Practical n°1

stoehr@ceremade.dauphine.fr

Exercise 1 *(EM algorithm).*

Consider $(X, Z) \in \mathscr{X} \times \mathscr{Z}$ random variables such that $\mathscr{X}$ corresponds to an observed sample space and $\mathscr{Z}$ to a hidden sample space. The observed/incomplete data likelihood writes as

$$g(x \mid \theta) = \int_{\mathscr{Z}} f(x, z)\mathrm{d}z,$$

and is related to the unobserved/complete data likelihood $L^c(\theta \mid X, Z)$ through

$$\log g(x \mid \theta) = \log L^c(\theta \mid x, z) - \log k(z \mid x, \theta).$$

The EM algorithm [1] is an optimisation algorithm that yields a maximum likelihood estimate by iterating two steps: consider $\theta^{(t)}$ the value of the parameter at iteration t

- **E step:** compute $Q\left(\theta; \theta^{(t)}\right) = \mathbb{E}_{\theta^{(t)}}\left[\log L^c(\theta \mid x, Z) \mid x\right]$;
- **M step:** set $\theta^{(t+1)} = \arg\max_\theta Q\left(\theta; \theta^{(t)}\right)$.

A finite mixture model with $K \in \mathbb{N}^*$ components is a convex combination of K densities with unknown parameter θ_k, that is,

$$\sum_{k=1}^K p_k f(x \mid \theta_k), \quad \text{where} \quad \sum_{k=1}^K p_k = 1 \quad \text{and} \quad p_k > 0,\ k \in [\![1, K]\!].$$

Mixture models are a special instance of latent variable model. The information missing regarding the observed sample is the assignment of the data points to the different component of the model. Mixture models can hence be reformulated Mixture models incomplete data problems by introducing a random variable $Z = (Z_1, \ldots, Z_K)$ distributed according to the multinomial distribution $\mathscr{M}(1, p_1, \ldots, p_K)$. Denote $\psi = (\theta_1, \ldots, \theta_K, p_1, \ldots, p_K)$, the complete likelihood corresponding to the missing data structure is

$$L^c(\psi \mid x, z) = \prod_{i=1}^n p_{z_i} f(x_i \mid \theta_{z_i}).$$

Preliminary work

1. Show that the objective function for a mixture model is

$$Q\left(\psi, \psi^{(t)}\right) = \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(t)} \log[p_k f(x_i \mid \theta_k)], \quad \text{where} \quad \tau_{ik}^{(t)} = \frac{p_k^{(t)} f\left(x_i \,\middle|\, \theta_k^{(t)}\right)}{\sum_{\ell=1}^K p_\ell^{(t)} f\left(x_i \,\middle|\, \theta_\ell^{(t)}\right)}.$$

2. Show that the update for estimating the mixing probabilities is given by

$$p_k^{(t+1)} = \frac{N_k^{(t)}}{n}, \quad \text{where} \quad N_k^{(t)} = \sum_{i=1}^{n} \tau_{ik}^{(t)},$$

and the update for parameter θ is solution of

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \frac{\partial}{\partial \theta} \log f(x_i \mid \theta_k) = 0.$$

Application: Gaussian mixture models. Let assume that $f(\cdot \mid \theta_k)$, $k \in [\![1, K]\!]$, is a normal distribution with parameter $\theta_k = (\mu_k, \sigma_k^2)$.

3. Show that the updates of parameters μ_k and σ_k^2 are given by

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} \tau_{ik}^{(t)} x_i \quad \text{and} \quad \sigma_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} \tau_{ik}^{(t)} \left(x_i - \mu_k^{(t+1)} \right)^2.$$

4. **(a)** Write the EM algorithm for Gaussian mixture models.

(b) In order to check the code, run the algorithm on a sample of size $n = 1000$ that you will simulate from a two-component normal mixture with $p = 0.7$, $\mu_1 = 2.5$, $\mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$.

5. Estimate the parameters of a normal mixture for the faithful dataset from R.

Exercise 2 *(Metropolis Hastings)*.

Assume we aim at sampling from a distribution π (eventually known up to a constant). Metropolis Hastings algorithm [2] is a simple algorithm that produces a Markov Chain $(X_n)_{n \in \mathbb{N}^*}$, whose stationary distribution is π that may otherwise be difficult, if not impossible, to sample from.

To implement the algorithm, we must provide an irreducible transition kernel Q, that is a distribution that describes how to move randomly from one point $\mathbf{x}$ of the state space $\mathscr{X}$ to another point $\mathbf{y}$. The kernel of the algorithm is then described as follows : given a current point $\mathbf{X}_n = \mathbf{x}$ at iteration $n \in \mathbb{N}^*$

(a) Generate $\mathbf{y} \sim Q(\cdot, \mathbf{x})$,

(b) Set

$$\mathbf{X}_{n+1} = \begin{cases} \mathbf{y} & \text{with probability } \rho = 1 \wedge \dfrac{\pi(y) Q(x, y)}{\pi(x) Q(y, x)}, \\ \mathbf{x} & \text{otherwise.} \end{cases}$$

Remark. When Q is a gaussian kernel, we refer to it as random walk Metropolis Hastings.

Example 1: sampling from a density function

We aim at sampling from the Wald distribution whose density with respect to the Lebesgue measure on

$\mathbb{R}_+^*$ is given by

$$\pi(x) \propto \frac{1}{\sqrt{x^3}} \exp\left[-\frac{(x-1)^2}{2x}\right].$$

1. Sample from π using a random walk Metropolis Hastings.

Example 2: estimating an allele frequency

The haptoglobine has 3 different possible configurations AA, aa, aA. Plato, *et al.* (1964) observed the haptoglobine type on a sample of $n = 190$ people:

Genotype	AA	aa	aA
Count	10	112	68

Assuming the population is randomly mating, genes frequencies are following Hardy Weinberg Equilibrium. Given $\theta \in]0,1[$ the frequence of allele a, we have

$$\mathbb{P}[X = \mathrm{AA}] = (1-\theta)^2, \quad \mathbb{P}[X = \mathrm{aa}] = \theta^2, \quad \text{and} \quad \mathbb{P}[X = \mathrm{aA}] = 2\theta(1-\theta).$$

2. Asumming that the prior distribution on θ is uniform on $[0,1]$, sample from the posterior distribution of θ using a random walk Metropolis Hastings.

3. **(a)** Show that the posterior distribution for θ is a Beta distribution $B(293, 89)$.

(b) Compare your sample from the posterior distribution with the theoretical posterior of θ.

Exercise 3 *(Sudoku puzzle solver and simulated annealing).*

Introduced by [2], it can be used to minimize a function $\mathbf{x} \mapsto h(\mathbf{x})$ (referred to as loss function) defined on a finite set. Given a current value $\mathbf{x}^{(0)}$ and a temperature T, an iteration of the algorithm starts by proposing a new candidate $\mathbf{x}$ drawn from a uniform distribution in the vicinity of $\mathbf{x}^{(0)}$. Then the new value $\mathbf{x}^{(1)}$ is generated as follows

$$\mathbf{x}^{(1)} = \begin{cases} \mathbf{x} & \text{with probability } \rho = 1 \wedge \exp\left(-\dfrac{h(\mathbf{x}) - h\left(\mathbf{x}^{(0)}\right)}{T}\right), \\ \mathbf{x}^{(0)} & \text{otherwise.} \end{cases}$$

The method iterates the above by decreasing the temperature T at each step.

Simulated annealing algorithm is used to solve a Sudoku grid represented by a 9×9 matrix where missing values are coded by 0. Entries of the matrix are numbered from top to bottom and left to right.

1. **(a)** Give three possible functions h whose minimum corresponds to the solution of the puzzle and specify how to propose a new candidate (*i.e.* a complete grid) for each.

(b) In what follows, a candidate is drawn by randomly selecting a sub-block and then randomly flipping two of the entries that where originally empty. Write a code for the corresponding loss function.

2. Explain the function `empty_cells` and `missing_values` from the file source_sudoku.R.
3. In order to initialize the algorithm, we need a complete grid. Write a function `init_puzzle(x, empty, missing)` that randomly fills the grid consistently with the proposal mechanism.
4. Write a function `candidate(x, missing)` that provides a candidate accordingly to the aforementioned proposal mechanism.
5. Write the simulated annealing algorithm to solve a puzzle, with a temperature $T = 0.5$ that geometrically decreases at each step with common ratio 0.99999.
6. Use your program to solve the following Sudoku:

3						5	2	
1					5			6
				6	7			
6			4		2			
	1			7			8	
			8		3			7
			2	3				
8			5					4
	9	2						1

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of Chemical Physics*, 21(6):1087–1092, 1953.