expectation–maximization (EM)

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

The expectation–maximization (EM) algorithm is an iterative optimization method for approximately solving maximum likelihood optimization problems that are difficult to solve directly, such as fitting a Gaussian mixture model (GMM). Each iteration alternates an E-step, which computes the posterior distribution of an unobserved auxiliary attribute under the current model parameters, and an M-step, which minimizes a surrogate objective function derived from this posterior distribution. The surrogate upper-bounds the negative log-likelihood and is tight at the current iterate, so each iteration never increases the negative log-likelihood: EM is a majorize–minimize (MM) method. The method is a fixed-point iteration whose fixed points are model parameters that minimize their own surrogate.

Definition

The nightly minimum temperatures recorded over one year at the weather station of Krems an der Donau (Austria) scatter around two regimes — cold-season and warm-season nights. A Gaussian mixture model (GMM) with two components captures such a two-regime distribution, but maximizing its likelihood has no closed-form solution. The standard method for fitting it is the EM algorithm (Dempster et al., 1977): an iterative optimization method for approximately solving certain maximum likelihood optimization problems that are difficult to solve directly (Bishop, 2006, Sec. 9.4; Murphy, 2012, Sec. 11.4.7). To motivate the EM algorithm and explain its construction, consider a machine learning (ML) application involving a single observed data point with feature $x \in \mathcal{X}$, where $\mathcal{X}$ is a finite feature space. The data generation is modeled via a probabilistic model that consists of a random variable (RV) x' with a probability mass function (pmf) $p^{(x')}(\cdot; \mathbf{w})$. Here, the actual model parameters $\mathbf{w} \in \mathcal{W}$—used for the data generation via sampling from $p^{(x')}(\cdot; \mathbf{w})$—are unknown. A widely used approach for estimating these model parameters is via the solutions of the following maximum likelihood problem:

$$\min_{\mathbf{w} \in \mathcal{W}} - \log p^{(x')}(x; \mathbf{w}) . \tag{1}$$

For some probabilistic models, such as a GMM, this optimization problem can be difficult to solve directly. As a work-around, one can often introduce an auxiliary attribute $y \in \mathcal{Y}$, generated via some RV y'. For a suitable choice of

this attribute, the corresponding probabilistic model $p^{(x',y')}(\cdot,\cdot;\mathbf{w})$ yields the following, much easier maximum likelihood problem:

$$\min_{\mathbf{w}\in\mathcal{W}} -\log p^{(x',y')}(x,y;\mathbf{w})\,. \tag{2}$$

The attribute y is introduced solely to simplify (2), but it is not observed in practice—only the feature x is available. Thus, (2) cannot be solved directly, as the value of y to plug into the pmf is unknown $p^{(x',y')}(x,y;\mathbf{w})$. The EM method resolves this dilemma by alternating between two steps. The E-step computes a "soft" estimate of the auxiliary attribute y: the posterior distribution $p^{(y'|x')}(\cdot;\widehat{\mathbf{w}})$, i.e., the probability of each value of y given the observed feature, under the current choice $\widehat{\mathbf{w}}$ for the model parameters. The M-step minimizes a surrogate objective function derived from this posterior distribution. Together, one E-step and one M-step constitute one full iteration of the EM method. In more detail, the E-step produces the following function:

$$Q(\mathbf{w}\mid\widehat{\mathbf{w}}) := -\sum_{y\in\mathcal{Y}} p^{(y'|x')}(y;\widehat{\mathbf{w}})\log\Big(p^{(x',y')}(x,y;\mathbf{w})\,/p^{(y'|x')}(y;\widehat{\mathbf{w}})\Big),$$

and the M-step minimizes $Q(\mathbf{w}\mid\widehat{\mathbf{w}})$ over $\mathbf{w}\in\mathcal{W}$. This function satisfies the following two key properties (Bishop, 2006, Sec. 9.4; Murphy, 2012, Sec. 11.4.7): 1) upper bound $Q(\mathbf{w}\mid\widehat{\mathbf{w}}) \geq -\log p^{(x')}(x;\mathbf{w})$ for all $\mathbf{w}\in\mathcal{W}$; and 2) tightness $Q(\widehat{\mathbf{w}}\mid\widehat{\mathbf{w}}) = -\log p^{(x')}(x;\widehat{\mathbf{w}})$. To summarize, during each iteration, EM minimizes an upper-bounding surrogate objective function that is tight at the current iterate $\widehat{\mathbf{w}}$. Thus, EM is a majorize–minimize (MM) method for approximately solving (1).

The EM method is also a fixed-point iteration $\widehat{\mathbf{w}}^{(t+1)} = \mathcal{F}\big(\widehat{\mathbf{w}}^{(t)}\big)$ with the operator

$$\mathcal{F}(\widehat{\mathbf{w}}) := \arg\min_{\mathbf{w}\in\mathcal{W}} Q(\mathbf{w}\mid\widehat{\mathbf{w}}).$$

By the upper-bound and tightness properties above, each application of $\mathcal{F}$ never increases the objective function of (1) — the negative log-likelihood — so monotone descent, rather than a contractive operator property, justifies convergence. The fixed points of $\mathcal{F}$ are model parameters that minimize their own surrogate: there, the surrogate touches the negative log-likelihood from above, so no further descent step is available to the method. The above construction and analysis of EM can be extended to more general settings involving multiple data points and infinite feature spaces such as $\mathbb{R}^d$ (see Murphy (2012), Sec. 11.4.7 for further details).

Fig. 1 shows EM at work on real data: a GMM with two components, fitted to the 366 nightly minimum temperatures of 2024 at Krems, recovers the cold-season and warm-season regimes, and the negative log-likelihood descends monotonically to a fixed point.

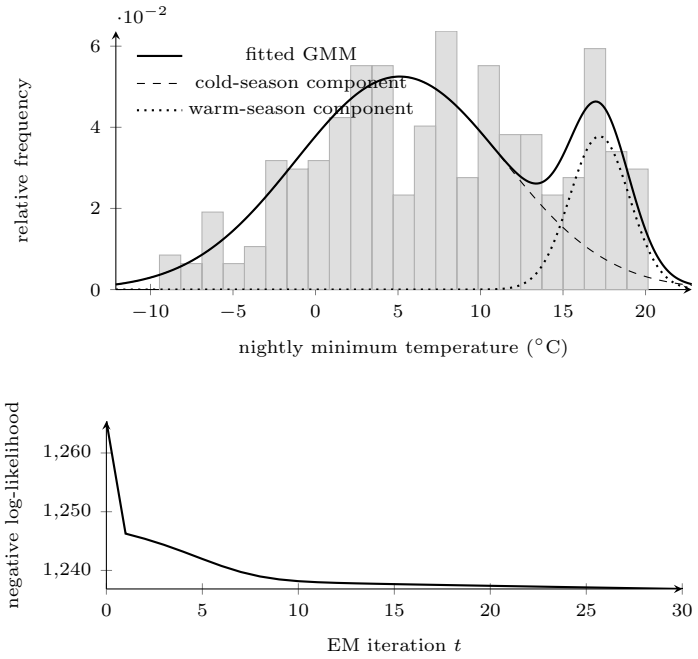

Figure 1: EM fitting a GMM with two components to the 366 nightly minimum temperatures of 2024 at the weather station Krems. Top: histogram of the temperatures with the fitted mixture and its two weighted components. Bottom: the negative log-likelihood never increases and settles at a fixed point of the EM operator $\mathcal{F}$. Data generated by `pythondemos/em.py`.

Cross References

maximum likelihood, GMM, MM, fixed point, posterior distribution, optimization problem, probabilistic model

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

3. Dempster et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc.: Ser. B (Methodological), 39(1), pp. 1–22.