CMPT 825: Natural Language Processing **Spring 2008**

Lecture 6 — Feb 18-22, 2008

Lecturer: Anoop Sarkar *Scribe: Stephen Fagan*

The EM Algorithm

The Expectation-Maximization (EM) algorithm is a general method for deriving maximum likelihood parameter estimates from incomplete (i.e. partially unobserved) data. The algorithm is a two-step iterative method that begins with an initial guess of the model parameters, θ. In the first step (the E or Expectation step), the observed data and the current estimate of θ is used to find the distribution of the unobserved data. In the second step (the M or Maximization step), a re-estimation of the θ parameters is performed under the assumption that the distribution of unobserved data from the previous step is the true distribution. So, the idea behind each step of the algorithm is:

Expectation: If we knew the value of θ, then we could compute the distribution of the hidden data of the model.

Maximization: If we knew the the distribution of the hidden data, then we could compute the maximum likelihood value of θ.

It will be shown that iterating between the E-step and M-step improves the likelihood of the estimates of θ.

The EM algorithm is a very general method with many applications. It was first presented in its general form in 1977 by Dempster, Laird, and Rubin. Before presenting the algorithm, several examples are presented to secure an intuitive understanding.

6.1 Example - Three Coins

The setup of this example is as follows: There are three possibly biased coins numbered 0, 1, and 2. A sequence of heads (H) and tails (T) from coins 1 and 2 are generated by first tossing coin 0, and if coin 0 shows H, then three

observations are generated by tossing coin 1 three times. If coin 0 shows T, then three observations are generated by tossing coin 2. This process of tossing coin 0 to determine which of coins 1 or 2 will generate the next three observations is repeated several times. Note that the results of tossing coin 0 is not a part of the observation sequence - this is the "hidden data." An example of output from this setup is:

$$HHH, TTT, HHH, TTT, HHH.$$

The hidden component of the data is the results of tossing coin 0. Suppose that the hidden data observed from the above observed sequence is

$$H, T, H, T, H.$$

There are three model parameters to be estimated in this model

$$\theta = (\lambda, p_1, p_2)$$

where λ is the probability of coin 0 showing H (so $1 - \lambda$ is the probability of it showing T), p_1 is the probability of coin 1 showing H, and p_2 is the probability of coin 2 showing H.

Three Coins - The Fully Observed Case: In order to fully grasp the impact of having hidden data, the simple case of no hidden data is first explored. This case is simple since an estimate of θ is derived from the counts of the data:

$$\hat{\lambda} = \frac{count(coin_0 = H)}{count(tosses\_of\_coin_0)} = \frac{3}{5}$$

$$\hat{p}_1 = \frac{count(coin_1 = H)}{count(tosses\_of\_coin_1)} = \frac{9}{9} = 1$$

$$\hat{p}_2 = \frac{count(coin_2 = H)}{count(tosses\_of\_coin_2)} = \frac{0}{6} = 0$$

The hats, $\hat{}$, over the parameters indicate that these are estimates of the true values based on the available data. The difficulty of the case of hidden data comes from the fact that we don't know which coin is being tossed throughout the sequence of observed data.

Three Coins - Hidden Data, but θ is Known: In the case where the results of tossing coin 0 are hidden, then it is uncertain which coin (1 or 2) generated each result in the observed sequence. However, if θ was known, then we could compute the distribution of the hidden data. Let $x = THT$ be a sequence of observed coin-toss data from a single coin, and let y be the hidden value of the coin 0 toss that determined the coin for those tosses (so $y = H$ or $y = T$). Assume that the parameters $\theta = (\lambda, p_1, p_2)$ are known. Then given θ and the observed data $x = THT$, the probability distribution of y can be calculated as follows:

$$\begin{aligned}
\Pr(y = H|x = THT, \theta) &= \frac{\Pr(x{=}THT, y{=}H|\theta)}{\Pr(x{=}THT|\theta)} \\
&= \frac{\Pr(x{=}THT, y{=}H|\theta)}{\Pr(x{=}THT, y{=}H|\theta) \cdot \Pr(x{=}THT, y{=}T|\theta)} \\
&= \frac{\lambda p_1(1-p_1)^2}{\lambda p_1(1-p_1)^2 + (1-\lambda)p_2(1-p_2)^2}
\end{aligned}$$

$$\begin{aligned}
\Pr(y = T|x = THT, \theta) &= \frac{\Pr(x{=}THT, y{=}T|\theta)}{\Pr(x{=}THT|\theta)} \\
&= \frac{\Pr(x{=}THT, y{=}T|\theta)}{\Pr(x{=}THT, y{=}H|\theta) \cdot \Pr(x{=}THT, y{=}T|\theta)} \\
&= \frac{(1-\lambda)p_2(1-p_2)^2}{\lambda p_1(1-p_1)^2 + (1-\lambda)p_2(1-p_2)^2}
\end{aligned}$$

where the first equality follows from the definition of conditional probability, the second equality follows from the fact that

$$\Pr(x) = \sum_y \Pr(x, y),$$

and the third equality follows from the definition of the parameter values. This example shows how the Expectation step of the EM algorithm works: Given a θ, the probability distribution of the hidden data can be estimated from the observed data. To start the algorithm, an initial θ is needed, but once the algorithm is started, new values of θ can be calculated that increases their likelihood.

Three Coins - How to Get a New θ: Given the probability distribution of the hidden data and the previous estimate of θ, denoted θ^0, a new estimate θ^1 can be constructed that has a higher likelihood. In this case, since there are only two types of observed triples in this example, $x = HHH$ or $x = TTT$, the new value for λ is easily calculated by

$$\lambda^1 = \frac{3 \times \Pr(y = H|x = HHH, \theta^0) + 2 \times \Pr(y = H|x = TTT, \theta^0)}{5}$$

where the numerator is the expected number of heads from coin 0 in 5 trials and the denominator is the number of trials. Thus the new value of λ is calculated conditional on the old θ values. The new parameter estimates for coins 1 and 2 are similarly calculated by

$$p_1^1 = \frac{3 \times 3 \times \Pr(y = H|x = HHH, \theta^0) + 0 \times 2 \times \Pr(y = H|x = TTT, \theta^0)}{3 \times 3 \times \Pr(y = H|x = HHH, \theta^0) + 3 \times 2 \times \Pr(y = H|x = TTT, \theta^0)}$$

$$p_2^1 = \frac{3 \times 3 \times \Pr(y = T|x = HHH, \theta^0) + 0 \times 2 \times \Pr(y = T|x = TTT, \theta^0)}{3 \times 3 \times \Pr(y = T|x = HHH, \theta^0) + 3 \times 2 \times \Pr(y = T|x = TTT, \theta^0)}$$

These calculations are an example of the Maximization step, and with these new parameter estimates, the Estimations step can be performed again to yield better estimates of the probability distribution of the hidden data. An implementation of the EM algorithm for the three coins problem can be found in the directory:

$$/cs/fac1/anoop/cmpt825/demos/three\_coins.py$$

6.2 ML Estimation

The goal of maximum likelihood (ML) parameter estimation is to find the parameters of a model that maximize the probability (i.e. likelihood) of the sample data. In the EM algorithm, ML estimation occurs in the Maximization step. To set up the required notation, let $x_1, x_2, \cdots, x_n$ be the observed

data sequence drawn from a set $\mathcal{X}$, and let $y_1, y_2, \cdots, y_n$ be the corresponding sequence of unobserved data drawn from a set $\mathcal{Y}$. In the case of the Three Coins example, these sets are

$$\mathcal{X} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

$$\mathcal{Y} = \{H, T\}.$$

For computational tractability, the data are assumed to be independently and identically distributed. This assumption allows us to write

$$\Pr(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} \Pr(x_i|\theta).$$

As before, let θ be a parameter vector taken from a parameter space Ω. In the fully observed case, where the sequences of both x's and y's are observed, then maximizing

$$L(\theta) = \sum_{i=1}^{n} \log \Pr(x_i, y_i|\theta)$$

would be the goal. Note that taking logs allows the product from above to be written as a sum. In the case where the y's are hidden, then all possible y's must be considered. Thus, the problem is to maximize

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{n} \log \Pr(x_i|\theta) \\ &= \sum_{i=1}^{n} \log \sum_{y \in \mathcal{Y}} \Pr(x_i, y|\theta) \end{aligned}$$

So, we need to find

$$\theta_{ML} = \arg\max_{\theta \in \Omega} \sum_{i=1}^{n} \log \sum_{y \in \mathcal{Y}} \Pr(x_i, y|\theta).$$

If $L(\theta)$ is a convex function with a single θ such that

$$\frac{dL(\theta)}{d\theta} = 0,$$

then finding the global solution to the maximum likelihood problem is straightforward. However, in more complex likelihood functions, there may be many local maxima that are less than the global maximum, and so finding the true θ_{ML} may be tricky. There are some strategies for dealing with multiple local maxima such as running the algorithm with different starting values or step sizes.

6.3 Convergence of the EM Algorithm

While the EM algorithm might only find a parameter estimate that has a locally maximum likelihood, it does have the desirable property that the likelihood is non-decreasing with each step of the algorithm and it is guaranteed to converge to a local maximum of the likelihood function.

The EM algorithm produces a new estimate of θ with each iteration,

$$\theta^1, \theta^2, \theta^3, \cdots, \theta^t, \cdots$$

and for all t, $L(\theta^{t+1}) \geq L(\theta^t)$. To prove this result, the concept of *expected values* and a result known as *Jensen's inequality* are used:

Expected Values: The expected value of a discrete random variables, (also called the expectation or the mean) is the probability-weighted average value of the random variable. It is calculated by summing each possible realization of the random variable multiplied by the probability of that realization. For example, let x be a random variable that can take on values from the set $\{2, 3, 4\}$ and let the probability of taking each of these values is 0.25, 0.25, and 0.50, respectively. Then the expected value of x, denoted $E[x]$ is $\sum_{x \in \{2,3,4\}} x \times \Pr(x) = 3.25$.

Jensen's Inequality: This result states that the expected value of the function is not, in general, equal to the function of the expected value. So $E[f(x)] \neq f(E[x])$. In particular, if f is a convex function, then $E[f(x)] \geq f(E[x])$. Note that the fact that the log function is convex will be used in the following proof in conjunction with Jensen's Inequality.

The proof of EM convergence proceeds in three steps:

1. Show that the log-likelihood function, $L(\theta)$, has the form

$$L(\theta) = Q(\theta, \theta^t) - H(\theta, \theta^t)$$

for some functions Q and H.

2. Show that $Q(\theta^{t+1}, \theta^t) \geq Q(\theta^t, \theta^t)$.

3. Show that $H(\theta^t, \theta^t) \geq H(\theta^{t+1}, \theta^t)$.

Proof of Convergence - Step 1: As before, let x be the observed data, y be the unobserved data, and θ be the vector of parameter values that we are estimating. By the definition of conditional probabilities, we have the following equality:

$$\Pr(y|x, \theta) = \frac{\Pr(x, y|\theta)}{\Pr(x|\theta)}.$$

At any step, t, in the EM algorithm, we have an estimate of the parameter vector θ^t as well as a working estimate of the distribution of the unobserved data $\widetilde{p}(y|x, \theta^t)$. Taking expectations of both sides of the above equality at time t conditional on $\widetilde{p}(y|x, \theta^t)$ gives:

$$E_{\theta^t}[\log \Pr(y|x, \theta)|\widetilde{p}(y|x, \theta^t)] = E_{\theta^t}[\log \Pr(x, y|\theta)|\widetilde{p}(y|x, \theta^t)] - E_{\theta^t}[\log \Pr(x|\theta)|\widetilde{p}(y|x, \theta^t)].$$

The second term on the right-hand side is not dependant on y and is not a random variable, and so the expectation operator can be removed giving:

$$E_{\theta^t}[\log \Pr(y|x, \theta)|\widetilde{p}(y|x, \theta^t)] = E_{\theta^t}[\log \Pr(x, y|\theta)|\widetilde{p}(y|x, \theta^t)] - \log \Pr(x|\theta).$$

This second term is now expressed as the log-likelihood of a parameter vector θ as described in the previous section. Thus,

$$E_{\theta^t}[\log \Pr(y|x, \theta)|\widetilde{p}(y|x, \theta^t)] = E_{\theta^t}[\log \Pr(x, y|\theta)|\widetilde{p}(y|x, \theta^t)] - L(\theta).$$

Rearranging this equality gives

$$L(\theta) = E_{\theta^t}[\log \Pr(y|x, \theta)|\widetilde{p}(y|x, \theta^t)] - E_{\theta^t}[\log \Pr(x, y|\theta)|\widetilde{p}(y|x, \theta^t)]$$

which is of the desired form,

$$L(\theta) = Q(\theta, \theta^t) - H(\theta, \theta^t).$$

Proof of Convergence - Step 2: In the E-step of the EM algorithm, we compute an estimate $\widetilde{p}(y|x,\theta^t)$ which is then used in the M-step. In this step, we are actually calculating the arg max of $Q(\theta,\theta^t)$. That is,

$$\begin{aligned}\theta^{t+1} &= \arg\max_{\theta\in\Omega} Q(\theta,\theta^t) \\ &= \arg\max_{\theta\in\Omega} E_{\theta^t}[\log \Pr(x,y|\theta)|\widetilde{p}(y|x,\theta^t)].\end{aligned}$$

So by the definition of arg max, $Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t)$ as required.

Proof of Convergence - Step 3: The proof of this step uses the definition of expected values and conditional probability, the properties of logs, and Jensen's Inequality.

$$\begin{aligned}& H(\theta^t,\theta^t) - H(\theta^{t+1},\theta^t) \\ =\ & E_{\theta^t}[\log \Pr(y|x,\theta^t)|\widetilde{p}(y|x,\theta^t)] - E_{\theta^t}[\log \Pr(y|x,\theta^{t+1})|\widetilde{p}(y|x,\theta^t)] \\ =\ & \sum_y \log \Pr(y|x,\theta^t)\widetilde{p}(y|x,\theta^t) - \sum_y \log \Pr(y|x,\theta^{t+1})\widetilde{p}(y|x,\theta^t) \\ =\ & \sum_y \log \tfrac{\Pr(x,y|\theta^t)}{\Pr(x|\theta^t)}\widetilde{p}(y|x,\theta^t) - \sum_y \log \tfrac{\Pr(x,y|\theta^{t+1})}{\Pr(x|\theta^{t+1})}\widetilde{p}(y|x,\theta^t) \\ =\ & \sum_y \log \tfrac{\Pr(x,y|\theta^t)}{\Pr(x,y|\theta^{t+1})}\widetilde{p}(y|x,\theta^t) + \sum_y \log \tfrac{\Pr(x|\theta^{t+1})}{\Pr(x|\theta^t)}\widetilde{p}(y|x,\theta^t) \\ \geq\ & \log \tfrac{\Pr(x|\theta^t)}{\Pr(x|\theta^{t+1})} + \log \tfrac{\Pr(x|\theta^{t+1})}{\Pr(x|\theta^t)} \\ =\ & 0\end{aligned}$$

So, give these three steps, we see that the likelihood is non-decreasing, and the convergence of the EM algorithm is proven. ■

6.4 Baum-Welch as an Instance of the EM Algorithm

The Baum-Welch algorithm (also called the forward-backward algorithm) is a special case of the EM algorithm. As we saw earlier in the course, the Baum-

Welch algorithm is used to find the unknown parameters of a hidden Markov model (HMM). To see Baum-Welch as an instance of EM, we need to identify the hidden data and state what happens in the E and M-steps. The hidden data are the unobserved state transitions. So in the E-step, given a parameter estimate, we compute the expected values of (i) the number of transitions from each state i in the observed data, and (ii), for each state pair (i, j), the number of transitions from state i to state j. The parameters of the model are (1) the initial state probabilities, (2) the state transition probabilities, and (3) the symbol emission probabilities. The M-step computes a new estimate of these parameters given the expected values (i) and (ii).

6.5 Variants of the EM Algorithm

Dempster, Laird, and Rubin (DLR) (1977) first presented the EM algorithm in its general form. Wu (1983) extended the results of DLR, including a proof that $L(\theta^{t+1}) > L(\theta^t)$ provided that θ^t is not a stationary point (i.e. $\frac{dL(\theta^t}{d\theta} \neq 0$). Wu's result is a more useful result than DLR's since it says that EM is gautenteed to strictly improve the likelihood as long as you're not at a local maximum.

DLR also presented a Generalized EM algorithm (GEM), which is like the EM algorithm except that one picks θ^{t+1} such that $Q(\theta^{t+1}, \theta^t) \geq Q(\theta^t, \theta^t)$. Note that we need only choose a θ^{t+1} that improves Q rather than one that maximizes it. This variant of EM helps when finding the true maximum is very expensive. The convergence results still apply to GEM.

Neal and Hinton (1998) give an alternative proof of convergence and allow other generalizations. They define a function $F(\widetilde{p}, \theta)$ which enables a new definition of EM. They define the maximum likelihood estimate of θ as

$$\theta_{ML} = \arg\max_{\theta}[\max_{\widetilde{p}} F(\widetilde{p}, \theta)].$$

This implies that the likelihood function is defined to be

$$L(\theta) = \max_{\widetilde{p}} F(\widetilde{p}, \theta).$$

We can find the probability of the hidden data in the following manner:

$$p(y|x, \theta) = \arg\max_{\widetilde{p}} F(\widetilde{p}, \theta).$$

So the new definition of EM is:

E-Step: $\widetilde{p}^t = \arg\max_{\widetilde{p}} F(\widetilde{p}, \theta^{t-1})$

M-Step: $\theta^t = \arg\max_{\theta} F(\widetilde{p}^t, \theta)$

This is a rephrasing of EM in which we're maximizing one variable at a time. The convergence properties of EM still apply and the proof is more compact than earlier versions (but it's also harder to understand). The useful aspect of Neal and Hinton's F function is that it can be modified to prove that many variants EM also converge. For example, they can show that on-line processing, various sampling procedures, and learning only some parameters in each iteration are all variants of EM for which the convergence results also apply.

6.6 References

Collins, Michael (1997) "The EM Algorithm." manuscript.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society B, vol. 39, pp. 1-38.

Neal, Radford and Hinton, Geoffrey (1998). "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants." In M. I. Jordan (editor) Learning in Graphical Models, pp. 355-368, Dordrecht: Kluwer Academic Publishers. 1998.

Wu, C. F. J. (1983). "On the Convergence Properties of the EM Algorithm." The Annals of Statistics, Vol. 11, No. 1, pp. 95-103.