U. Grenoble-Alpes Master MSIAM **2016 - 2017**

Ensimag-Grenoble INP

Computational Biology – 2

Introduction to the EM algorithm

Change point detection and model-based clustering

Contact e.mail: olivier.francois@grenoble-inp.fr

Background: The EM algorithm

The EM algorithm is a method for estimating parameters in models with unobserved variables. Classical examples of applications are found in model-based clustering and in sequence analysis. EM stands for *Expectation-Maximization*, and it describes an iterative method that maximizes an *expected value* at each iteration.

Problem statement. Assume that we observe data, y, from a probability distribution which is defined in hierarchical way, as follows

$$p(y|\theta) = \int p(y|z,\theta)p(z|\theta)dz = \mathrm{E}[p(y|z,\theta)|\theta]\,.$$

In this formula, θ is the parameter of interest, and z is an unobserved (hidden) variable (The parameter θ and the hidden variable z can have large dimensions). The integral symbol is a generic symbol that represents the summation symbol when z is a discrete variable, and the multiple integration symbol when z is a continuous multidimensional variable.

A way to estimate θ is by maximizing the log-likelihood function, where the likelihood represent the probability distribution of the observed variables, y, given θ. The solution of the optimzation problem can be formalized as follows

$$\theta^* = \arg\max L(\theta) = \arg\max \log p(y|\theta)\,.$$

The concern with this approach is that the summation that appears in the formula of the log-likelihood

$$\log p(y|\theta) = \log \left(\int p(y|z,\theta) p(z|\theta) dz \right).$$

is very difficult to evaluate in general.

Algorithm. To overcome the above problem, the EM algorithm repeats an iterative process that is guaranteed to increase the likelihood of the parameter at each iteration, and that converges to a local maximum of the likelihood function. Let θ^0 denote the current value of the parameter θ. The EM update rule replaces θ^0 by θ^1, the value of θ that maximizes the following quantity

$$Q(\theta, \theta^0) = \mathrm{E}[\log p(y, z|\theta)|y, \theta^0] = \int \log p(y, z|\theta) p(z|y, \theta^0) dz,$$

and, more generally, we have

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta, \theta^t).$$

The EM algorithm is useful when the quantity $\log p(y, z|\theta)$ has a simple expression, for example, a linear function of the hidden variable z, and when the probability $p(z|y, \theta^0)$ can be easily obtained from the Bayes formula

$$p(z|y, \theta^0) \propto p(y|z, \theta^0) p(z|\theta^0).$$

Exercise. Basic arguments and remarks. Answer the following questions.

1. Find justifications for why the likelihood increases at each iteration and summarize the key arguments of the proof.

2. There are obvious limitations of the EM algorithm. Describe two potential concerns with this method.

Problem 1. An EM algorithm for detecting frequency change points in a binary sequence. We observe a sequence of binary data that consists of n observations, $y = (y_1, \ldots, y_n)$, $y_i \in \{0, 1\}$. The n binary variables correspond to independent signals from a random source, for which the frequency of 1's is unknown and may be modified at an unknown time point from θ_1 to θ_2. For example, the sequence of observations can be as follows

$$y = 01100 \ldots 00110 \| 1110011 \ldots 1100111$$

In this representation, the $\|$ symbol indicates that a change occurred at position z, that can represent any point between 2 et n. for all $i < z$, the frequency of 1's is θ_1, and for $i \geq z$ the frequency of 1's is θ_2. By convention, $z = 1$ corresponds to the situation where no change occurs. In this case, the frequency of 1's is constant, and it is equal to θ_2.

Challenge and evaluation rule. Download the data from the following URL:

`http://membres-timc.imag.fr/Olivier.Francois/sequence.data`

The data consists of a sequence of 321 binary items. The objective of the challenge is to provide a list of (one or more) change points with the following information

- Most likely change point position, z, in the range $[1, 321]$.
- Lower and upper values z_l and z_u, such that $p(z \in [z_l, z_u]) = 0.75$.
- Estimates of frequencies θ_1 and θ_2 before and after z.
- Number of iterations of the EM algorithm.

The output file must be formatted as follows

```
number position lower upper theta1 theta2 iter
1 42 35 56 .72 .43 13
2
3
```

A `README` file including comments and describing the options used when analyzing the data is required. The results will be evaluated on the basis of the 1) number of correct detections, 2) evaluation of uncertainty on each correctly detected position (ie, correctness of the difference `(upper - lower)`).

Derivation of an EM algorithm. We use the following notations

$$\theta = (\theta_1, \theta_2).$$

For $z = 2, \ldots, n$, we have

$$p(y_i|z, \theta) = \theta_1^{y_i}(1-\theta_1)^{1-y_i} \quad i = 1, \ldots, z-1$$

et

$$p(y_i|z, \theta) = \theta_2^{y_i}(1-\theta_2)^{1-y_i} \quad i = z, \ldots, n.$$

A priori, we assume that the change point z is sampled from the uniform distribution

$$p(z) = \frac{1}{n}, \quad z = 1, \ldots, n.$$

We call this distribution the *prior distribution* on z. The goal of this problem is to propose an EM algorithm for estimating the model parameter θ and for evaluating the conditional probabilities $p(z|y)$ for all $z = 1, \ldots, n$.

1. Let $z = 1$. Give a formula for the probability $p(y|z, \theta)$. Same question for $z > 1$.

2. Show that, for $z > 1$, we have

$$\begin{aligned} \log p(y|z,\theta) &= \log\theta_1 \left(\sum_{i=1}^{z-1} y_i\right) + \log(1-\theta_1)\left(z-1-\sum_{i=1}^{z-1} y_i\right) \\ &+ \log\theta_2 \left(\sum_{i=z}^{n} y_i\right) + \log(1-\theta_2)\left(n-z+1-\sum_{i=z}^{n} y_i\right) \end{aligned}$$

3. Suppose θ_1 and θ_2 are known. Using the Bayes formula, show that

$$\forall z = 2, \ldots, n, \quad R(z) = \frac{p(z|y,\theta)}{p(z=1|y,\theta)} = \prod_{j=1}^{z-1} \frac{p(y_j|\theta_1, z)}{p(y_j|\theta_2, z=1)}$$

4. Show that there is a relationship between $R(z)$ and $R(z-1)$ for all $z = 2, \ldots, n$, and propose an algorithm for computing $p(z|y)$ for all $z = 1, \ldots, n$.

5. Propose an algorithm for computing the expected value of $\sum_{i=1}^{z-1} y_i$

$$E_1 = \mathrm{E}[\sum_{i=1}^{z-1} y_i | \theta^0].$$

by averaging over all values of z. Apply the same approach to the 3 other quantities found in question 2. What is the complexity of the algorithm?

6. Describe the EM algorithm for estimating θ from y.

7. Generate simulated data for known values of θ and z. Apply the EM algorithm to the simulated data, and evaluate the convergence of the algorithm by testing several values of θ^0. Plot histograms for $p(z|y)$ using `plot(.,type ='h')`.

8. Discuss the choice of a uniform prior distribution for z. How could the EM algorithm be modified to account for informative prior distributions?

Problem 2. Gaussian mixture models. Consider a population P consisting of 2 subpopulations P_0 and P_1 with equal sizes. We sample n individuals from P, but their origins are not observed. A quantity y_i is measured for each individual. Grouping individuals based on the observations is called an *unsupervised clustering* task.

In order to group individuals into clusters, we assume that subpopulation labels are missing data. We want to estimate the cluster localization m_0 and m_1 for each group and the proportion of individuals sampled from subpopulation P_0 or P_1. In addition, we want to estimate the probability that each individual is sampled from population P_1 (or P_0).

Let $\theta = (m_0, m_1)$. We define a Gaussian mixture model as follows

$$\forall y_i \in \mathbb{R}, \quad p(y_i|\theta) = pp_1(y_i|\theta) + qp_0(y_i|\theta)$$

where $p_k(y_i|\theta) = \mathcal{N}(y_i|m_k, \sigma^2 = 1)$ is the Gaussian density function, for $k = 0, 1$. The probability p is the probability of sampling from P_1, and $q = 1 - p$.

1. To begin, we consider that the variance σ^2 is a known parameter equal to $\sigma^2 = 1$, and $p = 1/2$. For all individuals, we consider hidden variables $z_i \in \{0, 1\}$ representing their unobserved label of source population. Show that

   $$p(y_i|\theta) = p(z_i = 1)p(y_i|z_i = 1, \theta) + p(z_i = 0)p(y_i|z_i = 0, \theta)\,,$$

   where $p(z_i = 1) = 1/2$.

2. Write a computer program for drawing samples from $p(y_i|\theta)$ of size $n = 200$ (for fixed values of (p, m_0, m_1)). Check your the program is correct by drawing a histogram of the simulated data.

3. Considering the unobserved vector $z = (z_1, \ldots, z_n)$, show that

   $$\log p(y, z|\theta) = -\frac{1}{2}\sum_{i=1}^{n}(1 - z_i)(y_i - m_0)^2 + z_i(y_i - m_1)^2 + C_n$$

4. Let $n_1 = \sum_{i=1}^{n} z_i$ and $n_0 = n - n_1$. Show that the above expression is maximized for

   $$\hat{m}_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i z_i \qquad \hat{m}_0 = \frac{1}{n_0}\sum_{i=1}^{n} y_i(1 - z_i) \quad (\star)$$

5. We suppose $p(z_i = 1) = 1/2$. Show that the conditional probability of $z_i = 1$ given y_i and θ^0 is equal to

$$p(z_i = 1|y_i, \theta^0) = \frac{\exp(-(y_i - m_1^0)^2/2)}{\exp(-(y_i - m_0^0)^2/2) + \exp(-(y_i - m_1^0)^2/2)} .$$

6. In equations $(\star)$, replace the hidden variable z_i by $p(z_i = 1|y_i\theta^0)$, and show that this operation corresponds to writing the EM algorithm for estimating θ.

7. Write the EM algorithm in the `R` programming language. Generate simulated data that for known values of θ and z. Apply the EM algorithm to the simulated data, and evaluate the convergence of the algorithm by testing several initial values θ^0.

8. Extend the EM algorithm to the case where the variance σ^2 is unknown, and p is arbitrary. Extend it further to the case where the two classes have unequal (unknown) variances, and p is arbitrary.

9. Download the data from the following URL:

   `http://membres-timc.imag.fr/Olivier.Francois/data2.txt`

10. Apply the EM algorithm to the data. Evaluate the convergence of the algorithm by testing several initial values of θ^0. Report estimates for θ_1 and θ_2, and display $p(z|y)$ by using the `barplot` command to visualize the probability matrix of size $n \times 2$.

11. Install the `R` package `mclust` from the CRAN web site. Look at the different options of the `Mclust` function (models and outputs), and run the `Mclust` command on the data for $G = 1$ to 5.

12. Find a definition of the Bayesian Information Criterion (BIC). Discuss the choice of a model for the data using the BIC.

Challenge and evaluation rule. Download the data from the following URL:

`http://membres-timc.imag.fr/Olivier.Francois/matrix.data`

The data consists of a matrix of 482 rows and 3083 columns with entries in $0, 1, 2$. The objective of the challenge is to evaluate the number of clusters (for rows) in the data set and to assign a cluster label to each row. The result is a list of 482 cluster labels, one for each row. The output file must be contain the resulting list formatted as a sequence of integer values separated by space characters as follows

```
12 12 1 6 7 6 6 3 11 11 ...
```

A `README` file describing the options used when analyzing the data is required. The results will be evaluated on the basis of the confusion matrix and the number of wrongly classified rows.

Important comments: Use a dimension reduction algorithm such as *principal component analysis* or *multidimensional scaling* to reduce the dimension of the data set **before** applying model-based clustering algorithms. Then, prefer using the`Mclust` algorithm rather than reprogramming your own EM method.

Problem 3. ABO groups and genetics

A geneticist studies genotypes at the ABO locus (alleles A, B, O) for blood samples from n individuals, and gets observations for phenotypes for each individual (blood groups). Four distinct phenotypes can be observed

- type A corresponds to genotypes A/A and A/O (sample size n_A),
- type B corresponds to genotypes B/B and B/O (sample size n_B),
- type AB corresponds to genotype A/B (sample size n_{AB}),
- type O corresponds to genotype O/O (sample size n_O)

where we suppose that A/O and O/A are a same genotype (the same for B), and

$$n = n_A + n_B + n_{AB} + n_O.$$

We say that types A et B are codominants, whereas type O is recessive and is observed only if an individual carries to copies of allele O. We want to estimate the frequency p_A of allele A in the sampled population.

1. We first assume that all allele frequencies are known parameters. Using the Hardy-Weinberg principle, show that the expected number of genotypes A/A in a sample of size n is equal to
$$n_{A/A} = n_A \frac{p_A^2}{p_A^2 + 2p_A p_O}.$$
2. Find a similar equation for $n_{A/O}$.
3. Suppose we know $n_{A/A}, n_{A/O}, n_{AB}$. Give an estimate $\hat{p}_A$ of the frequency p_A
4. Use a circular (iterative) argument to compute an estimate $\hat{p}_A$ from the data.
5. Write an EM algorithm for estimating all allele frequencies.
6. Application: We observe $n = 521$ cases of peptic ulcer disease, for which $n_A = 186$, $n_B = 38$, $n_{AB} = 13$ et $n_O = 284$. Find estimates for p_A, p_B et p_O.