$$

The EM-algorithm itself then reads

$$\text{E-step} \qquad \alpha_i^{(t)}(y) = p_{\theta^{(t)}}(y|x_i) \tag{11}$$

$$\text{M-step} \qquad \theta^{(t+1)} = \operatorname*{argmax}_\theta \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \alpha_i^{(t)}(y) \log p_\theta(x_i, y). \tag{12}$$

History and applications The classic paper [5] is often cited as the first one introducing the EM algorithm in its general form. It should be noted, however, that the method was introduced and analysed substantially earlier for a broad class of pattern recognition tasks in [6] and for exponential families in [7].

A comprehensive discussion of the EM algorithm can be found in [8] and in the context of pattern recognition in [9,10]. Standard application examples are parameter estimation for mixtures of Gaussians [8] and the Mean Shift algorithm [11]. Another important application is parameter estimation for Hidden Markov Models. This model class is extensively used for automated speech recognition. The corresponding EM algorithm is known as Baum-Welch algorithm in

this context [12]. Rather complex applications of the EM algorithm arise in the context of parameter estimation for Markov Random Fields [13].

Recommended Readings

[1] Meintrup, D., Schäffler, S. (2005). Stochastik. Springer
[2] Papoulis, A. (1990). Probability and Statistics. Prentice-Hall
[3] Minka, T. (1998). Expectation-maximization as lower bound maximization. tutorial, MIT
[4] Dellaert, F. (2002). The expectation maximization algorithm. Technical Report GIT-GVU-02-20, Georgia Institute of Technology
[5] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**(1) 1–38
[6] Schlesinger, M.I. (1968). The interaction of learning and self-organization in pattern recognition. Kibernetika **4**(2) 81–88
[7] Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scandinavian Journal of Statistics **1**(2) 49–58
[8] McLachlan, G.J., Krishnan, T. (1997). The EM Algorithm and Extensions. John Wiley & Sons
[9] Schlesinger, M.I., Hlavac, V. (2002). Ten Lectures on Statistical and Structural Pattern Recognition. Kluwer Academic Press
[10] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer
[11] Cheng, Y. (1995). Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(8) 790 – 799
[12] Jelinek, F. (1998). Statistical methods for speech recognition. MIT Press
[13] Li, S.Z. (2009). Markov Random Field Modeling in Image Analysis. 3. edn. Advances in Pattern Recognition. Springer Verlag