Introduction to Computational Biology  
Lecture # 19: EM clustering algorithm  
Ruty Rinott  
14/07/2008

1 Review of EM clustering

Let \( X_{ga} \) represent the measurement of gene \( g \) in array \( a \). For each gene there is a cluster variable \( C_g \) that denotes the cluster of gene \( g \). We define \( X = \{ X_{ga} : g \in G, a \in A \} \) as the observations vector, and \( C = \{ C_g : g \in G \} \) as the vector of cluster variables. Given the following parameters:

- \( \theta^k \in \mathbb{R}^k \) a multinomial distribution specifying the prior of the probability of a gene to belong to each cluster.
- \( \theta^{x}_{ca} = (\mu_{ca}, \sigma^2_{ca}) \) a gaussian distribution of samples from array \( a \) for genes in cluster \( c \).

We define

\[
P(X, \overline{C} : \theta) = \prod_g \left[ P(C_g : \theta^k) \prod_a P(X_{ga} : \theta^{x}_{ca}) \right]
\]

and the likelihood of the observations is

\[
\ell(\theta : \overline{X}) = \sum_c P(X, \overline{C} : \theta)
\]

We would like to find the parameters \( \theta \) that will maximize this likelihood. Since we have hidden variables, \( \overline{C} \), which control the cluster of each gene, EM can be a good strategy.

The EM algorithm is:

- guess an initial \( \overline{\theta} \)
- Loop
  - M-step: compute \( \theta = \arg \max_{\theta} \ell(\theta : ESS) \).

1.1 Defining Sufficient Statistics

We will start by defining 3 sufficient statistics, and then show that they are indeed sufficient.

\[
N_c = \sum_g 1(C_g = c) \text{ denotes the number of genes in cluster } C
\]

\[
M^{1}_{ca} = \sum_g 1(C_g = c) \cdot x_{ga}
\]

\[
M^{2}_{ca} = \sum_g 1(C_g = c) \cdot x^2_{ga}
\]
Using these sufficient statistics and changing the order so that we go over the genes by clusters the probability function can be written as:

\[
P(\mathbf{X}, \mathcal{C} : \theta) = \prod_c N_c \cdot \theta_c \prod_g \prod_a \frac{1}{\sqrt{2\pi\sigma_{ca}}} e^{-\frac{(x_{ga} - \mu_{ca})^2}{2\sigma_{ca}^2}}
\]

\[
= \prod_c N_c \cdot \theta_c \prod_a \frac{1}{\sqrt{2\pi\sigma_{ca}}} e^{-\frac{\sum_g (x_{ga} - \mu_{ca})^2}{2\sigma_{ca}^2}}
\]

\[
= \prod_c N_c \cdot \theta_c \prod_a \frac{1}{\sqrt{2\pi\sigma_{ca}}} e^{-\left(\frac{\sum_g x_{ga}^2 - 2\sum_g x_{ga} \cdot \mu_{ca} + \sum_g \mu_{ca}^2}{\sigma_{ca}^2}\right)}
\]

\[
= \prod_c N_c \cdot \theta_c \prod_a \frac{1}{\sqrt{2\pi\sigma_{ca}}} e^{-\left(\frac{M^2 - 2n \cdot \mu^2 + N \cdot n^2}{\sigma^2}\right)}
\]

(6)

1.2 E step

Notice that in EM we do not guess an assignment for the hidden variables, and therefore we can’t calculate the sufficient statistics directly. Instead we use the expectation of the sufficient statistics.

\[
E[N_c \mid \mathbf{X}, \theta] = \sum_g E[1(C_g = c) \mid \mathbf{X}, \theta] = \sum g E[1(C_g = c) \mid \mathbf{X}, \theta]
\]

by definition

\[
E[\sum_g 1(C_g = c) \mid \mathbf{X}, \theta] = \sum g E[1(C_g = c) \mid \mathbf{X}, \theta]
\]

linearly of \( E \)

\[
E \text{ of ind function}
\]

\[
= \sum_g P((C_g = c) \mid \mathbf{X}, \theta) = \sum g P((C_g = c) \mid \mathbf{X}, \theta)
\]

assuming genes are i.d.

\[
E [\mathbf{X}_g] = \frac{P(\mathbf{X}_g \mid C_g = c, \theta) \cdot P(C_g = c \mid \theta)}{P(\mathbf{X}_g \mid \theta)}
\]

bayes law

\[
E[\mathbf{X}_g] = \frac{\prod_a P(X_{ga} \mid C_g = c, \theta) \cdot P(C_g = c \mid \theta)}{\sum_c \prod_a P(X_{ga} \mid C_g = c, \theta)}
\]

(7)

Where \( \mathbf{X}_g \) denotes the vector of observations of gene \( g \).

Given \( \theta \) we know how to calculate this expression since \( P(X_{ga} \mid C_g = c) \) is distributed normally with known parameters and \( P(C_g = c) = \theta_c \). Calculating the expectation of the other two sufficient statistics is very similar, and we will just show the first steps.

\[
E[M^1_{ca} \mid \mathbf{X}, \theta] = \sum_g E[1(C_g = c) \cdot X_{ga} \mid \mathbf{X}, \theta] = \sum_g X_{ga} \cdot E[1(C_g = c) \mid \mathbf{X}, \theta]
\]

of constant \( g \)

\[
= \sum_g X_{ga} \cdot \frac{\prod_a P(X_{ga} \mid C_g = c, \theta) \cdot P(C_g = c \mid \theta)}{\sum_c \prod_a P(X_{ga} \mid C_g = c, \theta)}
\]

(8)

1.3 M step

In the M-step we are looking for \( \hat{\theta} = \arg \max_\theta \ell(\theta : \text{ESS}) \). \( \theta \) is composed of a gaussian distribution and a multinomial distribution. As we have seen many times in the past the MLE for gaussian distribution are:

\[
\hat{\mu} = \mathbf{X}_n
\]

(9)
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \frac{2 \bar{X}_n \sum_{i=1}^{n} X_i}{n} + \sum_{i=1}^{n} \bar{X}_n^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}_n \] 

(10)

The MLE for a multinomial distribution is

\[ \hat{\theta}_c = \frac{\sum_g 1(C_g = c)}{\text{number of genes}} \] 

(11)

Using the sufficient statistics we get a closed formula for the maximization of \( \theta \). In the EM we will use of course the expectation of these statistics.

\[ \hat{\mu}_{ca} = \frac{M_{ca}^1}{N_c} \] 

(12)

\[ \hat{\sigma}_{ca}^2 = \frac{M_{ca}^2}{N_c} - \hat{\mu}_{ca}^2 \] 

(13)

\[ \hat{\theta}_c = \frac{N_c}{\text{number of genes}} \] 

(14)