Expectation Maximization [3]

Qi Zhao

August 7, 2018

1. Introduction

The goal of the expectation maximization (EM) algorithm is to provide just enough information to use this technique for fitting models. The EM algorithm is a general-purpose tool for fitting parameters θ in models of the form of Equation 1 where the EM algorithm works by defining a lower bound B[$\{q_i(h_i)\}$, θ] on the log likelihood in Equation 2 and iteratively increasing this bound. The lower bound is simply a function that is parameterized by θ and some other quantities and is guaranteed to always return a value that is less than or equal to the log likelihood L[θ] for any given set of parameters θ (see Figure 1).
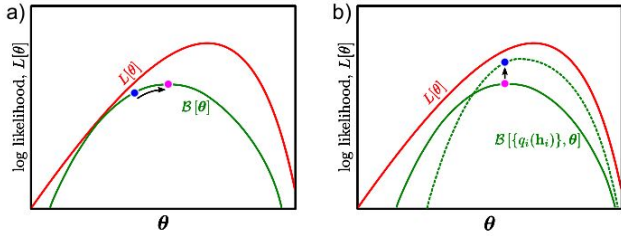

Figure 1. Manipulating the lower bound. a) Consider the log likelihood L[θ] of the data $\{x\}_{i=1}^{I}$ as a function of the model parameters θ (red curve). In maximum likelihood [1] learning our goal is to find the parameters θ that maximize this function. A lower bound on the log likelihood is another function B[θ] of the parameters that is everywhere lower or equal to the log likelihood (green curve). One way to improve the current estimate (blue dot) is to manipulate the parameters so that B[θ] increases (purple dot). This is the goal of the maximization step of the EM algorithm. b) The lower bound B [$\{q_i(h_i)\}_{i=1}^{I}$, θ] also depends on a set of probability distributions $\{q_i(h_i)\}_{i=1}^{I}$ over hidden variables $\{h_i\}$. Manipulating these probability distributions [2] changes the value that the lower bound returns for every θ (e.g., green curve). So a second way to improve the current estimate (purple dot) is to change the distributions in such a way that the curve increases for the current parameters (blue dot). This is the goal of the expectation step of the EM algorithm.

$$Pr(x|\theta) = \int Pr(x, h|\theta)dh \qquad (1)$$

$$\hat{\theta} = argmax_\theta[\sum_{i=1}^{I} log[\int Pr(x_i, h_i|\theta)dh_i]] \qquad (2)$$

2. Descriptions

For the EM algorithm, the particular lower bound chosen is in Equation 3.

$$B[q_i(h_i), \theta] = \sum_{i=1}^{I} \int q_i(h_i) log[\frac{Pr(x_i, h_i|\theta)}{q_i(h_i)}]dh_i$$
$$\leq \sum_{i=1}^{I} log[\int Pr(x_i, h_i|\theta)dh_i] \qquad (3)$$

It is not obvious that this inequality is true, making this a valid lower bound;

In addition to the parameters θ, the lower bound B[$\{q_i(h_i)\}$, θ] also depends on a set of I probability distributions $\{q_i(h_i)\}_{i=1}^{I}$ over the hidden variables $\{h_i\}_{i=1}^{I}$. When it varies these probability distributions, the value that the lower bound returns will change, but it will always remain less than or equal to the log likelihood.

The EM algorithm manipulates both the parameters θ and the distributions $\{q_i(h_i)\}_{i=1}^{I}$ to increase the lower bound. It alternates between:

1.updating the probability distributions $\{q_i(h_i)\}_{i=1}^{I}$ to improve the bound in Equation 3. This is called the expectation step or E-step. 2.updating the parameters θ to improve the bound in Equation 3. This is called the maximization step or M-step. Each of these steps is guaranteed to improve the bound, and iterating them alternately is guaranteed to find at least a local maximum with respect to θ.

References

[1] 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

[2] E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *International Conference on Machine Learning and Natural Language Processing*, pages 255–262, 2000. 1

[3] T. K. Moon. *The expectation maximization algorithm.* Springer, 2000. 1