MSF100
Statistical Inference Principles - Lecture 11

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

February 29, 2012

1 EM algorithm

- When direct maximization of the likelihood is difficult
- Look for *augmentation*/additional data, to make the problem easy.

Example:

- observed data $y_i, i = 1, \cdots, n$ from a mixture of densities $f_k, k = 1, \cdots, K$, $f_k(y|\theta_k)$
- Estimating θ_k and the mixture proportions is difficult, $f(\tilde{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_y(y|\theta_k)$
- However, if I knew which y_i belonged to which mixture component f_k, the problem is easy:

$$\hat{\theta}_k = argmax_\theta \prod_{i \in k} f_k(y_i|\theta_k)$$

$$\hat{\pi}_k = \sum_i 1\{i \in k\}$$

- augmentation = group or mixture memberships

Augmentation doesn't have to be of the form of additional data, though it frequently is. The augmentation has to be of the form

$$\Theta \rightarrow X \rightarrow Y$$

where X is the augmented data and Y is the observed. That is, we have a Markov dependency such that $p(y|x, \theta) = p(y|x)$. Common forms of X are $X : Y = T(X)$ (so the observed is some summary or truncation of X) or $X = (Z, Y)$ where Z is called the hidden or latent component of the data.

- We want to maximize $l(\theta|\tilde{y}) = \prod_{i=1}^{n} p(y_i|\theta)$, HARD.
- Complete data $X = (Z, y)$
- Idea: a two-step process. "Guess" the Z and use it as if it was true to maximize the likelihood of the complete data.

1.1 The algorithm

1. Starting value θ^0
2. E-step Compute the *expectation* of the complete likelihood *conditioned* on observed data $\tilde{y}$ and a current estimate θ^m
3. M-step *Maximize* this expectation wrt θ
4. Iterate until convergence.

Example:

- y_i from f_1 and f_2 densities (special case with 2 mixtures).
- $Z_i = 1$ if y_i belongs to component f_1 and 0 otherwise (augmentation).
- Observed likelihood $\prod_i \sum_{k=1}^{2} \pi_k f_k(y_i|\theta_k)$
- Complete likelihood

$$L(X) = f(\tilde{y}, \tilde{Z}) = \prod_i \prod_{k=1}^{K} (\pi_k f_k(y_i|\theta_k))^1\{Z_i = k\}$$

- Note, now the likelihood is a more simple form, just products.
-

$$l(X) = \sum_i 1\{Z_i = 1\}(\log(\pi_1) + \log f_1(y_i|\theta_1)) + (1 - 1\{Z_i = 1\})(\log(1 - \pi_1) + \log f_2(y_i|\theta_2))$$

- Compute the expectation

$$E[l(X)|\tilde{y}, \pi^m, \theta^m] =$$

$$= \sum_i P(Z_i = 1|\tilde{y}, \pi^m, \theta^m)(\log(\pi_1)+\log f_1(y_i|\theta_1))+(1-P(Z_i = 1|\tilde{y}, \pi^m, \theta^m))(\log(1-\pi_1)+\log f_2(y_i|\theta_2))$$

  using previous estimate $\pi^m = (\pi_1)^m$ and $\theta_m = (\theta_1^m, \theta_2^m)$

- Now,

$$P(Z_i = 1|\tilde{y}, \pi^m, \theta^m) = \frac{\pi_1^m f_1(y_i|\theta_1^m)}{\pi_1^m f_1(y_i|\theta_1^m) + (1 - \pi_1^m) f_2(y_i|\theta_2^m)} = \pi_1^{m+1}(i)$$

  This quantity: $\pi_1^{m+1}(i)$ is the posterior probability that observation i belongs to the component f_1, i.e. it is an observation level mixture probability as opposed to the π_1 which is the population parameter, the *proportion* of samples y_i that came from the component f_1.

- Given the expression for $\pi_1^{m+1}(i)$ we now maximize the expected value $* = E[l(X)|\tilde{y}, \pi^m, \theta^m]$ above wrt to π_1 and θ_k.

- First, taking the derivative wrt π_1 we obtain

$$\frac{d*}{d\pi} = \sum_i \pi_1^{m+1}(i)\pi^{-1} + \sum_i -(1 - \pi_1^{m+1})(1 - \pi)^{-1} = 0$$

  and solving for π_1:

$$\pi_1^{m+1} = \sum_{i=1}^{n} \pi_1^{m+1}(i)/n = \text{"}n_1\text{"}/n$$

  where we use n_1 to denote the total posterior probability mass in component 1 as representing an estimate of the number of samples from this component.

- Similarly, we obtain estimate for θ_k as

$$\theta_k^{m+1} = argmax_{\theta_k} \sum_{i=1}^{n} \pi_k^{m+1}(i) f_k(y_i|\theta_k)$$

- Notice that *all* observations contribute to both components weighted by their posterior probability mass.

- An alternative called the CEM (classification EM) uses a threshold on $\pi_k^{m+1}(i)$ such that observations contribute only to the component that they have maximum probability wrt to. This is tantamount to replacing the E-step with a MaP step

1.2 General formulation

We want the MLE $\theta^* = argmax_\theta L(\theta|\tilde{y})$ but this is hard to obtain. We augment the data $X = (Z, y)$ s.t. $\theta \to X \to y$ holds (Markov) and thus $p(x|\theta) = p(y|z,\theta)p(z|\theta)$ or $p(y|x,\theta) = p(y|x)$.

We start with value θ^0 and compute the E-step:

- E-step Compute
$$Q(\theta, \theta^m) = E[\log \frac{f(y, Z|\theta)}{f(y, Z|\theta^m)}|y, \theta^m]$$
(note, use of the relative loglikelihood in the expectation)
- M-step View $Q(\theta, \theta^m)$ as a function of θ with θ^m fixed and maximize to obtain
$$\theta^{m+1} = argmax_\theta Q(\theta, \theta^m)$$

The point is that if we choose Z carefully, the Mstep has closed form/is easy to compute.

Why does this work? Well, we want to maximize $l(\theta|y)$. We can write

$$l(\theta|y) = \log p(y|\theta) = \log \int_{X(y)} p(x, y|\theta)dx = (**)$$

where $X(y)$ refers to x-values for which the joint density $p(x, y|\theta)$ is defined.
Now,
$$(**) = \log \int_{X(y)} \frac{p(x, y|\theta)}{p(x|y, \theta^m)} p(x|y, \theta^m)dx = \log E_{X|y,\theta^m}[\frac{p(X, y|\theta)}{p(X|y, \theta^m)}] = (***)$$

By Jensen's inequality we have

$$(***) \geq E_{X|y,\theta^m}[\log \frac{p(X, y|\theta)}{p(X|y, \theta^m)}] = E_{X|y,\theta^m}[\log \frac{p(X|\theta)p(y|\theta)}{p(X|\theta^m)p(y|X)/p(y|\theta^m)}] = (o)$$

where the last expression follows from the Markov property of X and y and Bayes theorem. This simplifies to

$$(o) = E_{X|y,\theta^m}[\log \frac{p(X|\theta)p(y|\theta^m)}{p(X|\theta^m)}] = E_{X|y,\theta^m}[\log p(X|\theta)] - E_{X|y,\theta^m}[\log p(X|\theta^m)] + \log p(y|\theta^m) =$$

$$= Q(\theta|\theta^m) - Q(\theta^m|\theta^m) + l(\theta^m)$$

Now, since we maximize the Q function in each iteration of the EM algorithm, we have $Q(\theta|\theta^m) - Q(\theta^m|\theta^m) \geq 0$ and so

$$l(\theta) \geq l(\theta^m)$$

That is, the EM steps are guaranteed to not decrease the likelihood at least. Another way to interpret the above result and what the EM does is to think of the Q function as a function of the iteration step m as better and better lower bounds of the loglikelihood that we want to estimate.

1.3 Convergence

We have that the EM always generates a nondecreasing sequence of loglikelihood values. Using the "imputed" likelihood Q has the following property:
Given $\theta \to X \to y$, if $Q(\theta|\theta^m) \geq Q(\theta^m|\theta^m)$ then $l(\theta) \geq l(\theta^m)$. Now, no general convergence result exists for EM since the behaviour of the sequence $\{\theta^m\}$ depends on both l, Q and the starting value. However, there are some results that show, under certain regularity conditions, that EM converges to a stationary point of $l(\theta)$. It is therefore important to try different starting values. Convergence can also be rather slow, slower than e.g. Newton-Rhapson, but the benefit is that the M-step usually is a very simple, closed form. If you are curious about convergence results for EM I recommend the paper by Jeff Wu ('83) though it heavy on optimization.

1.4 Practical considerations

Besides the convergence issues and multiple starting points you need also to worry about nonrobustness in the M-step. This frequently happens in e.g. mixture modeling where essentially one or more of the components have very little probability mass and the corresponding mixture parameters are poorly estimated. One can incorporate priors into EM to deal with this problem. The prior enters in the Mstep as

$$argmax_\theta Q(\theta|\theta^m) + \log \pi(\theta)$$

In multivariate gaussian mixture modeling, we commonly use a prior that "shrinks" the covariance estimate Σ_k toward a global value Δ. That is,

$$\pi_k^{m+1} = \sum_{i=1}^{n} \pi_k^m(i)/n, \quad n_k = \sum_{i=1}^{n} \pi_k^m(i)$$

$$\mu_k^{m+1} = \sum_{i=1}^{n} \pi_k^m(i) x_i / n_k$$

$$\Sigma_k^{m+1} = \frac{1}{n_k} \sum_{i=1}^{n} \pi_k^{m+1}(i)(x_i - \mu_k^{m+1})(x_i - \mu_k^{m+1}) + \nu\Delta$$

The ν is a tuning parameter (controlled by the shape of the prior - how focused it is around Δ.). A common choice for Δ is a scaled version of the covariance of y - i.e. estimated covariance from the data ignoring that mixture components are present. If you are interested in mixture modeling I recommend you check out the book by G. J. McLachlan. For papers on mixture modeling and Bayesian estimates, check out e.g. Adrian Raftery's homepage (U. Washington).