CS 410: Technology Review
Expectation Maximization

ritikd2

November 7, 2021

One of the advantages of probabilistic methods has been their efficiency and robustness in learning parameters of the model from observations. The parameters are typically used to model the joint probability distribution from which the observed data was generated. Being able to do so has a host of applications in different fields, from text mining to computational biology.

Estimating the parameters of a model by relying on observed data falls under the idea of Maximum Likelihood Estimation. For eg, let's suppose there are two jars filled with different proportions of blue and red candies. We can represent the probability of picking a red candy from jar A as θ_A and the probability of picking a red candy from jar B as θ_B. We wish to estimate θ_A, θ_B.

We do this by repeating the following procedure for 10 times: with equal probability, we pick a jar randomly (with equal probability), and then pick a candy 10 times from the chosen jar. The entire procedure thus involves us picking 100 candies.

We keep track of our observations by maintaining two vectors, $x = (x_{1,x_2,\ldots,x_{10}})$ and $z = (z_1, z_2, \ldots, z_{10})$, where $x_i \in 0, 1, \ldots, 10$ is the number of red candies picked in the i^{th} iteration, and $z_i \in A, B$ is the jar picked in the i^{th} iteration. Given the above information, we can infer the probabilities as:

$$\theta_A = \frac{\text{\# red candies from jar A}}{\text{total \# of candies picked from jar A}}$$
$$\theta_B = \frac{\text{\# red candies from jar B}}{\text{total \# of candies picked from jar B}}$$

The method of parameter estimation is called Maximum Likelihood Estimation (MLE). The above method is useful, but there could be cases in which we might not be aware of the values of some data that is being used to generate the actual distribution. For instance, in the above example suppose we do not collect z_i's, i.e., which jar were the candies coming from.

In such cases, we usually try to follow an iterative method. We start from some initial estimates of the parameters ${\theta_A^t}', {\theta_B^t}'$, use these to assign labels to the observation data (determine whether jar A or jar B was more likely to generate a set of observations) and then assume those probabilities to be true. We then use estimated values to apply regular MLE to get $\theta_A^{t+1\prime}, \theta_B^{t+1\prime}$. Such an iterative method to estimate parameters in the presence of latent (hidden) parameters is called Expectation Maximization.

Thus, expectation maximization alternates between the following two steps:

1. Guessing a probability distribution of missing data given the current parameters (*E-step*)
2. Re-estimating the model parameters used these guessed distributions (*M-step*)

The above steps are repeated one after the other until some convergence criteria is met (for eg, the estimated parameters do not change over two iterations, or the change is negligible).

The name *E-step* comes from the fact that we do not need to guess the exact probability distributions, rather just the expected values. The name *M-step* comes from the fact that we re-estimate the parameters by maximizing the expected log-likelihood of generating the data.

In some sense the EM algorithm is the generalization of the MLE algorithm to the case where observed data is incomplete. It should be noted that the EM algorithm converges to a local optimum, and not necessarily the global one. This means that depending on the initial conditions it is possible to reach an entirely different solution. There are also no good bounds on the convergence rate of the basic EM algorithm, i.e., it is possible for the algorithm to continuously iterate for a significant amount of time before it finally reaches a local optimum. Good starting points (that is, the initial estimates) therefore need to be chosen carefully.

There are many variations on the basic EM algorithm that provide better guarantees in terms of either converging faster or providing the global optimum. The EM algorithm can be further generalized as well. When trying to estimate the parameters, rather than trying to maximize likelihood, we can simply choose parameters that the likelihood is non-decreasing. This is called the Generalized Expectation Maximization (GEM), and is useful in cases where maximization might be difficult [2].

Expectation Maximization has a host of use cases from document classification to computation biology applications such as finding motifs when given a set of unaligned DNA sequences [1].

References

[1] Do, C., Batzoglou, S. What is the expectation maximization algorithm?. Nat Biotechnol 26, 897–899 (2008). https://doi.org/10.1038/nbt1406

[2] Borman, Sean. "The expectation maximization algorithm-a short tutorial." Submitted for publication 41 (2004).