Expectation Maximization [2] in Detail

Qi Zhao

August 29, 2018

1. Introduction

The EM algorithm is used to find maximum likelihood [3] or MAP estimates of model parameters θ where the likelihood Pr(x|θ) of the data x can be written as in Equation 1.

$$Pr(x|\theta) = \sum_k Pr(x, h=k|\theta) = \sum_k Pr(x|h=k,\theta)Pr(h=k)$$
$$Pr(x|\theta) = \int Pr(x, h|\theta)dh = \int Pr(x|h,\theta)Pr(h)dh \quad (1)$$

Where for discrete and continuous hidden variables, respectively. In other words, the likelihood Pr(x|θ) is a marginalization of a joint distribution [4] over the data and the hidden variables.

2. Descriptions

The EM algorithm relies on the idea of a lower bounding function (or lower bound), B[θ] on the log likelihood. This is a function of the parameters θ that is always guaranteed to be equal to or lower than the log likelihood [1]. The lower bound is carefully chosen so that it is easy to maximize with respect to the parameters. This lower bound is also parameterized by a set of probability distributions $\{q_i\ (h_i)\}_{i=1}^{I}$ over the hidden variables, so we write it as B[$\{q_i\ (h_i)\}$, θ]. Different probability distributions $q_i\ (h_i)$ predict different lower bounds B[$\{q_i\ (h_i)\}$, θ] and hence different functions of θ that lie everywhere below the true log likelihood (Figure 1).

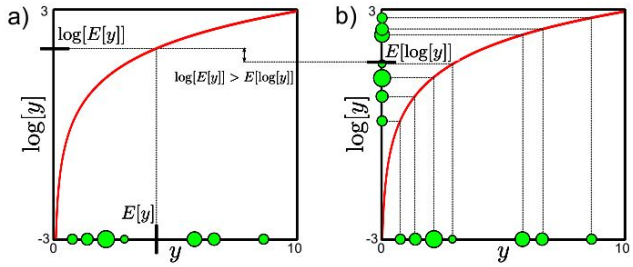

Figure 1. Jensens inequality for the logarithmic function (discrete case). a) Taking a weighted average of examples E[y] and passing them through the log function. b) Passing the samples through the log function and taking a weighted average E[log[y]]. The latter case always produces a smaller value than the former (E[log[y]] $\leq$ log(E[y])): higher valued examples are relatively compressed by the concave log function.

By iterating steps, the (local) maximum of the actual log likelihood is approached (Figure 2). To complete our picture of the EM algorithm, it must:

1.define B[$\{q_i\ (h_i)\}$, $\theta^{[t?1]}$] and show that it always lies below the log likelihood;

2.show which probability distribution $q_i\ (h_i)$ optimizes the bound in the E-step;

3.show how to optimize the bound with respect to θ in the M-step;

References

[1] A. Abou-Elailah, I. Bloch, and V. Gouet-Brunet. Unsupervised detection of ruptures in spatial relationships in video sequences based on log likelihood ratio. *Pattern Analysis & Applications*, 21(3):829–846, 2018. 1

[2] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1-2):51–80, 1995. 1

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

[4] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *IEEE International Conference on Computer Vision*, pages 2200–2207, 2014. 1

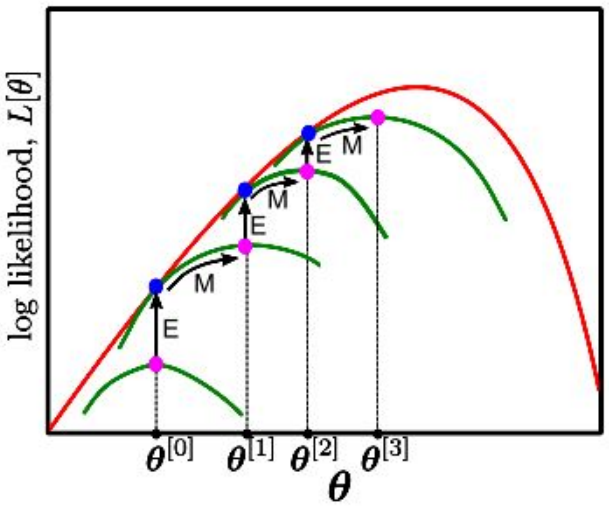

Figure 2. Expectation maximization algorithm. We iterate the expectation and maximization steps by alternately changing the distributions q_i (h_i) and the parameter θ so that the bound increases. In the E-step, the bound is maximized with respect to q_i (h_i) for fixed parameters θ: the new function with respect to touches the true log likelihood at θ. In the M-step, we find the maximum of this function. In this way we are guaranteed to reach a local maximum in the likelihood function.