A Gentle Introduction to the EM Algorithm

Ted Pedersen
Department of Computer Science
University of Minnesota Duluth
tpederse@d.umn.edu

A unifying methodology

- Dempster, Laird & Rubin (1977) unified many strands of apparently unrelated work under the banner of **The EM Algorithm**
- EM had gone incognito for many years
  - Newcomb (1887)
  - McKendrick (1926)
  - Hartley (1958)
  - Baum et. al. (1970)

A general framework for solving many kinds of problems

- Filling in missing data in a sample
- Discovering the value of latent variables
- Estimating parameters of HMMs
- Estimating parameters of finite mixtures
- Unsupervised learning of clusters
- …

EM allows us to make MLE under adverse circumstances

- What are Maximum Likelihood Estimates?
- What are these adverse circumstances?
- How does EM triumph over adversity?
- PANEL: When does it really work?

Maximum Likelihood Estimates

- Parameters describe the characteristics of a population. Their values are estimated from samples collected from that population.
- A MLE is a parameter estimate that is most consistent with the sampled data. It maximizes the likelihood function.

Coin Tossing!

- How likely am I to toss a head? A series of 10 trials/tosses yields (h,t,t,t,h,t,t,h,t,t)
  - (x1=3, x2=7), n=10
- Probability of tossing a head = 3/10
- That's a MLE! This estimate is absolutely consistent with the observed data.
- A few underlying details are masked…

Coin tossing unmasked

- Coin tossing is well described by the binomial distribution since there are ***n*** independent trials with two outcomes.
- Given 10 tosses, how likely is 3 heads?

$$L(\theta) = \binom{10}{3} \theta^3 (1-\theta)^7$$

Maximum Likelihood Estimates

- We seek to estimate the parameter such that it maximizes the likelihood function.
- Take the first derivative of the likelihood function with respect to the parameter theta and solve for 0. This value maximizes the likelihood function and is the MLE.

Maximizing the likelihood

$$L(\theta) = \binom{10}{3} \theta^3 (1-\theta)^7$$

$$\log L(\theta) = \log \binom{10}{3} + 3 \log \theta + 7 \log(1-\theta)$$

$$\frac{d \log L(\theta)}{d\theta} = \frac{3}{\theta} - \frac{7}{1-\theta} = 0$$

$$\frac{3}{\theta} = \frac{7}{1-\theta} \quad \Rightarrow \quad \theta = \frac{3}{10}$$

Multinomial MLE example

- There are ***n*** animals classified into one of four possible categories (Rao 1973).
  - Category counts are the sufficient statistics to estimate multinomial parameters
- Technique for finding MLEs is the same
  - Take derivative of likelihood function
  - Solve for zero

Multinomial MLE example

There are $n = 197$ animals classified into one of 4 categories:
$Y = (y1, y2, y3, y4) = (125, 18, 20, 34)$

The probability associated with each categoryis given as:
$\Theta = (\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$

The resulting likelihood function for this multinomial is:
$L(\pi) = \frac{n!}{y1!\,y2!\,y3!\,y4!} * (\frac{1}{2} + \frac{1}{4}\pi)^{y1} * (\frac{1}{4}(1-\pi))^{y2} * (\frac{1}{4}(1-\pi))^{y3} * (\frac{1}{4}\pi)^{y4}$

Multinomial MLE example

$$\log L(\pi) = y1 * \log(\frac{1}{2} + \frac{1}{4}\pi) + y2 * \log(\frac{1}{4}(1-\pi)) + y3 * \log(\frac{1}{4}(1-\pi)) + y4 * \log(\frac{1}{4}\pi)$$

$$\frac{d \log L(\pi)}{d\pi} = \frac{y1}{2+\pi} - \frac{y2+y3}{1-\pi} + \frac{y4}{\pi} = 0$$

$$\frac{d \log L(\pi)}{d\pi} = \frac{125}{2+\pi} - \frac{38}{1-\pi} + \frac{34}{\pi} = 0 \quad \Rightarrow \quad \pi = 0.627$$

Multinomial MLE runs aground?

- Adversity strikes! The observed data is incomplete. There are really 5 categories.
- y1 is the composite of 2 categories (x1+x2)
  - p(y1)= ½ + ¼ *pi, p(x1) = ½, p(x2)= ¼* pi
- How can we make a MLE, since we can't observe category counts x1 and x2?!
  - Unobserved sufficient statistics!?

EM triumphs over adversity!

- E-STEP: Find the expected values of the sufficient statistics for the complete data X, given the incomplete data Y and the current parameter estimates
- M-STEP: Use those sufficient statistics to make a MLE as usual!

MLE for complete data

$X = (x1, x2, x3, x4, x5) = (x1, x2, 18, 20, 34)$ where x1+x2=125

$\Theta = (\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$

$L(\pi) = \frac{n!}{x1!x2!x3!x4!x5!} * (\frac{1}{2})^{x1} * (\frac{1}{4}\pi)^{x2} * (\frac{1}{4}(1-\pi))^{x3} * (\frac{1}{4}(1-\pi))^{x4} * (\frac{1}{4}\pi)^{x5}$

MLE for complete data

$\log L(\pi) = x2 * \log(\frac{1}{4}\pi) * x3 * \log(\frac{1}{4}(1-\pi)) * x4 * \log(\frac{1}{4}(1-\pi)) * x5 * \log(\frac{1}{4}\pi)$

$\frac{d \log L(\pi)}{d\pi} = \frac{x2+x5}{\pi} - \frac{x3+x4}{1-\pi} = 0$

$\frac{d \log L(\pi)}{d\pi} = \frac{\mathbf{x2}+34}{\pi} - \frac{38}{1-\pi} = 0$

E-step

- What are the sufficient statistics?
  - X1 => X2 = 125 – x1
- How can their expected value be computed?
  - E [x1 | y1] = n*p(x1)
- The unobserved counts x1 and x2 are the categories of a binomial distribution with a sample size of 125.
  - p(x1) + p(x2) = p(y1) = ½ + ¼*pi

E-Step

- E[x1|y1] = n*p(x1)
  - p(x1) = ½ / (½+ ¼*pi)
- E[x2|y1] = n*p(x2) = 125 – E[x1|y1]
  - p(x2)= ¼*pi / (½ + ¼*pi)
- Iteration 1? Start with pi = 0.5 (this is just a random guess…)

E-Step Iteration 1

- E[x1|y1] = 125* (½ / (½+ ¼*0.5)) = 100
- E[x2|y1] = 125 – 100 = 25
- These are the expected values of the sufficient statistics, given the observed data and current parameter estimate (which was just a guess)

M-Step iteration 1

- Given sufficient statistics, make MLEs as usual

$$\frac{d \log L(\pi)}{d\pi} = \frac{\mathbf{x2} + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\frac{\mathbf{25} + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\pi = .608$$

E-Step Iteration 2

- E[x1|y1] = 125* (½ / (½+ ¼*0.608)) = 95.86
- E[x2|y1] = 125 – 95.86 = 29.14
- These are the expected values of the sufficient statistics, given the observed data and current parameter estimate (from iteration 1)

M-Step iteration 2

- Given sufficient statistics, make MLEs as usual

$$\frac{d \log L(\pi)}{d\pi} = \frac{\mathbf{x2} + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\frac{\mathbf{29.14} + 34}{\pi} - \frac{38}{1-\pi} = 0$$

$$\pi = .624$$

Result?

- Converge in 4 iterations to pi=.627
  - E[x1|y1] = 95.2
  - E[x2|y1] = 29.8

Conclusion

- Distribution must be appropriate to problem
- Sufficient statistics should be identifiable and have computable expected values
- Maximization operation should be possible
- Initialization should be good or lucky to avoid saddle points and local maxima
- Then…it might be safe to proceed…