

Stat 992: Lecture 32

Gaussian mixture model I.

Moo K. Chung mchung@stat.wisc.edu

December 14, 2003

1. *Gaussian mixture model.* Download data and run the following MATLAB code.

```
load sagittal.data;
intensity=reshape(sagittal,256*124,1)
intensity=intensity(find(
(intensity >=50)&(intensity<=250)));
hist(intensity,100);
```

We will get 19,660 values for intensity which are gray scale values between 50 and 250 of the magnetic resonance image (MRI) of the human brain. Let $f(y_i)$ be the image intensity at pixel position y_i . From the intensity histogram, we see that the image intensity can be modelled as a two component Gaussian mixture

$$f(y) = pf_1(y) + (1 - p)f_2(y)$$

where $f_1 \sim N(\mu_1, \sigma_1^2)$ and $f_2 \sim N(\mu_2, \sigma_2^2)$. We will estimate parameters by maximizing the likelihood function using the EM algorithm.

2. *Expectation-maximization (EM)* method is an iterative method for maximizing difficult likelihood problems. It was first introduced by Dempster *et al.* (J. Roy. Statist. Soc. 1977). Read Robert and Casella's Monte Carlo Statistical Methods for the introduction to EM. Flury's A First Course in Multivariate Statistics for the detailed discussion on EM applied to Gaussian mixture model. See Little and Rubin (1987) and McLachlan and Krishnan (1997). EM is widely used in imaging. Instead of classifying pixels into classes, it is possible to assign the probability belong to classes (SPM'99 segmentation algorithm).

Suppose we have a random sample $X = \{X_1, \dots, X_n\} \sim f(x|\theta)$. We wish to find the maximum likelihood estimator of θ :

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n f(x_i|\theta) = \arg \max_{\theta} \sum_{i=1}^n \ln f(x_i|\theta).$$

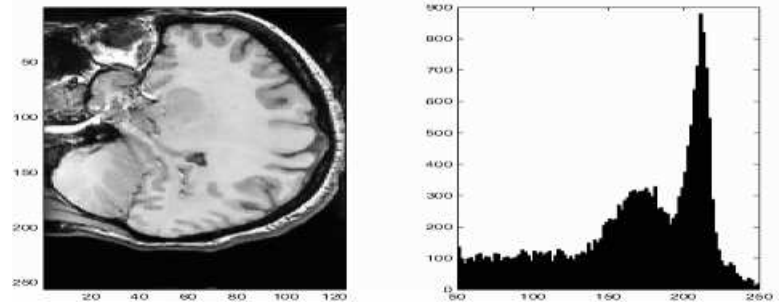


Figure 1: Left: mid sagittal brain image. Right: histogram of image intensity.

This optimization problem is nontrivial when f is complex and it gives a motivation for the development of the EM algorithm. We augment the data with with *latent* (unobserved) data X^m such that the complete data $X^c = (X, X^m)$. The density of X^c is denoted as

$$X^c = (X, X^m) \sim f(x^c) = f(x, x^m).$$

The conditional density for the missing data X^m is

$$f(x^m|x, \theta) = \frac{f(x, x^m|\theta)}{f(x|\theta)}.$$

Rearranging terms, $f(x|\theta) = \frac{f(x, x^m|\theta)}{f(x^m|x, \theta)}$. Taking logarithm, we get log-likelihood for the observed data

$$\ln f(X|\theta) = \ln f(X^c|\theta) - \ln f(X^m|X, \theta).$$

Take expectation with respect to $f(x^m|x, \theta_0)$ on both sides so that the left side is taken as constant.

$$\begin{aligned} \ln f(X|\theta) &= \mathbb{E}[\ln f(X^c|\theta)|X, \theta_0] \\ &\quad - \mathbb{E}[\ln f(X^m|X, \theta)|X, \theta_0]. \end{aligned}$$

3. Following Robert and Casella, denote the expected log-likelihood for the complete data by

$$Q(\theta|\theta_0, X) = \mathbb{E}[\ln f(X^c|\theta)|X, \theta_0].$$

We maximize the likelihood in the following fashion.

1. E-step: compute $Q(\theta|\hat{\theta}_{j-1}, X)$.
2. M-step: maximize $Q(\theta|\hat{\theta}_{j-1}, X)$ and take

$$\hat{\theta}_j = \arg \max_{\theta} Q(\theta|\hat{\theta}_{j-1}, X).$$

If the above procedure is iterated, we get the sequence of estimators $\hat{\theta}_0, \hat{\theta}_1, \dots$ and it can be shown that it converges to the maximum likelihood estimator $\hat{\theta}$.

Proof. By the definition of $\hat{\theta}_{j+1}$,

$$Q(\hat{\theta}_{j+1}|\hat{\theta}_j, x) \geq Q(\hat{\theta}_j|\hat{\theta}_j, x).$$

Let $R(\theta|\theta_0, X) = \mathbb{E}[\ln f(X^m|X, \theta)|X, \theta_0]$. It can be shown that

$$\theta_0 = \arg \max_{\theta} R(\theta|\theta_0, x) \quad (1)$$

Then $R(\hat{\theta}_{j+1}|\hat{\theta}_j, x) \leq R(\hat{\theta}_j|\hat{\theta}_j, x)$. Consequently

$$\begin{aligned} \ln f(x|\hat{\theta}_j) &= Q(\hat{\theta}_j|\hat{\theta}_j, x) - R(\hat{\theta}_j|\hat{\theta}_j, x) \\ &\leq Q(\hat{\theta}_{j+1}|\hat{\theta}_j, x) - R(\hat{\theta}_{j+1}|\hat{\theta}_j, x) \\ &\leq \ln f(x|\hat{\theta}_{j+1}). \end{aligned}$$

This inequality guarantees the the sequence of estimators $\hat{\theta}_j$ monotonically increase the likelihood. Since the monotonically increasing sequence is bounded, i.e. $\ln f(x|\hat{\theta}_j) \leq \ln f(x|\hat{\theta})$, it must be converging to a stationary point but it is not clear if the limit is $\ln f(x|\hat{\theta})$. To guarantee that the limit converges to the maximum likelihood estimator, additional conditions are needed. See Boyles (J. Roy. Statist. Soc. B. 1983) and Wu (Ann. Statist. 1983)

The difficulty of EM algorithm is at the E-step where we need to compute the conditional expectation $Q(\theta|\hat{\theta}_{j-1}, x)$. The *Monte-Carlo EM* algorithm overcome this by simulating missing data $X^m \sim f(x^m|x, \theta)$. so that

$$\hat{Q}(\theta|\theta_0, x) = \frac{1}{l} \sum_{j=1}^l \ln f(X, X^m|\theta).$$

Problem 46. Prove inequality (1). Hint. Use Jensen's inequality.

4. *Multinomial example.* $X = (X_1, \dots, X_n) \sim M(n, p_1, p_2, \dots, p_{k-1}, p_k)$ is multinomial if

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k},$$

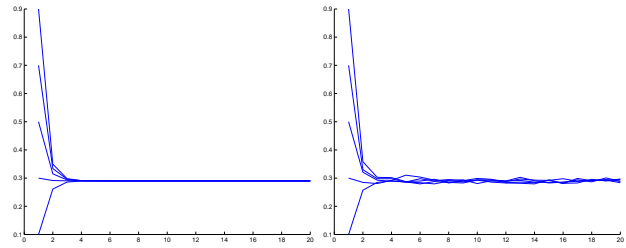


Figure 2: Left: convergence of EM algorithm. Right: Monte-Carlo EM algorithm.

where $\sum_{i=1}^k x_i = n$ and $\sum_{i=1}^k p_i = 1$. The following example is motivated from Rao C.R. (Linear statistical inference and its applications, 1973) and simplified from $k = 4$ to $k = 3$. Suppose observations $x = (x_1, x_2, x_3)$ are from the multinomial $M(n, \frac{1}{2} + \frac{p}{4}, \frac{1-p}{2}, \frac{p}{4})$. Find the MLE of p .

Solution. The observed likelihood is

$$L(p|x) \propto (2+p)^{x_1} (1-p)^{x_2} p^{x_3}.$$

MLE can be computed if one is willing to do algebraic manipulation. Instead we augment data by additional data such that

$$x^c = (z_1, z_2, x_2, x_3) \sim M(n, \frac{1}{2}, \frac{p}{4}, \frac{1-p}{2}, \frac{p}{4})$$

with $x_1 = z_1 + z_2$. The complete likelihood function is $L(p|x^c) \propto p^{z_2+x_3} (1-p)^{x_2}$. This is easier to manipulate than $L(p|x)$. The conditional expected complete log-likelihood $Q(p|p_0, x)$ is

$$\mathbb{E}[\log L(p|X^c)|p_0, x] \propto \mathbb{E}_{p_0}[(Z_2+x_3) \log p + x_2 \log(1-p)]$$

We can also show that $Z_2|X_1 = x_1 \sim B(x_1, \frac{p}{2+p})$ (check it yourself). So the E-step is

$$Q(p|p_0, x) = \left(\frac{p_0}{2+p_0} x_1 + x_3 \right) \log p + x_2 \log(1-p).$$

The M-step is also easy. Computing $\partial Q/\partial p = 0$, we get

$$\hat{p}_1 = \frac{\frac{p_0}{2+p_0} x_1 + x_3}{\frac{p_0}{2+p_0} x_1 + x_2 + x_3}.$$

This gives a recursive formula to find \hat{p}_i .

```
clear p; x1=35;x2=50;x3=16;
p(1)=0.9;
for i=2:20
    pr= p(i-1)/(2+p(i-1))
    p(i) = (pr*x1 + x3)/(pr*x1+x2+x3);
end; hold on; plot(p)
```